Learning Theory in the AI for Science Era: From Classical Foundations to Operator Learning

by

Unique Subedi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2026

Doctoral Committee:

Professor Ambuj Tewari, Chair
Assistant Professor Saptarshi Chakraborty
Associate Professor Mahdi Cheraghchi
Associate Professor Yuekai Sun

Unique Subedi
subedi@umich.edu
ORCID iD: 0009-0004-9587-5391

DEDICATION

Of all the lessons I learned along the way, the most important was that optimism is the ultimate virtue. This thesis is dedicated to all the radical optimists out there who choose to see and create the possibility of a better world for all of us, even when the world does not make that choice easy.

ACKNOWLEDGEMENTS

It feels surreal to sign off on this thesis, and more broadly, on graduate school. I am thankful to everyone who supported me on this journey. If there is a designer, I owe my deepest gratitude for a design in which this story was possible. If, instead, everything is just the result of random evolution, I remain equally grateful to all the random bits that were flipped in just the right order to give rise to this particular instance of reality.

First, I would like to extend my sincerest gratitude to my advisor, Ambuj Tewari, for all his support. His wide-ranging interests and enthusiasm for research have been a great source of inspiration. I will be forever indebted to him for guiding me through my ever-changing interests and project directions. I would also like to thank Vinod Raman for being a great friend and collaborator. Our wide-ranging discussions of problems in machine learning, mathematics, and occasionally about the meaning of life have been an important part of my intellectual development over the last five years.

I would also like to thank Yash Patel for a fun collaboration that eventually became Chapter 9 of this thesis, and for many stimulating conversations ranging from Alexander's campaign in Persia to Dirac's life and what we could understand of his work. I am also thankful to Seamus Somerstep, Yuekai Sun, Steve Hanneke, Amirezza Shaeri, Florence Regol, and Thomas Markovich, all of whom I had the opportunity to work with and learn from during my PhD.

I would like to thank my other committee members, Saptarshi Chakraborty, Mahdi Cheragchi, and Yuekai Sun, for their guidance and feedback. I am also grateful to Karthik Sridharan for his guidance and feedback on some of the early work during my PhD.

Finally, I am deeply grateful to Micah Milinovich for his mentorship and for setting me on this path.

I am also thankful to the staff at the Department of Statistics for making my life easier, especially Becca Ussoff and Judy McDonald. I would also like to thank my friends in the department, including but not limited to Josh Wasserman, Jake Trauger, Saptarshi Roy, Sahana Rayan, Abhiti Mishra, Gabriel Patron, Marc Brooks, and Paolo Borello, for making my time in Ann Arbor memorable.

All of this would have been impossible without my friends. First, I am deeply grateful

to Bibash Dallakoti for being the friend and brother I could have ever asked for, and for encouraging me to dream big, think carefully about all aspects of life beyond work, and above all to be a judicious person. My life in Michigan was much easier because I could always count on Sanjay Barati, whose kindness, clarity of thought, and ability to remain calm even in the face of intense pressure and setbacks have been a constant source of inspiration. I am also grateful to Nischal Aryal for being such a great and reliable friend over the years in this meandering pursuit of who knows what. Finally, I am deeply thankful to Rasika Adhikari for her companionship, for being a constant source of support, for patiently listening to the never-ending stream of my random facts, and above all for bringing a sense of spontaneity to my otherwise rigid life.

I have been away from home for many years now, and I am thankful to Ayush Paudel and Priyanka Pokharel for making my first few years in Michigan feel like home. I am also thankful to Roshani Dhakal, Sanju Dhakal, Milan Khatiwada, and Roman Dhakal, who became the closest thing to family for me in Michigan. I am also grateful to Aakankshya Rijal, Sanjeev Adhikari, Rashmi Baral, Saugat Adhikari, Samiksha Gautam, and Sudip Sharma for creating a home away from home in Michigan. I would also like to thank my friends Santosh Dharel, Sagun Kandel, and Milind Osti, who have supported me from far away.

Finally, I thank my parents, my sister, and my broader family for their love and unwavering belief in me, which continually encouraged me to be a more responsible human being.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES ix

LIST OF TABLES x

LIST OF APPENDICES xi

ABSTRACT xii

CHAPTER

1 Introduction **1**

1.1 Foundations of Learning Theory 3
1.1.1 Classical Results 3
1.1.2 Contributions of This Thesis 5
1.2 Operator Learning: A Learning Theoretic Perspective 7
1.2.1 Motivation and Background 8
1.2.2 Contributions of This Thesis 10

Part I. Foundations of Learning Theory **13**

2 A Characterization of Multioutput Learnability **13**

2.1 Related works 14
2.2 Preliminaries 15
2.2.1 Batch Setting 16
2.2.2 Online Setting 17
2.3 Batch Multioutput Learnability 19
2.3.1 Batch Multilabel Classification 19
2.3.2 Batch Multioutput Regression 24
2.4 Online Multioutput Learnability 31
2.4.1 Online Agnostic-to-Realizable Reduction 32
2.4.2 Online Multilabel Classification 37
2.4.3 Bandit Online Multilabel Classification 39

2.4.4 Online Multioutput Regression . . . 40
2.5 Discussion . . . 50

3 Online Learning with Set-Valued Feedback . . . 51

3.1 Related Works . . . 53
3.1.1 Relation to List Online Classification . . . 53
3.2 Preliminaries . . . 54
3.2.1 Notation . . . 54
3.2.2 Online Learning . . . 54
3.3 Combinatorial Dimensions . . . 55
3.3.1 Relations Between Combinatorial Dimensions . . . 59
3.4 Realizable Setting . . . 60
3.4.1 A Separation Between Deterministic and Randomized Learnability . 60
3.4.2 Deterministic Learnability . . . 61
3.4.3 Randomized Learnability . . . 61
3.5 Agnostic Setting . . . 62
3.6 Applications . . . 64
3.6.1 Online Multilabel Ranking . . . 64
3.6.2 Online Multilabel Classification . . . 65

4 A Unified Theory of Supervised Online Learnability . . . 67

4.1 Related Works . . . 69
4.2 Preliminaries . . . 70
4.2.1 Notation . . . 70
4.2.2 Supervised Online Learning . . . 71
4.2.3 Combinatorial dimensions . . . 72
4.3 A Unifying Combinatorial Dimension . . . 74
4.4 Bounding Minimax Expected Regret . . . 76
4.4.1 Proof sketch of lower bound . . . 76
4.4.2 The ε_t-realizable setting . . . 77
4.4.3 Realizable-to-Agnostic conversion . . . 78
4.5 SMdim and the Finite Character Property . . . 81

5 Online Infinite-Dimensional Regression: Learning Linear Operators . . . 85

5.1 Related Works . . . 87
5.2 Preliminaries . . . 87
5.2.1 Hilbert Space Basics . . . 87
5.2.2 Online Learning . . . 89
5.3 Schatten Operators are Online Learnable . . . 90
5.3.1 Examples of p-Schatten class . . . 92
5.4 Lower Bounds and Hardness Results . . . 92
5.4.1 Lower Bounds in the Batch Setting . . . 94
5.5 Online Learnability without Sequential Uniform Convergence . . . 95
5.5.1 Batch Learnability without Uniform Convergence . . . 98
5.6 Discussion . . . 99

Part II. Operator Learning: A Learning Theoretic Perspective 101

6 Controlling Statistical, Discretization, and Truncation Errors in Learning Fourier Linear Operators . . . 101

6.1 Neural Operators . . . 102
6.1.1 Fourier Neural Operator (FNO) . . . 103
6.1.2 Our Contribution . . . 104
6.2 Related Works . . . 105
6.3 Preliminaries . . . 106
6.3.1 Notation . . . 106
6.3.2 L^2-Spaces and Fourier Analysis . . . 106
6.3.3 Sobolev Spaces . . . 107
6.4 Learning Fourier Linear Operators . . . 108
6.4.1 Problem Setting and Error Types . . . 109
6.4.2 A Constrained Least-Squares Estimator . . . 111
6.4.3 Error Bounds . . . 112
6.4.4 On Possible Extensions and Refinements of our Error Bounds . . . 114
6.5 Experiments . . . 115
6.5.1 Statistical Error . . . 116
6.5.2 Truncation Error . . . 116
6.5.3 Discretization Error . . . 116
6.5.4 Summary of Experimental Findings . . . 117
6.6 Discussion . . . 118

7 On the Benefits of Active Data Collection for Operator Learning . . . 120

7.1 Related Works . . . 122
7.2 Preliminaries . . . 123
7.2.1 Notation . . . 123
7.2.2 Distribution Over Function Space . . . 123
7.2.3 Problem Setting and Goal . . . 124
7.3 Upper Bounds Under Active Data Collection . . . 126
7.3.1 Data Collection Strategy and The Estimator . . . 127
7.3.2 Sketch of a Proof of Theorem 27 . . . 128
7.3.3 Examples of Covariance Kernels . . . 128
7.3.4 Comparison to Traditional Active Learning . . . 130
7.4 Lower Bounds on Passive Learning . . . 130
7.5 Experiments . . . 133
7.5.1 Poisson Equation . . . 133
7.5.2 Heat Equation . . . 135
7.6 Discussion . . . 136

8 Is Zero-Shot Super-Resolution Possible in Operator Learning? . . . 138

8.1 Related Works . . . 140
8.2 Problem Formulation . . . 141
8.2.1 Zero-Shot Super-Resolution vs Discretization Invariance . . . 142

8.3 Impossibility of Zero-Shot Super-Resolution: A Lower Bound 143
8.4 A Generalization Bound for Zero-Shot Super-Resolution 144
8.5 On the Assumption of Hölder Continuity of Outputs 147
8.5.1 Hölder Smoothness of Ground Truth Output Functions 148
8.5.2 Hölder Smoothness of Predicted Output Functions 149
8.6 When Inputs Are Available Only on a Discrete Grid 151
8.7 Experiments 153
8.7.1 Synthetic Data: Lower-Bound Setup 153
8.7.2 Inviscid Burgers Equation 154
8.8 Discussion 155

9 Operator Learning for Schrödinger Equation: Unitarity, Error Bounds, and Time Generalization 156

9.1 Related Works 158
9.2 Preliminaries 159
9.2.1 Time-Dependent Schrödinger Equation 159
9.2.2 Problem Formulation and Goal 160
9.3 Data Collection Strategy and Estimator 160
9.3.1 Estimator 161
9.3.2 On Unitarity of the Estimator 162
9.4 Error Analysis and Convergence Rates 162
9.4.1 Upper Bounds 163
9.4.2 Lower Bounds 163
9.4.3 Refined Upper Bound Under Stronger Assumptions on PDE Solver 164
9.5 Time Generalization 165
9.6 Experiments 167
9.6.1 Setup 167
9.6.2 Estimator Accuracy 169
9.6.3 Estimator Under Partial Observation 170
9.6.4 Time Generalization 170
9.7 Discussion 172

10 Future Directions 173

10.1 Learnability, Uniform Convergence, and Empirical Risk Minimization 173
10.2 A General Statistical Theory of Operator Learning 174
10.3 Active Data Collection in Operator Learning 175
10.4 Theory of Time Generalization 175

APPENDICES 177

BIBLIOGRAPHY 320

LIST OF FIGURES

FIGURE

1.1 Ground truth solution of the heat equation and its prediction by FNO 9

6.1 Statistical error of the DFT-based estimator for linear core of FNO 116
6.2 Truncation error of the DFT-based estimator for linear core of FNO 117
6.3 Discretization error of the DFT-based estimator for linear core of FNO 118

7.1 Error plots of active vs passive data collection protocols for Poisson equation . . 134
7.2 Log scale error plots for active vs passive data protocols for Poisson equation . 135
7.3 Error convergence for test functions with varying smoothness (Poisson equation). 136
7.4 Error plots of active vs passive data protocols for Heat equation in Log scale . . 137

8.1 Zero-shot predictions vs ground truth at multiple test resolutions (synthetic data) 139
8.2 Zero-shot predictions vs ground truth at multiple test resolutions (Burgers eqn.) 155

9.1 Squared amplitude of the initial wave, the true solution wave, and our estimator's prediction for the barrier potential with double slits. 157

E.1 Statistical error decay across sample sizes for different smoothness levels. 261
E.2 Truncation error plotted against truncation mode for various smoothness levels. 262
E.3 Discretization error as a function of grid resolution for various smoothness levels. 262

F.1 Error Plots for various estimators and data protocols for Poisson Equation . . . 285
F.2 Error Plots for various estimators and data protocols for Heat Equation 286
F.3 Error convergence for test functions with varying smoothness (Heat equation) . 286

LIST OF TABLES

TABLE

8.1 Performance of the model at various finer testing resolutions (synthetic data) . 154
8.2 Performance of the model at various finer testing resolutions (Burgers equation) 154

9.1 Summary of potential functions used for experiments 168
9.2 Relative errors of estimator across different Hamiltonians for 0.1% noise 169
9.3 Relative errors of estimator across different Hamiltonians for 10% masking . . . 171
9.4 Time-generalization errors across different Hamiltonians for 0.1% noise 171

H.1 Parameter values used in the implementation of each potential. 316
H.2 Relative errors across different Hamiltonians for 0.01% noise 317
H.3 Relative errors across different Hamiltonians for 1% noise 317
H.4 Relative errors across different Hamiltonians for 20% masking 318
H.5 Time-generalization errors across different Hamiltonians for 1% noise 318
H.6 Relative errors across different Hamiltonians within observed spectrum 319

LIST OF APPENDICES

APPENDIX

A A Characterization of Multioutput Learnability 177

B Online Learning with Set-Valued Feedback 198

C A Unified Theory of Supervised Online Learnability 219

D Online Infinite-Dimensional Regression: Learning Linear Operators . . . 229

E Controlling Statistical, Discretization, and Truncation Errors in Learning Fourier Linear Operators 237

F On the Benefits of Active Data Collection in Operator Learning 263

G Is Zero-Shot Super-Resolution Possible In Operator Learning? 287

H Operator Learning for Schrödinger Equation: Unitarity, Error Bounds, and Time Generalization 298

ABSTRACT

The ability to make reliable predictions using data lies at the core of modern machine learning. Over the past several decades, learning theory has provided precise characterizations of when such prediction is possible, along with principled learning rules and optimal error guarantees. However, many modern machine learning applications fall in the regime not considered within the classical theory. This dissertation develops new results in learning theory and uses them to build learning-theoretic foundations for operator learning, an emerging paradigm within AI for Science.

The first part of the dissertation advances the foundations of learning theory itself. Chapter 2 provides a characterization of learnability for multioutput prediction problems. Chapter 3 introduces a new setting of set-valued feedback to capture practical scenarios in which multiple answers may be correct for a given instance and predicting any one of them suffices. In this setting, we propose a combinatorial parameter that characterizes learnability and establish minimax-optimal rates. Chapter 4 then develops a general theory of supervised learnability in the online setting, unifying nearly four decades of results in online learning theory. With this unified view of classical theory in hand, the remaining chapters move beyond its traditional scope.

Chapter 5 serves as a bridge between the first and second parts of this dissertation and studies the problem of learning linear operators between infinite-dimensional spaces. We identify several new learning-theoretic phenomena that arise in this setting. In particular, we show that in infinite-dimensional settings different classes of bounded linear operators exhibit distinct optimal error rates, and that these rates can be arbitrarily slow. This contrasts with the finite-dimensional case, where the class of bounded linear operators can be learned at Monte Carlo rates. We also establish separations between learnability and uniform convergence, a property that underlies the validity of empirical risk minimization as a general learning principle in many classical tasks.

The second part of the dissertation discusses on operator learning, with the primary focus on learning solution operators of partial differential equations for surrogate modeling and scientific computation. To ensure the reliability of methods through rigorous guarantees, we adopt a learning-theoretic perspective to study operator learning. Chapter 6 begins by

identifying two additional sources of error specific to operator learning: discretization error, arising from functions being observed only on finite grids, and truncation error, arising from restricting functions to low-frequency representations, in addition to the usual statistical and approximation errors present in classical settings. We then show how these errors can be systematically controlled using the linear core of a popular Fourier neural operator architecture as an illustrative example. Chapter 7 studies the role of data collection protocols and show that transitioning from passive (i.i.d sampling) to active data collection can fundamentally change which operator classes are learnable. Moreover, even for operator classes learnable under both protocols, active data collection can significantly improve sample efficiency, sometimes yielding exponential gains over passive approaches. Chapter 8 investigates the limits of zero-shot spatial generalization to previously unseen grid points and provides both impossibility results and sufficient conditions under which such extrapolation is possible. Finally, in Chapter 9, we study physics-informed operator learning for time-dependent Schrödinger equations and show that incorporating structural constraints such as unitarity in the learning process can lead to both improved empirical performance and stronger theoretical guarantees.

CHAPTER 1

Introduction

The central goal of learning theory is to understand when learning from data is possible. A positive resolution of this objective naturally leads to the next question of *how* to learn. An even more fundamental question that precedes both is what it means to learn successfully. For the purpose of this thesis, we define successful learning as the ability to make reliable predictions on new, previously unobserved examples. The question of when learning is possible turns out to be particularly deep, and meaningful progress toward this question led to the foundations of modern learning theory in the mid 1960s. In contrast, the questions of how to learn from data have a much longer intellectual history.

For much of human history, the primary mode of "learning" consisted of constructing lookup tables from observations and using them to make predictions. For example, many ancient societies systematically recorded the positions of celestial bodies at regular intervals and compiled astronomical catalogs. These records were then used to predict eclipses, seasonal cycles, and other recurring astronomical phenomena.

Moving beyond lookup-table-based prediction, perhaps one of the earliest attempts to learn a mathematical relationship directly from observational data is due to Kepler [1619]. Using the extensive and highly precise planetary position measurements recorded by the Danish astronomer Tycho Brahe, Kepler inferred mathematical laws governing planetary motion. In many respects, his approach bears conceptual similarities to modern data-driven learning. Beginning with a broad model class that included various possible geometric descriptions of planetary motion, he used observational data to systematically rule out competing hypotheses, such as perfect circular orbits, epicycles, and equants, ultimately converging on elliptical orbits as the correct description. Based on the observed data, he then formulated his famous law relating the orbital period (T) of a planet to semi-major axis (a) of its orbit; that is, $T^2 \propto a^3$. Kepler's work turned out to be remarkably influential in the history of science as it set the stage for Newtonian mechanics and the modern mathematical formulation of predictive physical science.

A more algorithmic approach to learning based on a well-defined learning principle is Legendre's work on the method of least squares [Legendre, 1805], where he proposed fitting linear models to observed data using least squares. A few years later, Gauss published his work on least squares [Gauss, 1809] and claimed that he had been using the method as early as 1795, nearly a decade before Legendre's publication. Notably, both Legendre and Gauss developed the method of least squares to fit mathematical models to astronomical and measurement data, highlighting the longstanding connection between learning from data and scientific application.

Rather than viewing least squares merely as a computational rule, Gauss also sought to justify it from first principles. First, he provided a probabilistic justification by showing that least squares maximizes likelihood under normality of observational errors. In his later work, Gauss [1823] moved beyond purely probabilistic arguments and justified least squares in terms of predictive reliability. In particular, he showed that among all linear unbiased estimators formed from linear combinations of measurements, the least squares estimator is optimal in the sense that it minimizes the root mean squared error. This is arguably one of the earliest instances of a rigorous error analysis of a learning principle.

As for when learning is possible, systematic investigation of this question led to the modern mathematical formulation of learning theory in the 1960s, most notably through the work of Vapnik and Chervonenkis, now known as Vapnik–Chervonenkis (VC) theory [Vapnik and Chervonenkis, 1971, 1974, Vapnik, 1982]. See Vovk, Papadopoulos, and Gammerman [2015] for a detailed historical account of the development of the theory, including commentary by Chervonenkis himself. Due to both its elegance and intellectual depth, the theory was rapidly adopted by broader mathematics community and became a central tool in the development of empirical process theory [Dudley, 1978, Steele, 1978, Assouad, 1983, Pollard, 1990, Talagrand, 1994].

Following the introduction of the probably approximately correct (PAC) learning model by Valiant [1984], the work of Blumer, Ehrenfeucht, Haussler, and Warmuth [1989] showed that VC dimension provides necessary and sufficient conditions for distribution-free learnability of classes of Boolean concepts under Valiant's PAC model. These developments brought VC theory into the computer science community and gave rise to the field of computational learning theory. Subsequent works such as [Van Der Vaart and Wellner, 1996, Geer, 2000] established VC theory as an important tool in the statistical analysis of estimation problems. In parallel, Littlestone [1987] initiated the development of an analog of VC theory for online learning and sequential decision-making problems. Additionally, development of algorithmic ideas such as boosting [Freund and Schapire, 1997] and support vector machines [Boser, Guyon, and Vapnik, 1992, Cortes and Vapnik, 1995] brought core learning-theoretic ideas to

applied machine learning. Thus, learning theory and its tools have had broad influence across probability theory, computer science, statistics, sequential decision making, and applied machine learning.

As foreshadowed in the early works of Legendre [1805] and Gauss [1809, 1823], an algorithmic approach to learning has long been closely tied to scientific applications. Yet, despite the central role of data-driven inference in modern scientific methodology, learning theory itself has not historically played as prominent a role in shaping methods used in scientific practice. This gap motivates the central theme of this thesis: developing learning-theoretic foundations for the emerging paradigm of operator learning, which has become a key approach within modern AI for Science. We first apply classical learning-theoretic tools to analyze operator learning methods. We then show that this setting gives rise to fundamentally new questions in learning theory, which motivates further development of its foundational principles.

Accordingly, this thesis is divided into two parts. Part I discusses classical foundations of learning theory and presents our contributions towards extending them. Part II turns to operator learning and discusses our work on developing learning-theoretic foundations for this emerging paradigm.

1.1 Foundations of Learning Theory

In this section, we first review the classical foundations of learning theory. We then provide a brief overview of Part I of this thesis that extend these classical foundations.

1.1.1 Classical Results

To formally discuss classical results in learning theory, let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively, and let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote a class of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. In statistical learning theory, a learner is provided with n i.i.d. samples $S_n := \{(x_i, y_i)\}_{i=1}^n$ drawn from a distribution μ on the product space $\mathcal{X} \times \mathcal{Y}$. Using this sample and a predefined learning rule, the learner constructs then an estimator $\widehat{f}_n$. For simplicity, we use $\widehat{f}_n$ to denote both the estimator and the learning rule itself. The estimator $\widehat{f}_n$ is evaluated relative to the best function within the class $\mathcal{F}$, namely in terms of excess risk.

For a prespecified loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the excess risk is defined as

$$\mathcal{E}_n(\widehat{f}_n, \mathcal{F}) := \sup_{\mu} \left(\mathop{\mathbb{E}}_{S_n \sim \mu^n} \left[\mathop{\mathbb{E}}_{(x,y)\sim\mu} \left[\ell\Big(\widehat{f}_n(x), y\Big) \right] - \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim\mu} \left[\ell\Big(f(x), y\Big) \right] \right] \right), \tag{1.1}$$

where the supremum is taken over all distributions μ supported on $\mathcal{X} \times \mathcal{Y}$. This setting

is commonly referred to as the *agnostic* model of learning. In the special case where there exists an $f^\star \in \mathcal{F}$ such that $y = f^\star(x)$ for all $(x, y) \sim \mu$, we say that we are in the *realizable* setting.

Within this framework of learning, we say that the class $\mathcal{F}$ is *learnable* if and only if there exists a learning rule for which

$$\mathcal{E}_n(\widehat{f}_n, \mathcal{F}) \longrightarrow 0 \quad \text{as } n \to \infty.$$

Equivalently, one can define this objective using the (ε, δ)-formulation of the PAC model by Valiant [1984] (see also [Shalev-Shwartz and Ben-David, 2014, Chapter 3]). Under this formulation, a class $\mathcal{F}$ is learnable if for every $\varepsilon, \delta > 0$ there exists a sample size $m(\varepsilon, \delta) \in \mathbb{N}$ and a learning rule such that for all $n \geq m(\varepsilon, \delta)$,

$$\sup_{\mu} \mathbb{P}\left(\mathop{\mathbb{E}}_{(x,y)\sim\mu}\Big[\ell(\widehat{f}_n(x), y)\Big] - \inf_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim\mu}\Big[\ell(f(x), y)\Big] > \varepsilon \right) \leq \delta.$$

For bounded loss functions, which will always be the case in this thesis, these two definitions of learnability are *qualitatively* equivalent.

Now that we have formally defined *what* it means to learn, we turn to the questions of when learning is possible and how it should be carried out. A central result of VC theory addresses the case of binary classification, where $\mathcal{Y} = \{0, 1\}$ and the loss is $\ell(y, y') = \mathbb{1}\{y \neq y'\}$. In the framework described above, a learning rule with vanishing excess risk exists if and only if $\mathrm{VC}(\mathcal{F}) < \infty$ (see [Blumer, Ehrenfeucht, Haussler, and Warmuth, 1989] and [Shalev-Shwartz and Ben-David, 2014, Chapter 28]). In this sense, the VC dimension characterizes the learnability of binary classification problems by providing a precise answer to the question of *when* learning is possible.

Beyond this qualitative characterization, the VC dimension also provide a quantitative characterization of optimal error rates. In the agnostic setting, we have

$$\mathcal{E}_n(\widehat{f}_n, \mathcal{F}) = \Theta\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}\right),$$

while in the realizable setting the optimal rate improves to

$$\Theta\left(\frac{\mathrm{VC}(\mathcal{F})}{n}\right).$$

Moreover, whenever $\mathrm{VC}(\mathcal{F}) < \infty$ and learning is possible, empirical risk minimization (ERM) is a optimal learning principle in the agnostic setting. In the realizable setting,

standard ERM remains optimal up to an additional logarithmic factor in the error rate (see [Hanneke, 2016] and [Shalev-Shwartz and Ben-David, 2014, Chapter 28]). Thus, VC theory not only determines when learning is possible, but also provides principled answers to how learning should be performed and what error rates and notions of optimality can be achieved.

Although the original VC theory was formulated for binary classification in statistical settings, its central insight of providing qualitative and quantitative characterizations of learnability via combinatorial dimension has had a profound and lasting influence on the development of learning theory. For example, Natarajan [1989a,b], Ben-David, Cesa-Bianchi, and Long [1995] characterized learnability in multiclass classification with finitely many labels using a combinatorial quantity now known as the Natarajan dimension. This characterization was further extended to the case $|\mathcal{Y}| = \infty$ by Brukhim, Carmon, Dinur, Moran, and Yehudayoff [2022] via the Daniely–Shalev-Shwartz (DS) dimension proposed by Daniely and Shalev-Shwartz [2014]. Similarly, for scalar-valued prediction with $\mathcal{Y} = [0, 1]$ and loss $\ell(y, y') = |y - y'|^p$ for $p \in \{1, 2\}$, Bartlett, Long, and Williamson [1996] and Alon, Ben-David, Cesa-Bianchi, and Haussler [1997] established learnability in terms of the fat-shattering dimension, originally introduced by Kearns and Schapire [1990].

1.1.2 Contributions of This Thesis

In Chapter 2, we extend characterizations of learnability to multioutput prediction problems, where each instance is labeled by a vector-valued target. In particular, we extend classification results to target spaces of the form $\mathcal{Y} = \{0, 1\}^K$ for finite $1 < K < \infty$. Our characterization holds for any loss ℓ satisfying the identity of indiscernibles property, that is, $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$. We further extend this characterization to bounded continuous output spaces $\mathcal{Y} \subseteq \mathbb{R}^K$. For these continuous prediction problems, our characterization holds for all ℓ_p norms and their natural variants. More precisely, we show that a multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is learnable if and only if each single-output component class $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is learnable. Here, a multioutput function $f \in \mathcal{F}$ is written as $f = (f_1, \ldots, f_K)$, with targets $y = (y_1, \ldots, y_K)$ for $y \in \mathcal{Y}$. Beyond statistical settings, we also extend these characterizations to an adversarial online model.

The adversarial model of online learning was introduced by Littlestone [1987] as a sequential counterpart to VC theory, motivated by problems in online prediction and sequential decision making. In the online setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, the adversary selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner, let's call it $\mathcal{A}$, then makes a (possibly randomized) prediction $\hat{y}_t \in \mathcal{Y}$. The adversary subsequently reveals y_t, and the learner incurs loss $\ell(y_t, \hat{y}_t)$.

Given a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{y}_t$ such that its regret,

$$\mathrm{R}_T(\mathcal{A}, \mathcal{F}) := \sup_{(x_1,y_1),\ldots,(x_T,y_T)} \left(\sum_{t=1}^{T} \mathbb{E}\big[\ell(y_t, \hat{y}_t)\big] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(y_t, f(x_t)) \right),$$

is small. Here, the expectation is taken over the internal randomness of the learner $\mathcal{A}$. We say that the class $\mathcal{F}$ is learnable with respect to ℓ if there exists a learning rule such that

$$\limsup_{T \to \infty} \frac{\mathrm{R}_T(\mathcal{A}, \mathcal{F})}{T} = 0.$$

In this model, a combinatorial parameter introduced by Littlestone [1987], now referred to as Littlestone dimension, characterizes learnability in binary classification. A generalization of the Littlestone dimension to multiclass settings was introduced by Daniely, Sabato, Ben-David, and Shalev-Shwartz [2015] and was later shown by Daniely, Sabato, Ben-David, and Shalev-Shwartz [2015], Hanneke, Moran, Raman, Subedi, and Tewari [2023] to characterize learnability in this setting. Similarly, the sequential analog of the fat-shattering dimension characterizes learnability for regression problems [Rakhlin, Sridharan, and Tewari, 2015a].

Chapter 3 studies learnability under such adversarial online model. In this chapter, we further develop the theory of online learning for a more general setting of set-valued feedback, where multiple ground-truth answers may be correct and the learner succeeds as long as it predicts one of them. More precisely, the learner predicts a single label $\hat{y}_t \in \mathcal{Y}$, while the ground truth is given by a set of correct labels $S_t \subseteq \mathcal{Y}$. The loss is defined as

$$\ell(S_t, \hat{y}_t) = \mathbb{1}\{\hat{y}_t \notin S_t\}.$$

This phenomenon of multiple valid ground-truth outputs arises naturally in many settings. For example, in molecular structure prediction, many molecules admit multiple feasible conformations, and predicting any physically valid conformation is considered correct. Interestingly, this setting reveals several novel learning-theoretic phenomenon. One of them is a separation of deterministic and randomized learnability in the realizable setting. More precisely, in standard multiclass classification for online learning, every learnable class is learnable using a deterministic learning rule in the realizable setting. However, in this setting, we show that there exist classes that are learnable with randomized learning rules but not learnable with any deterministic learning rules, even in the realizable setting. To capture this separation, we introduce two new combinatorial dimensions that tightly characterize deterministic and randomized online learnability. We then use these characterizations to es-

tablish minimax regret bounds for known practical learning tasks such as online multilabel ranking and online multilabel classification.

Our effort to push the frontiers of learning theory culminates in a unified theory of online learnability in Chapter 4. In this chapter, we study the online learnability of function classes with respect to arbitrary, but bounded, loss functions. Prior to this work, no characterization of online learnability was known at this level of generality. We close this gap by showing that existing techniques can be extended to characterize online learnability for any supervised learning problem with bounded loss. Towards this endeavor, we introduce a new scale-sensitive combinatorial dimension, called the Sequential Minimax dimension, which generalizes all previously known combinatorial dimensions in online learning theory and provides matching upper and lower bounds on the minimax value. Thus, the results generalize four decades of work in online learning theory dating back to Littlestone [1987].

Continuing our theme of learning under abstract target spaces, Chapter 5 studies regression problems where the target space $\mathcal{Y}$ is an infinite-dimensional Hilbert space. More precisely, we consider the setting where $\mathcal{X}$ and $\mathcal{Y}$ are infinite-dimensional Hilbert spaces and $\mathcal{F}$ is a class of bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$. This problem is closely related to operator learning methods used in surrogate modeling for partial differential equations, which will be discussed in greater detail in the following section. Here, however, we study this setting as a representative example of a learning problem that leads to a fundamentally new learning-theoretic landscape. We show that the class of linear operators with uniformly bounded p-Schatten norm (the ℓ_p norm of the singular values) is online learnable for any $p \in [1, \infty)$ with a regret rate of $\Theta(T^{1-\frac{1}{p}})$ for $p \geq 2$. In contrast, we prove an impossibility result by showing that a class of linear operators with uniformly bounded operator norm, corresponding to the ℓ_∞ norm of the singular values, is not online learnable. This result is surprising in light of the fact that in finite dimensions the class of linear operators with uniformly bounded operator norm is always learnable with rate $\Theta(T^{\frac{1}{2}})$, where the constant depends on d. Remarkably, such dependence of constants lead to rate separation when $d \to \infty$. In addition, for almost all classical learning problems, learnability is equivalent to the property of uniform convergence. This equivalence underlies the validity of ERM as a learning principle in the batch setting. However, in this chapter, we show that learnability and uniform convergence are not equivalent for regression with infinite-dimensional target.

1.2 Operator Learning: A Learning Theoretic Perspective

Artificial Intelligence (AI) has seen rapid advances across domains ranging from language and vision modeling to applications in the physical sciences. In particular, AI is transforming

scientific research by enabling new methods for modeling complex systems [Tang, Kurths, Lin, Ott, and Kocarev, 2020] and optimizing scientific workflows [Wang et al., 2023]. A prominent example is protein structure prediction with AlphaFold [Jumper et al., 2021], an achievement recognized with the 2024 Nobel Prize in Chemistry [The Nobel Committee, 2024]. These developments have given rise to the emerging paradigm of "AI for Science" [Zhang et al., 2025b], which aims to systematically integrate AI into scientific discovery. Operator learning is one of the central approaches within this paradigm. In this section, we provide a brief overview of operator learning and discuss our works on developing a learning theoretic foundation of this emerging field.

1.2.1 Motivation and Background

In mathematics, a mapping between infinite dimensional function spaces is often called an operator. Operator learning is an area at the intersection of applied mathematics, computer science, and statistics which studies how we can learn such operators from data. Its primary application is the development of fast and accurate surrogate models [Bhattacharya, Hosseini, Kovachki, and Stuart, 2021] for the solution operators of partial differential equations (PDEs). Additionally, as a data-driven approach, operator learning techniques can be used to develop black-box simulators that simulate system behavior based on observed experimental data [You, Zhang, Ross, Lee, Hsu, and Yu, 2022a, You, Zhang, Ross, Lee, and Yu, 2022b], even when the underlying mathematical model is unknown.

Before formally defining the problem of operator learning, let us discuss a motivating example that illustrates its relevance. Many physical systems are governed by PDEs, which describe how the system evolves given particular initial conditions. A classic example is the heat equation [Evans, 2022, Section 2.3]

$$\frac{\partial u}{\partial t} = \tau \nabla^2 u, \tag{1.2}$$

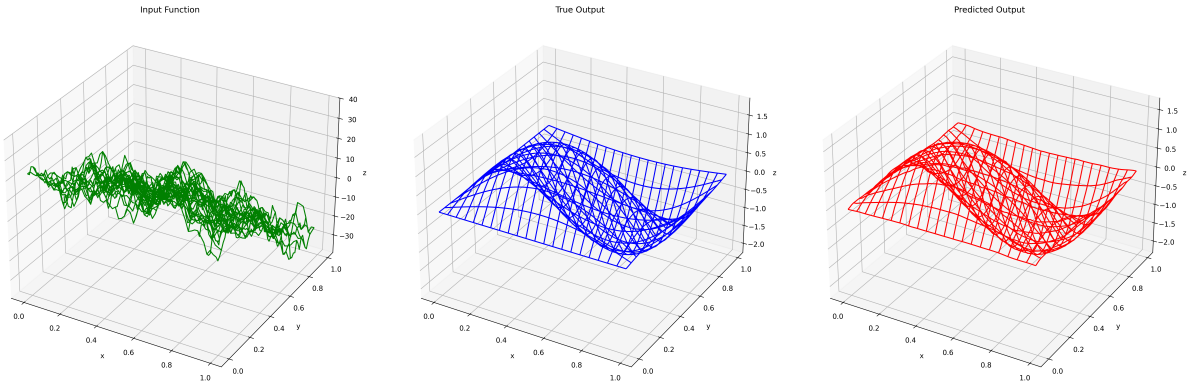
that arises in heat conduction and diffusion problems. Here, $\tau > 0$ could be the thermal conductivity of a material, $u : \Omega \times [0, \infty) \to \mathbb{R}$ for some set $\Omega \subseteq \mathbb{R}^d$ could define a temperature profile at any given space-time coordinate, and ∇^2 is the Laplacian operator defined as $\nabla^2 u := \sum_{j=1}^{d} \partial^2 u / \partial x_j^2$. The solution of the heat equation can be written using a linear operator defined as

$$\exp(\tau t \nabla^2) := \sum_{k=0}^{\infty} \frac{(\tau t \, \nabla^2)^k}{k!}.$$

That is, given an initial condition u_0, the solution function can be written as $u_t = \exp(\tau t \nabla^2) u_0$ for any time point $t > 0$ [Hunter, 2023, Chapter 5.4].

This solution operator is useful primarily for conceptual understanding and cannot be used to obtain the solution function in all but a few cases of simple domain geometry and simple initial conditions. The solution is generally obtained using PDE solvers which use numerical methods to map the initial conditions u_0 to u_t at some desired time point $t > 0$. Such solver starts from scratch for every new initial condition u_0 of interest. Since the solver is computationally slow and expensive, this ab initio approach to evaluating solutions can be limiting in applications such as engineering design where the solution needs to be evaluated for many different initial conditions [Umetani and Bickel, 2018]. To solve this problem, operator learning aims to learn the solution operator directly from the data. By amortizing the computational cost through upfront training, these learned operators allow for significantly efficient solution evaluation compared to traditional solvers while sacrificing a small degree of accuracy.

More precisely, for some prespecified time point $t = T$, let $\mathrm{G} := \exp(\tau T \nabla^2)$ denote the solution operator of interest. Then, given training data $(v_1, w_1), \ldots, (v_n, w_n)$ where v_i is the initial condition and $w_i = \mathrm{G}(v_i)$ is the solution at time point T, operator learning involves estimating an approximation $\widehat{\mathrm{F}}_n$ of G by searching over a predefined operator class $\mathcal{F}$ [Kovachki, Li, Liu, Azizzadenesheli, Bhattacharya, Stuart, and Anandkumar, 2023, Section 2]. Once trained, the estimated operator can be used to predict an approximate solution $\widehat{w} = \widehat{\mathrm{F}}_n(v)$ for a new initial condition v. The objective is to design an estimation procedure such that $\widehat{w}$ closely approximates the true solution $w = \mathrm{G}(v)$ under a suitable metric. For illustration, Figure 1.1 compares the Fourier neural operator's prediction and the actual output from a PDE solver for heat equation.



(a) Input (v) (b) Output (w) (c) Prediction ($\widehat{w}$)

Figure 1.1: Input function, ground truth solution of the heat equation, and the predicted solution by a Fourier Neural Operator. Here, we use $\Omega = [0, 1)^2, \tau = 0.05$, and $T = 1$.

We approach operator learning from the perspective of statistical learning theory [Vapnik, 2000], with the key distinction that both the input and output spaces consist of functions. Specifically, the input and output spaces $\mathcal{X}$ and $\mathcal{Y}$ are typically Banach spaces of functions defined over some domain $\Omega \subseteq \mathbb{R}^d$. Within this framework, the statistical learning setup remains as described earlier, with the objective of minimizing the excess risk defined in (1.1).

A notable difference lies in the data-generating process. We assume that input functions are sampled as $x \sim \mu_{\mathcal{X}}$, and the corresponding outputs are generated through a ground-truth operator G, so that $y = \mathrm{G}(x)$. In practice, the outputs $\mathrm{G}(x)$ are typically obtained in discrete form using numerical PDE solvers. Unlike traditional machine learning applications, where $\mu_{\mathcal{X}}$ is treated as an unknown but fixed data-generating distribution, in operator learning the user often specifies $\mu_{\mathcal{X}}$ deliberately when constructing a surrogate model for a target application. A common choice is a Gaussian process prior with a covariance kernel designed to concentrate mass on the class of input functions of interest [Bhattacharya et al., 2021, Kovachki et al., 2023].

Although the ground-truth G need not belong to the candidate class $\mathcal{F}$, the learned operator $\widehat{\mathrm{F}}_n$ is typically chosen from $\mathcal{F}$ in practice. Common choices for $\mathcal{F}$ include classes of bounded linear operators, operator-valued reproducing kernel Hilbert spaces (RKHS), and neural operators [Subedi and Tewari, 2026]. As for the loss function, a common choice is $\ell(\widehat{y}, y) = \|\widehat{y} - y\|_{\mathcal{Y}}^q$ for $q \geq 1$, where $\|\cdot\|_{\mathcal{Y}}$ is the canonical norm of the Banach space $\mathcal{Y}$.

1.2.2 Contributions of This Thesis

Building on the preceding discussion, Chapter 6 turns to the question of how to build a rigorous learning-theoretic foundation for this paradigm. Rather than focusing on specific PDE models, we adopt a broader perspective and study the fundamental statistical principles underlying operator learning. As a concrete model problem, we consider the linear layer of the influential Fourier Neural Operator (FNO) architecture proposed by Li, Kovachki, Azizzadenesheli, Liu, Bhattacharya, Stuart, and Anandkumar [2021]. Our primary objective is to understand how operator learning differs from classical machine learning settings and to identify the new analytical tools required to analyze it rigorously. To this end, we start by identifying the distinct types of errors that are unique to operator learning. In addition to the standard statistical error due to finite sample size, operator learning introduces a discretization error, since functional data are typically observed only on a finite grid of domain points. Furthermore, truncating high-frequency Fourier modes gives rise to a truncation error. Finally, we show how these error components can be systematically quantified and controlled.

While a statistical theory of operator learning is of interest, it is not immediately clear that the standard i.i.d.-based statistical framework is the best model for studying operator learning in the context of PDE modeling. Unlike traditional supervised learning settings, here the learner can generate training data by querying a numerical solver, and therefore is not inherently restricted to i.i.d. samples from a fixed source distribution. Moreover, since generating training data typically requires computationally expensive numerical simulations, the learner should ideally generate data adaptively so that the cost of training is justified by improved performance at evaluation time.

Motivated by this observation, Chapter 7 studies active data collection strategies for operator learning when the target operator is linear and the input functions are drawn from a mean-zero stochastic process with continuous covariance kernel, a setting widely considered in applied works [Bhattacharya et al., 2021, Li et al., 2021, Kovachki et al., 2023]. First, we construct a natural problem setting and identify a class of operators that is learnable under an active learning protocol but not learnable under passive (i.i.d.) sampling. Moreover, we show that under sufficiently rapid eigenvalue decay of the covariance kernel, the active learning protocol can achieve arbitrarily fast convergence rates in the number of samples. For instance, for kernels such as the Gaussian (RBF) kernel, the error can decay exponentially fast at a rate of e^{-n}. In contrast, under passive data collection strategies, the convergence rate is cannot exceed a linear decay, n^{-1}. This work clearly establishes the benefit of active data collection protocol over passive ones for operator learning.

The remaining two chapters investigate a couple of new phenomena specific to operator learning that do not arise in traditional learning problems. One such phenomenon is the presence of discretization error. For this error, practitioners have observed an intriguing property known as zero-shot super-resolution, where a model trained on coarse grids produces accurate predictions on finer test grids without additional retraining. Despite strong empirical evidence, the theoretical foundations of this phenomenon remain poorly understood. Accordingly, Chapter 8 provides a systematic theoretical investigation of zero-shot super-resolution in operator learning. We first show that zero-shot super-resolution can be information-theoretically impossible even in seemingly benign settings, such as when input functions are available over the entire continuum and the ground truth is a simple rank-one linear operator. We then identify Hölder smoothness of the output functions as a sufficient condition for zero-shot super-resolution and derive corresponding generalization guarantees. Finally, we complement our theoretical findings with experimental results that validate the identified failure modes.

Another important distinction between operator learning and traditional machine learning lies in the existence of a well-defined ground-truth operator G. In fact, in many operator

learning problems, the learner possesses partial prior knowledge about G through the structure of the underlying PDE, together with strong oracle access via numerical solvers. This is in contrast to settings such as language or vision modeling, where a well-defined ground-truth function may not even exist. This additional structure in operator learning allows for the design of PDE-specific architectures and the incorporation of domain knowledge directly into the training process. Such approaches are commonly referred to as "physics-informed learning", reflecting the fact that many PDEs arise from physical science [Raissi, Perdikaris, and Karniadakis, 2019, Li, Zheng, Kovachki, Jin, Chen, Liu, Azizzadenesheli, and Anandkumar, 2024b]. The central hypothesis of physics-informed learning is that incorporating these PDE-based constraints can significantly improve sample efficiency of learning.

Chapter 9 puts this hypothesis to test by investigating PDE-informed operator learning for the time-dependent Schrödinger equation, where the Hamiltonian may vary with time. Beyond its wide-ranging applications and extensive literature on surrogate modeling, the Schrödinger solution operator possesses a special structural property: it is linear and unitary with respect to the L^2 norm. This naturally raises the question of whether enforcing unitarity can lead to more efficient operator learning methods for the Schrödinger equation. We show that the answer is affirmative. In this chapter, we introduce a linear estimator for the evolution operator that preserves a weak form of unitarity. Experiments across a range of physically relevant Hamiltonians, including hydrogen atoms, ion traps for qubit design, and optical lattices, demonstrate that our estimator achieves relative errors that are 10^{-2} to 10^{-3} times smaller than those of state-of-the-art off-the-shelf methods such as the Fourier Neural Operator and DeepONet that do not exploit such structure. Furthermore, unlike for neural-network-based approaches, we can establish rigorous theoretical guarantees on the prediction error of the learned operator.

Chapter 9 also uses the Schrödinger equation to study another intriguing phenomenon known as time generalization. For time-dependent PDEs, the ground-truth operator G is implicitly indexed by the terminal time T, mapping an initial condition u_0 to the solution u_T at time $T > 0$. A natural question is whether an operator estimator $\widehat{\mathrm{F}}$ trained using data up to time T can be used to extrapolate to future times $T' > T$ at test time in a zero-shot manner, that is, without additional retraining. Existing works have empirically observed time-extrapolation capabilities of Fourier Neural Operators for Schrödinger equation [Mizera, 2023] and, more broadly, for time-dependent quantum spin systems [Shah, Patti, Berner, Tolooshams, Kossaifi, and Anandkumar, 2024]. In this chapter, we formally define the problem of time extrapolation and establish generalization bounds for our proposed estimator. In particular, our results show that time generalization is indeed possible when the potential is time-independent and sufficiently smooth.

Part I
Foundations of Learning Theory

CHAPTER 2

A Characterization of Multioutput Learnability

In this chapter[1], we consider the problem of multioutput learning in the batch and online settings. Multioutput learning is a problem where an instance is labeled by a vector-valued target. This is a generalization of scalar-valued-target learning settings such as multiclass classification and regression. Multioutput learning has enjoyed a wide range of practical applications like image tagging, document categorization, recommender systems, an weather forecasting to name a few. This widespread applicability has motivated the development of several practical methods [Kapoor et al., 2012, Borchani et al., 2015, Yang et al., 2020, Xu et al., 2013, Nam et al., 2017], as well as theoretical analysis [Koyejo et al., 2015, Liu and Tsang, 2015]. However, the most fundamental question of learnability in a multioutput setting remains unanswered.

As highlighted in the previous chapter, characterizing learnability is the first step toward understanding any statistical learning problem. For a brief recap, the fundamental theorem of statistical learning states that binary classification is learnable if and only if the Vapnik–Chervonenkis (VC) dimension is finite [Vapnik and Chervonenkis, 1971, 1974, Blumer et al., 1989]. For multiclass problems with finitely many labels, the Natarajan dimension [Natarajan, 1989a] characterizes learnability [Ben-David et al., 1995]. In the infinite-label setting, multiclass learnability is characterized by the Daniely–Shalev-Shwartz (DS) dimension [Daniely and Shalev-Shwartz, 2014, Brukhim et al., 2022]. In online learning, the Littlestone dimension [Littlestone, 1987] and its multiclass generalization [Daniely et al., 2011] characterize learnability for binary and finite multiclass problems, respectively. For scalar-valued regression, learnability in batch and online settings is characterized by the fat-shattering [Alon et al., 1997, Bartlett et al., 1996] and sequential fat-shattering dimensions [Rakhlin

[1]This chapter is based: Vinod Raman*, Unique Subedi*, and Ambuj Tewari (2024). *A Characterization of Multioutput Learnability.* Journal of Machine Learning Research, 25(342): 1–54.

et al., 2015a] respectively. Surprisingly, to our best knowledge, no such characterization of the learnability of multioutput function classes exists in the literature.

In this paper, we close this gap by characterizing the learnability of function classes $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} \subseteq \mathbb{R}^K$ is vector-valued target space for some $K \in \mathbb{N}$. Let us define scalar-valued function classes $\mathcal{F}_k = \{x \mapsto \langle f(x), e_k \rangle : f \in \mathcal{F}\}$ for each $k \in [K]$, where $\{e_1, \ldots, e_K\}$ is the standard basis of $\mathbb{R}^K$. Similarly, define $\mathcal{Y}_k := \{\langle y, e_k \rangle : y \in \mathcal{Y}\}$. Our main result, informally stated below, asserts that $\mathcal{F}$ is learnable if and only if each coordinate restriction $\mathcal{F}_k$ is learnable.

Theorem. *(Informal) A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is learnable if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is learnable.*

We prove a version of this result in four canonical settings: batch classification, online classification, batch regression, and online regression. For the batch settings, we consider the PAC framework and for the online settings, we consider the fully adversarial model. In addition, our result holds for a wide family of loss functions. A unifying theme throughout all four learning settings is our ability to *constructively* convert a learning algorithm $\mathcal{A}$ for $\mathcal{F}$ into learning algorithm $\mathcal{A}_k$ for $\mathcal{F}_k$ for each $k \in \{1, ..., K\}$ and vice versa. We show that even for multioutput losses that tightly "couple" the K coordinates of a function class, their learnability still depends on the learnability of each coordinate. For the batch setting, our algorithmic techniques use the realizable-to-agnostic conversion introduced by Hopkins et al. [2022]. In the online setting, we provide a new realizable-to-agnostic conversion similar in the spirit of Hopkins et al. [2022]. In principle, both ours and Hopkins et al. [2022]'s realizable-to-agnostic conversion is based on the idea of using algorithms to construct a cover of function classes, originally introduced in the seminal work of Ben-David et al. [2009].

2.1 Related works

Multilabel classification has been extensively studied in the batch setting. We review a few works here and also refer the reader to the references therein. Dembczyński et al. [2010] quantify the 0-1 risk of multilabel classifiers trained by minimizing the Hamming loss and vice versa. Dembczyński et al. [2012] and Chekina et al. [2013] study how exploiting dependencies between labels can improve the predictive performance of multilabel classifiers and how such exploitation interacts with loss minimization. Jain et al. [2016] consider the case where the label set is extremely large and design new loss functions to handle these settings. Busa-Fekete et al. [2022] derive upper and lower bounds on the excess risk for non-parametric and parametric function classes for various loss functions assuming label sparsity. Gao and Zhou

[2011] and Koyejo et al. [2015] study the consistency of surrogate loss functions for multilabel classification. Finally, Gentile and Orabona [2012] consider online multilabel classification under partial feedback and present a novel algorithm based on second-order descent methods.

There is also a long history of studying least squares estimators for multioutput linear models in the statistical literature, see [Rao, 1965, Brown and Zidek, 1980] and references therein. The topic received widespread attention in learning theory following the seminal work of Micchelli and Pontil [2005] in RKHS methods for vector-valued regression. We refer the reader to a comprehensive review of kernel methods for vector-valued regression by Alvarez et al. [2012]. An early work of Gnecco and Sanguineti [2008] provides estimation and approximation error of vector-valued functions using Rademacher complexity. Following the influential work of Maurer [2016] on Rademacher contraction inequalities for vector-valued functions, there have been works on the Rademacher analysis of vector-valued functions (see Cortes et al. [2016], Reeve and Kaban [2020], Yousefi et al. [2018], Foster and Rakhlin [2019]). Finally, we point out a recent work by Park and Muandet [2023] towards developing empirical process theory for vector-valued functions.

2.2 Preliminaries

Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} \subseteq \mathbb{R}^K$ be the target space for some $K \in \mathbb{N}$. For a space $\mathcal{Z}$, we let $\mathcal{Z}^\star$ be the set of all finite sequences of elements from $\mathcal{Z}$. Consider a vector-valued function class $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$, where $\mathcal{Y}^\mathcal{X}$ denotes set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. For an unlabeled sample $S_U \in \mathcal{X}^\star$, let $\mathcal{F}_{|S_U}$ denote the projection of $\mathcal{F}$ onto S_U. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner-product on $\mathbb{R}^K$. Define scalar-valued function classes $\mathcal{F}_k = \{x \mapsto \langle f(x), e_k \rangle : f \in \mathcal{F}\}$ for each $k \in [K]$, where $\{e_1, \ldots, e_K\}$ is the standard basis of $\mathbb{R}^K$. Here, each $\mathcal{F}_k \subseteq \mathcal{Y}_k^\mathcal{X}$, where $\mathcal{Y}_k = \{\langle y, e_k \rangle : y \in \mathcal{Y}\}$ denotes the restriction of the target space to its k^{th} component.

For a function $f \in \mathcal{F}$, we use $f_k(x) := \langle f(x), e_k \rangle$ to denote the k^{th} coordinate output of $f(x)$. On the other hand, we use $y^k := \langle y, e_k \rangle$ to denote k^{th} coordinate of $y \in \mathcal{Y}$. Additionally, it is useful to distinguish between the range space $\mathcal{Y}$ and the image of functions $f \in \mathcal{F}$. We define the image of function class $\mathcal{F}$ as $\mathrm{im}(\mathcal{F}) := \cup_{f \in \mathcal{F}} \mathrm{im}(f)$, where $\mathrm{im}(f) = \{f(x) : x \in \mathcal{X}\}$. Finally, we take $[N] := \{1, 2, \ldots, N\}$.

In this work, we only consider bounded, non-negative loss functions $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ that satisfy the identity of the indiscernibles. For the remainder of the paper, we drop the adjectives "bounded" and "non-negative" when referring to loss functions.

Definition 1 (Identity of Indiscernibles)**.** *A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ satisfies identity of indiscernibles whenever $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$.*

Note that if ℓ_1 and ℓ_2 are two losses defined on $\mathcal{Y} \times \mathcal{Y}$ that satisfy the identity of indiscernibles, then $\ell_1(y_1, y_2) = 0$ if and only if $\ell_2(y_1, y_2) = 0$. We also define a notion of approximate subadditivity, although not all the loss functions we consider have this property.

Definition 2 (c-subadditive)**.** *A loss function ℓ is c-subadditive if there exists a constant $c > 0$ that only depends on the loss function ℓ such that $\ell(y_1, y_2) \leq c\,\ell(y_1, y) + \ell(y, y_2)$ for all $y, y_1, y_2 \in \mathcal{Y}$.*

If $|\mathcal{Y}| < \infty$, ℓ being c-subadditive is an immediate consequence of ℓ satisfying the identity of indiscernibles. In fact, the value of c in this case is $\frac{\max_{r \neq t} \ell(r,t)}{\min_{r \neq t} \ell(r,t)}$. To see why this is true, it suffices to only consider the case when $\ell(y_1, y_2) > \ell(y, y_2)$ because the inequality is trivially true otherwise. Since the loss values are distinct, we must have $y \neq y_1$. Using the identity of indiscernible, we obtain $\ell(y_1, y) \geq \min_{r \neq t} \ell(r, t)$, thus implying that $c\,\ell(y_1, y) \geq \max_{r \neq t} \ell(r, t) \geq \ell(y_1, y_2)$. The case when $|\mathcal{Y}| = \infty$ is a bit delicate because $\min_{r \neq t} \ell(r, t)$ may not exist. So, one needs extra structure in the loss function to infer c-subadditivity. For instance, if ℓ is a distance metric, then it is trivially 1-subadditive due to the triangle inequality.

2.2.1 Batch Setting

In the batch setting, we are interested in characterizing the learnability of $\mathcal{F}$ under the classical PAC models: both in the original realizable formulation [Valiant, 1984] and in the agnostic extension [Kearns et al., 1994].

Definition 3 (Agnostic Multioutput Learnability)**.** *A function class $\mathcal{F}$ is agnostic learnable with respect to loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, if there exists a function $m : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^\star \to \mathcal{Y}^{\mathcal{X}}$ with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, running algorithm $\mathcal{A}$ on $n \geq m(\epsilon, \delta)$ iid samples from $\mathcal{D}$ outputs a predictor $g = \mathcal{A}(S)$ such that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(g(x), y)] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f(x), y)] + \epsilon.$$

Note that we do not require the output predictor $\mathcal{A}(S)$ to be in $\mathcal{F}$, but only require $\mathcal{A}(S)$ to compete with the best predictor in $\mathcal{F}$. If we restrict distribution $\mathcal{D}$ to a class such that $\inf_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(x), y)] = 0$, then we get realizable learnability.

The learnability of a function class is generally characterized in terms of the complexity measure of the function class. As stated in the introduction, the VC dimension characterizes

the learnability of binary function classes [Vapnik and Chervonenkis, 1974], and so does the Natarajan dimension for multiclass classification [Ben-David et al., 1995]. In a scalar-valued regression problem, the fat-shattering dimension of the function class characterizes learnability with respect to the absolute-value and squared loss [Bartlett et al., 1996, Alon et al., 1997]. In particular, a real-valued function class $\mathcal{G} \subseteq [0, B]^{\mathcal{X}}$ is learnable if and only if its fat-shattering dimension, denoted as $\mathrm{fat}_\gamma(\mathcal{G})$, is finite for every scale $\gamma > 0$. We extend this characterization to a wide range of loss functions in Lemma 3. Finally, for a real-valued class $\mathcal{G}$, we also use a more general notion of complexity measure called Rademacher complexity, denoted as $\mathfrak{R}_n(\mathcal{G})$, that provides a sufficient condition for learnability [Bartlett and Mendelson, 2003]. Precise definitions of all these complexity measures are provided in Appendix A.1.1.

One recurring theme in this work is to first construct a realizable multioutput learner and then convert it into an agnostic multioutput learner. It is well known that realizable learnability and agnostic learnability are equivalent for multiclass classification problems with respect to 0-1 loss, (see Ben-David et al. [1995], [Shalev-Shwartz and Ben-David, 2014, Theorem 6.7]). Lemma 1, which is an immediate consequence of [Hopkins et al., 2022, Theorem 18], extends this equivalence between realizable and agnostic learning to general loss functions and target spaces.

Lemma 1 (Hopkins et al. [2022])**.** *Consider a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $|im(\mathcal{F})| < \infty$ and a general loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ that is c-subadditive. Then, $\mathcal{F}$ is realizable PAC learnable with respect to ℓ if and only if $\mathcal{F}$ is agnostic PAC learnable with respect to ℓ.*

The result of [Hopkins et al., 2022, Theorem 18] is stated for the case when $|\mathcal{Y}| < \infty$. However, in the regression setting, we need a slightly general version of their result to handle α-discretized function classes $\mathcal{F}^\alpha \subseteq \mathcal{Y}^{\mathcal{X}}$ (see Proof of Theorem 5) where $|\mathrm{im}(\mathcal{F}^\alpha)| < \infty$ but $|\mathcal{Y}| = |\,[0,1]^K| = \infty$. Nevertheless, the proof of Lemma 1 requires only a minor modification to that of [Hopkins et al., 2022, Theorem 18]. Given the central role of this result in our characterization, we provide full proof of Lemma 1 in Section 2.3.1.2. Finally, we note that agnostic-to-realizable conversions in the batch setting are also possible via boosting and compression-based arguments [Montasser et al., 2019, Attias and Hanneke, 2023]. We use the conversion of Hopkins et al. [2022] due to its generality and simplicity.

2.2.2 Online Setting

In the online setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, an adversary selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner makes a (potentially randomized) prediction $\hat{y}_t \in \mathcal{Y}$. Finally,

the adversary reveals the true label y_t, and the learner suffers the loss $\ell(y_t, \hat{y}_t)$, where ℓ is some pre-specified loss function. Given a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{y}_t$ such that its cumulative loss is close to the best possible cumulative loss over functions in $\mathcal{F}$. A function class is online learnable if there exists an algorithm such that for any sequence of labeled examples $(x_1, y_1), ..., (x_T, y_T)$, the difference in cumulative loss between its predictions and the predictions of the best possible function in $\mathcal{F}$ is small.

Definition 4 (Online Multioutput Learnability)**.** *A multioutput function class $\mathcal{F}$ is online learnable with respect to loss ℓ, if there exists an (potentially randomized) algorithm $\mathcal{A}$ such that for any adaptively chosen sequence of labelled examples $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$, the algorithm outputs $\mathcal{A}(x_t) \in \mathcal{Y}$ at every iteration $t \in [T]$ such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \leq R(T)$$

where the expectation is taken with respect to the randomness of $\mathcal{A}$ and that of the possibly adaptive adversary, and $R(T) : \mathbb{N} \to \mathbb{R}^+$ is the additive regret: a non-decreasing, sub-linear function of T.

If it is guaranteed that the learner always observes a sequence of examples labeled by some function $f \in \mathcal{F}$, then we say we are in the *realizable* setting. On the other hand, if the true label y_t is not revealed to the learner in each round $t \in [T]$ and the adversary only reveals the learner's *loss* $\ell(\mathcal{A}(x_t), y_t)$ then we say we are in the *bandit* setting.

The online learnability of scalar-valued function classes $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ has been characterized. For example, when $\mathcal{Y}$ is finite (i.e. $\mathcal{Y} = [K]$ for some $K \in \mathbb{N}$), the Multiclass Littlestone Dimension (MCLdim) of $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ characterizes online learnability with respect to the 0-1 loss. A function class $\mathcal{H}$ is online learnable with respect to the 0-1 loss if and only if MCLdim($\mathcal{H}$) is finite [Daniely et al., 2011]. Moreover, MCLdim($\mathcal{H}$) tightly captures the best achievable regret in both the realizable and agnostic settings [Daniely et al., 2011]. When the label space $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, the *sequential* fat-shattering dimension of a real-valued function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, denoted $\text{fat}_\gamma^{\text{seq}}(\mathcal{H})$, characterizes the online learnability of $\mathcal{H}$ with respect to the absolute value loss [Rakhlin et al., 2015a]. Unlike the Littlestone dimension, note that the sequential fat-shattering dimension is a scale-sensitive dimension. That is, $\text{fat}_\gamma^{\text{seq}}(\mathcal{H})$ is defined at every scale $\gamma > 0$. Accordingly, a real-valued function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is learnable with respect to the absolute value loss if and only if its sequential fat-shattering dimension is finite at *every* scale $\gamma > 0$ [Rakhlin et al., 2015a]. We extend this characterization to a wide range of loss functions in Lemma 6. Beyond scalar-valued learnability, for any label space $\mathcal{Y}$, function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the *sequential* Rademacher complexity of the loss class $\ell \circ \mathcal{H}$, denoted $\mathfrak{R}_T^{\text{seq}}(\ell \circ \mathcal{H})$, is a useful

tool for providing sufficient conditions for online learnability [Rakhlin et al., 2015a]. See Appendix A.1.2 for complete definitions.

Like the batch setting, a key technique we use to prove online learnability is to first construct a realizable online learner and then convert it into an agnostic online learner. However, unlike the batch setting, there is no known generic algorithm that converts a (potentially randomized) realizable online learner into an agnostic online learner. Thus, one of the contributions of this work is Theorem 6, informally stated below, which provides an online analog of the realizable-to-agnostic conversion from Hopkins et al. [2022].

Theorem. *(Informal) Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multioutput function class such that $|im(\mathcal{F})| < \infty$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any bounded, c-subadditive loss function. If $\mathcal{F}$ is online learnable with respect to ℓ in the realizable setting, then $\mathcal{F}$ is online learnable with respect to ℓ in the agnostic setting.*

2.3 Batch Multioutput Learnability

2.3.1 Batch Multilabel Classification

In this section, we study the learnability of batch multilabel classification. Accordingly, let $\mathcal{Y} = \{-1, 1\}^K$. First, we consider the learnability of a natural decomposable loss. Then, we extend the result to more general non-decomposable losses that satisfy the identity of indiscernible. We want to point out that a multilabel classification with K labels can be viewed as a multiclass classification with 2^K labels. With this viewpoint, the Natarajan dimension of $\mathcal{F}$ continues to characterize the batch multilabel learnability for any loss satisfying the identity of indiscernibles (see [Ben-David et al., 1995, Section 4]). For the sake of completeness, we also provide proof of this characterization in Appendix A.2. However, it is more natural to view a multilabel classification as K different binary classification problems as opposed to a multi-class classification problem with 2^K labels. Exploiting this natural decomposability of a multilabel function class, we relate the learnability of $\mathcal{F}$ to the learnability of each component $\mathcal{F}_k$.

2.3.1.1 Characterizing Batch Learnability for the Hamming Loss

A canonical and natural loss function for multilabel classification is the Hamming loss, defined as $\ell_H(f(x), y) := \sum_{i=1}^{K} \mathbb{1}\{f_i(x) \neq y^i\}$, where $f(x) = (f_1(x), \ldots, f_K(x))$ and $y = (y^1, \ldots, y^K)$. The following result establishes an equivalence between the learnability of $\mathcal{F}$ with respect to Hamming loss and the learnability of each $\mathcal{F}_k$ with respect to 0-1 loss.

Theorem 1. *A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ_H if and only if each of $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic PAC learnable with respect to the* 0-1 *loss.*

Proof. We first prove that learnability of each component is sufficient followed by the proof of necessity.

Part 1: Sufficiency. Here our goal is to prove that the agnostic PAC learnability of each $\mathcal{F}_k$ is sufficient for agnostic PAC learnability of $\mathcal{F}$. Our proof is constructive: given oracle access to agnostic PAC learners $\mathcal{A}_k$ for each $\mathcal{F}_k$ with respect to 0-1 loss, we construct an agnostic PAC learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. Let $\mathcal{D}$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ be iid samples from distribution $\mathcal{D}$. Denote $\mathcal{D}_k$ to be the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X} \times \mathcal{Y}_k$. Then, for all $k \in [K]$, the marginal samples $S_k = \{(x_i, y_i^k)\}_{i=1}^n$ with scalar-valued targets are iid samples from $\mathcal{D}_k$. For each $k \in [K]$, define $h_k = \mathcal{A}_k(S_k)$ to be the hypothesis returned by algorithm $\mathcal{A}_k$ when trained on S_k.

Let $m_k(\epsilon, \delta)$ denote the sample complexity of $\mathcal{A}_k$. Since $\mathcal{A}_k$ is an agnostic PAC learner for $\mathcal{F}_k$, we have that for $n \geq \max_k m_k(\frac{\epsilon}{K}, \frac{\delta}{K})$, with probability at least $1 - \delta/K$ over samples $S_k \sim \mathcal{D}_k^n$,

$$\mathop{\mathbb{E}}_{\mathcal{D}_k}\left[\mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \inf_{f_k \in \mathcal{F}_k} \mathop{\mathbb{E}}_{\mathcal{D}_k}\left[\mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \frac{\epsilon}{K}.$$

Summing these risk bounds over all coordinates k and using union bounds over probabilities, we get that with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^n$, we obtain $\sum_{k=1}^K \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{\mathcal{D}_k}\left[\mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \epsilon$. Now using the fact that $\mathcal{F} \subseteq \mathcal{F}_1 \times \ldots \times \mathcal{F}_K$ followed by the linearity of expectation gives

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\sum_{k=1}^K \mathbb{1}\left\{h_k(x) \neq y^k\right\}\right] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}\left[\sum_{k=1}^K \mathbb{1}\left\{f_k(x) \neq y^k\right\}\right] + \epsilon.$$

This completes our proof as it shows that the learning rule that concatenates the predictors returned by each $\mathcal{A}_k$ on the marginalized samples S_k is an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ_H with sample complexity at most $\max_k m_k(\epsilon/K, \delta/K)$.

Part 2: Necessity. Next, we show that if $\mathcal{F}$ is learnable with respect to ℓ_H, then each $\mathcal{F}_k$ is PAC learnable with respect to the 0-1 loss. Our proof is again based on reduction: given oracle access to agnostic PAC learner $\mathcal{A}$ for $\mathcal{F}$, we construct an agnostic PAC learner $\mathcal{A}_1$ for $\mathcal{F}_1$. A similar construction can be used for all other $\mathcal{F}_k$'s.

Let $\mathcal{D}_1$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}_1$ and $S = \{(x_i, y_i^1)\}_{i=1}^n$ be iid samples from $\mathcal{D}_1$. In order to use the algorithm $\mathcal{A}$, we first augment the samples S to create samples with K-variate target. Define an augmented sample $\widetilde{S} = \{(x_i, (y_i^1, \ldots, y_i^K))\}_{i=1}^n$ such that $y_{ik} \sim \{-1, 1\}$ each with probability $1/2$ for all $i \in [n]$ and $k \in \{2, \ldots, K\}$. Next, we run $\mathcal{A}$ on $\widetilde{S}$ and obtain the hypothesis $h = (h_1, \ldots, h_K) = \mathcal{A}(\widetilde{S})$. We now show that h_1 obtains

agnostic PAC bounds.

Consider a distribution $\widetilde{\mathcal{D}}$ on $\mathcal{X}\times\mathcal{Y}$ such that a sample $(x,(y^1,\ldots,y^K))$ from $\widetilde{\mathcal{D}}$ is obtained by first sampling $(x,y^1)\sim\mathcal{D}_1$ and appending y^k's sampled independently from uniform distribution on $\{-1,1\}$ for each $k\in\{2,\ldots,K\}$. Let $m(\epsilon,\delta,K)$ denote the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is an agnostic PAC learner for $\mathcal{F}$, for $n\geq m(\epsilon,\delta,K)$, with probability at least $1-\delta$, we have

$$\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}\left[\sum_{k=1}^{K}\mathbb{1}\left\{h_k(x)\neq y^k\right\}\right]\leq\inf_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}\left[\sum_{k=1}^{K}\mathbb{1}\left\{f_k(x)\neq y^k\right\}\right]+\epsilon.$$

For $k\geq 2$, since the target is chosen uniformly at random from $\{-1,1\}$, the 0-1 risk of any predictor is $1/2$. Therefore, the expression above can be written as

$$\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\mathbb{1}\left\{h_1(x)\neq y^1\right\}\right]+\sum_{k=2}^{K}1/2\leq\inf_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\mathbb{1}\left\{f_1(x)\neq y^1\right\}\right]+\sum_{k=2}^{K}1/2\right)+\epsilon,$$

which reduces to $\mathbb{E}_{\mathcal{D}_1}[\mathbb{1}\{h_1(x)\neq y^1\}]\leq\inf_{f_1\in\mathcal{F}_1}\mathbb{E}_{\mathcal{D}_1}[\mathbb{1}\{f_1(x)\neq y^1\}]+\epsilon$. Therefore, $\mathcal{F}_1$ is agnostic PAC learnable with respect to 0-1 loss with sample complexity at most $m(\epsilon,\delta,K)$. ■

2.3.1.2 Characterizing Batch Learnability for General Losses

In this section, we characterize the learnability for general multilabel losses. Our main technical tool in characterizing the learnability for general loss functions is the equivalence between realizable and agnostic learning guaranteed by Lemma 1. Thus, we first provide the proof of that lemma before we proceed further.

Proof. (of Lemma 1) Note that agnostic learnability implies realizable learnability by definition. So, it suffices to show that realizable learnability of $\mathcal{F}$ with respect to ℓ implies agnostic learnability. Our proof here is constructive. That is, given a realizable algorithm $\mathcal{A}$ for $\mathcal{F}$, we provide an algorithm, stated as Algorithm 1, that constructs an agnostic learner for $\mathcal{F}$.

Let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X}\times\mathcal{Y}$. Define

$$f^\star:=\inf_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f(x),y)]$$

to be the optimal predictor in $\mathcal{F}$. Now consider a predictor $g=\mathcal{A}(S_U,f^\star(S_U))\in C(S_U)$ returned by $\mathcal{A}$ when trained on samples S_U labeled by $f^\star$. Note that g exists in $C(S_U)$ because we consider all possible labelings of S_U by $\mathcal{F}$ and there must be a sample labeled

Algorithm 1 Agnostic learner for $\mathcal{F}$ with respect to ℓ

Require: Realizable learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ, unlabeled samples $S_U \sim \mathcal{D}_{\mathcal{X}}$, and different labeled samples $S_L \sim \mathcal{D}$ independent from S_U

1: Run $\mathcal{A}$ over all possible labelings of S_U by $\mathcal{F}$ to construct a concept class

$$C(S_U) := \left\{ \mathcal{A}\Big(S_U, f(S_U)\Big) \mid f \in \mathcal{F}_{|S_U} \right\}.$$

2: Return $\hat{g} \in C(S_U)$ with the lowest empirical error over S_L with respect to ℓ.

by $f^\star$ as well. Let $m_{\mathcal{A}}(\epsilon, \delta, K)$ be the sample complexity of the algorithm $\mathcal{A}$. Since $\mathcal{A}$ is a realizable learner for $\mathcal{F}$, for $|S_U| \geq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right)$ with probability $1 - \delta/2$ over sample $S_U \sim \mathcal{D}_{\mathcal{X}}$, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}} [\ell(g(x), f^\star(x))] \leq \frac{\epsilon}{2c},$$

where $\mathcal{D}_{\mathcal{X}}$ is the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X}$ and c is the subaddtivity constant of ℓ. Since the loss function is c-subadditive, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have $\ell(g(x), y) \leq \ell(f^\star(x), y) + c\,\ell(g(x), f^\star(x))$ pointwise. Taking expectation with respect to $(x, y) \sim \mathcal{D}$, we obtain

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(g(x), y)] \leq c \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}} [\ell(g(x), f^\star(x))] + \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f^\star(x), y)] \leq \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{\epsilon}{2},$$

where the inequality holds with probability $\geq 1 - \delta/2$. Thus, we have shown that there exists a predictor $g \in C(S_U)$ that achieves agnostic PAC bounds for $\mathcal{F}$ with respect to ℓ. Let $\ell(\cdot, \cdot) \leq b$ be the upper bound on ℓ. Recall that by Hoeffding's inequality and union bound, with probability at least $1 - \delta/2$ over sample $S_L \sim \mathcal{D}$, the empirical risk of every hypothesis in $C(S_U)$ on a sample of size $\geq \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its population risk. So, if $|S_L| \geq \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta}$, then with probability at least $1 - \delta/2$ over sample $S_L \sim \mathcal{D}$, we have

$$\frac{1}{|S_L|} \sum_{(x,y) \in S_L} \ell(g(x), y) \leq \mathop{\mathbb{E}}_{\mathcal{D}} [\ell(g(x), y)] + \frac{\epsilon}{4} \leq \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{3\epsilon}{4}.$$

Next, consider the predictor $\hat{g}$ returned by Algorithm 1. Since it is an empirical risk minimizer, its empirical risk can be at most the empirical risk of g. Given that the population risk of $\hat{g}$ can be at most $\epsilon/4$ away from its empirical risk, we have that

$$\mathop{\mathbb{E}}_{\mathcal{D}} [\ell(\hat{g}(x), y)] \leq \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f^\star(x), y)] + \frac{3\epsilon}{4} + \frac{\epsilon}{4} \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(f(x), y)] + \epsilon,$$

where the second inequality above uses the definition of $f^\star$. Note that this inequality holds with probability at least $1 - \delta$, where the probability is taken over both samples S_U and S_L.

Thus, we have shown that Algorithm 1 is an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ.

We now upper bound the sample complexity of Algorithm 1, denoted $m(\epsilon, \delta, K)$ hereinafter. Note that $m_{\mathcal{A}}(\epsilon, \delta, K)$ is at most the number of unlabeled samples required for the realizable algorithm $\mathcal{A}$ to succeed plus the number of labeled samples for the ERM step to succeed. Thus,

$$\begin{aligned} m(\epsilon, \delta, K) &\leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2} \log \frac{4|C(S_U)|}{\delta} \\ &\leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2}\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) \log\left(|\mathrm{im}(\mathcal{F})|\right) + \log \frac{4}{\delta}\right), \end{aligned}$$

where the second inequality follows due to $|C(S_U)| \leq |\mathrm{im}(\mathcal{F})|^{|S_U|}$ and we need $|S_U|$ to be of size $m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right)$. ■

With Lemma 1 in hand, we can now relate the learnability of $\mathcal{H}$ with respect to any ℓ satisfying identity of indiscernibles to the learnability of $\mathcal{H}$ with respect to ℓ_H. To that end, we prove a result establishing the equivalence of learnability between any two loss functions satisfying the identity of indiscernibles.

Lemma 2. *Let ℓ and ℓ' be any two loss functions satisfying the identity of indiscernibles. Then, a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ if and only if $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ'.*

The key idea behind the proof of Lemma 2 is to use a realizable learner for ℓ to construct a realizable learner for ℓ'. This is possible because $\ell'(y_1, y_2) = 0$ if and only if $\ell(y_1, y_2) = 0$ for any $y_1, y_2 \in \mathcal{Y}$. Given such realizable learner for ℓ', Lemma 1 guarantees the existence of an agnostic learner for ℓ'.

Proof. Since ℓ and ℓ' are arbitrary, it suffices to prove only one direction. So, let us assume that $\mathcal{F}$ is learnable with respect to ℓ. We will now show that $\mathcal{F}$ is learnable with respect to ℓ' as well. First, we show this for any realizable distribution $\mathcal{D}$ with respect to ℓ'. Since, for any $y_1, y_2 \in \mathcal{Y}$, we have $\ell'(y_1, y_2) = 0$ if and only if $\ell(y_1, y_2) = 0$, the distribution $\mathcal{D}$ is also realizable with respect to ℓ. Furthermore, as there are at most 2^{2K} distinct possible inputs to $\ell'(\cdot, \cdot)$, the loss function can only take a finite number of values. So, we can always find universal constants $a > 0$ and $b > 0$ (that only depends on ℓ and ℓ') such that $a\ell \leq \ell' \leq b\ell$. Given that $\mathcal{F}$ is learnable with respect to ℓ, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for $S \sim \mathcal{D}^n$, such that $n = m_{\mathcal{A}}(\frac{\epsilon}{b}, \delta, K)$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, we

have $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq \frac{\epsilon}{b}$. This inequality upon using the fact that $\ell'(h(x), y) \leq b\ell(h(x), y)$ pointwise reduces to $\mathbb{E}_{\mathcal{D}}[\ell'(h(x), y)] \leq \epsilon$. Therefore, any realizable learner $\mathcal{A}$ for ℓ is also a realizable learner for ℓ'. Finally, as ℓ' satisfies the identity of indiscernibles and thus c-subadditivity, Lemma 1 guarantees the existence of agnostic PAC learner $\mathcal{B}$ for $\mathcal{F}$ with respect to ℓ'. In particular, one such agnostic PAC learner $\mathcal{B}$ is Algorithm 1 that has sample complexity

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}\left(\frac{\epsilon}{2bc}, \frac{\delta}{2}, K\right) + \frac{b^2}{\epsilon^2}\, O\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2bc}, \frac{\delta}{2}, K\right) K + \log\frac{1}{\delta}\right),$$

where $c > 1$ is the subadditivity constant of ℓ' and $m_{\mathcal{A}}(\cdot, \cdot, K)$ is the sample complexity of any realizable algorithm $\mathcal{A}$.

■

An immediate consequence of Lemma 2 is that learnability with respect to any loss ℓ satisfying the identity of indiscernibles is equivalent to learnability with respect to the Hamming loss ℓ_H. Thus, given Theorem 1, we can deduce the following result.

Theorem 2. *Let ℓ be any multilabel loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic PAC learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic PAC learnable with respect to the* 0-1 *loss.*

Remark. Since the learnability of a binary function class with respect to the 0-1 loss is characterized by its VC dimension [Shalev-Shwartz and Ben-David, 2014, Theorem 6.7], Theorem 2 implies that $\mathcal{F}$ is learnable with respect to ℓ satisfying the identity of indiscernibles if and only if $\mathrm{VC}(\mathcal{F}_k) < \infty$ for each $k \in [K]$.

2.3.2 Batch Multioutput Regression

In this section, we consider the case when $\mathcal{Y} = [0, 1]^K \subseteq \mathbb{R}^K$ for $K \in \mathbb{N}$. For bounded targets (with a known bound), this target space is without loss of generality because one can always normalize each $\mathcal{Y}_k$ to [0,1] by subtracting the lower bound and dividing by the upper bound of $\mathcal{Y}_k$. As usual, we consider an arbitrary multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$. Following our outline in classification, we first study the learnability of $\mathcal{F}$ under decomposable losses and then non-decomposable losses.

2.3.2.1 Characterizing Learnability for Decomposable Losses

A canonical loss for the scalar-valued regression is the absolute value metric, $d_1(f_k(x), y^k) := |f_k(x) - y^k|$. Analogously, we define $d_p(f_k(x), y^k) := |f_k(x) - y^k|^p$ for $p > 1$ are other natural

scalar-valued losses. For multioutput regression, we consider decomposable losses that are natural multivariate extensions of the d_1 metric. In particular, we consider loss functions with the following properties.

Assumption 1. *The loss can be written as* $\ell(f(x), y) = \sum_{k=1}^{K} \psi_k \circ d_1(f_k(x), y^k)$ *where for each* $k \in [K]$*, the function* $\psi_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is L-Lipschitz and satisfies* $\psi_k(0) = 0$*.*

Here, $\psi_k \circ d_1$ is a composition function defined as $\psi_k \circ d_1(f_k(x), y^k) := \psi_k\Big(d_1(f_k(x), y^k)\Big)$. Note that $\psi_k \circ d_1$ is a large family of loss functions that effectively contains all natural decomposable multioutput regression losses. For instance, taking $\psi_k(z) = |z|^p$ for $p \geq 1$ gives ℓ_p norms raised to their p-th power. Considering $\psi_k(z) = |z|^2/2\, \mathbb{1}[|z| \leq \delta] + \delta(|z| - \delta/2)\mathbb{1}[|z| > \delta]$ for some $1 > \delta > 0$ yields multivariate extension of Huber loss used for robust regression. One may also construct a multioutput loss by considering different scalar-valued losses for each coordinate output. Next, we establish an equivalence between the learnability of $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to ℓ and the learnability of each $\mathcal{F}_k$ with respect to the loss $\psi_k \circ d_1$.

Theorem 3. *Let* ℓ *be any loss function that satisfies Assumption 1. Then, a function class* $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ *is agnostic learnable with respect to* ℓ *if and only if each of* $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ *is agnostic learnable with respect to* $\psi_k \circ d_1$*.*

Proof. The proof of the sufficiency direction is similar to that of Theorem 1, so we defer it to Appendix A.3.1. We now focus on the necessity direction. To that end, we show that if $\mathcal{F}$ is agnostic learnable with respect to ℓ, then each $\mathcal{F}_k$ is agnostic learnable with respect to $\psi_k \circ d_1$. In particular, given oracle access to agnostic learner $\mathcal{A}$ for $\mathcal{F}$, we construct agnostic learner $\mathcal{A}_1$ for $\mathcal{F}_1$. By symmetry, a similar reduction can then be used to construct an agnostic learner for each component $\mathcal{F}_k$.

Since we are given a sample with a single, univariate target, the main problem is to find the right way to augment samples to a K-variate target. In the proof of Theorem 1, we showed that randomly choosing $y_{ik} \sim \text{Uniform}(\{-1, 1\})$ for $k \geq 2$ results in all predictors having a constant $1/2$ risk–leaving only the risk of the first component on both sides. Unfortunately, in regression under general losses, no single augmentation works for every distribution on $\mathcal{X}$. Thus, we augment the samples by considering all possible behaviors of $(\mathcal{F}_2, \ldots, \mathcal{F}_K)$ on the sample. Since the function class maps to a potentially uncountably infinite space, we first discretize each component of the function class and consider all possible labelings over the discretized space. Fix $1 > \alpha > 0$. For $k \geq 2$, define the discretization

$$f_k^{\alpha}(x) = \left\lfloor \frac{f(x)}{\alpha} \right\rfloor \alpha \tag{2.1}$$

for every $f_k \in \mathcal{F}_k$ and the discretized component class $\mathcal{F}_k^\alpha = \{f_k^\alpha | f_k \in \mathcal{F}_k\}$. Note that a function in $\mathcal{F}_k$ maps to $\{0, \alpha, 2\alpha, \ldots, \lfloor 1/\alpha \rfloor \alpha\}$ and the size of the range of the discretized function class $\mathcal{F}_k^\alpha$ is $1 + \lfloor 1/\alpha \rfloor \leq (\alpha+1)/\alpha \leq 2/\alpha$. For the convenience of exposition, let us define $\mathcal{F}_{2:K}^\alpha$ to be $\mathcal{F}^\alpha$ without the first component, and we denote $f_{2:K}^\alpha$ to be an element of $\mathcal{F}_{2:K}^\alpha$.

Algorithm 2 Agnostic learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$

Require: Agnostic learner $\mathcal{A}$ for $\mathcal{F}$, samples $S = (x_{1:n}, y_{1:n}^1) \sim \mathcal{D}_1^n$, and another independent samples $\widetilde{S}$ from $\mathcal{D}_1$

1: Define $S_{\text{aug}} = \{(x_{1:n}, y_{1:n}^1, f_{2:K}^\alpha(x_{1:n})) \mid f_{2:K}^\alpha \in \mathcal{F}_{2:K}^\alpha\}$, all possible augmentations of S by $\mathcal{F}_{2:K}^\alpha$.

2: Run $\mathcal{A}$ over all possible augmentations to get

$$C(S) := \left\{ \mathcal{A}\Big(S_a\Big) \mid S_a \in S_{\text{aug}} \right\}.$$

3: Define $C_1(S) = \{g_1 \mid (g_1, \ldots, g_k) \in C(S)\}$, a restriction of $C(S)$ to its first coordinate output.

4: Return $\hat{g}_1$, the predictor in $C_1(S)$ with the lowest empirical error over $\widetilde{S}$ with respect to $\psi_1 \circ d_1$.

We now show that Algorithm 2 is an agnostic learner for $\mathcal{F}_1$. First, let us define

$$f_1^\star := \arg\min_{f_1 \in \mathcal{F}_1} \mathop{\mathbb{E}}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)],$$

to be optimal predictor in $\mathcal{F}_1$ with respect to $\mathcal{D}_1$. By definition of $\mathcal{F}_1$, there must exist $f_{2:K}^\star \in \mathcal{F}_{2:K}$ such that $(f_1^\star, f_{2:K}^\star) \in \mathcal{F}$. We note that $f_k^\star$ need not be optimal predictors in $\mathcal{F}_k$ for $k \geq 2$, but we use the $\star$ notation just to associate these component functions with the first component function $f_1^\star$. Define $f_{2:K}^{\star,\alpha} \in \mathcal{F}_{2:K}^\alpha$ to be the corresponding discretization of $f_{2:K}^\star$. At a high level, the key idea of this proof is to show that the algorithm $\mathcal{A}$ when run on the sample $(x_{1:n}, y_{1:n}^1, f_{2:K}^{\star,\alpha}(x_{1:n})$ produces a predictor $g = \mathcal{A}(x_{1:n}, y_{1:n}^1, f_{2:K}^{\star,\alpha}(x_{1:n}))$ such that its restriction g_1 is a valid agnostic learner for $\mathcal{F}_1$. Although one such augmentation is enough to produce an agnostic learner for $\mathcal{F}_1$, all possible augmentations are required in step 2 of Algorithm 2 because $f_1^\star$ is not known to the learner apriori. We now make this argument precise.

Suppose $g = \mathcal{A}((x_{1:n}, y_{1:n}^1, f_{2:K}^{\star,\alpha}(x_{1:n}))$ is the predictor obtained by running $\mathcal{A}$ on the sample augmented by $f_{2:K}^{\star,\alpha}$. Note that $g \in C(S)$ by definition. Let $m_{\mathcal{A}}(\epsilon, \delta, K)$ be the sample complexity of $\mathcal{A}$. Since $\mathcal{A}$ is an agnostic learner for $\mathcal{F}$ with respect to ℓ, we have that for

$n \geq m_{\mathcal{A}}(\epsilon/4, \delta/2, K)$, with probability at least $1 - \delta/2$,

$$\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] + \sum_{k=2}^{K} \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}\left[\psi_k \circ d_1(g_k(x), f_k^{\star,\alpha}(x))\right] \\ \leq \inf_{f \in \mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(f_1(x), y^1)\right] + \sum_{k=2}^{K} \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}\left[\psi_k \circ d_1(f_k(x), f_k^{\star,\alpha}(x))\right]\right) + \frac{\epsilon}{4}$$

Note that the quantity on the left is trivially lower bounded by the risk of the first component. To handle the right-hand side, we first note that the optimal risk is trivially upper bounded by the risk of $(f_1^\star, f_{2:K}^\star)$, yielding

$$\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] \leq \mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(f_1^\star(x), y^1)\right] + \sum_{k=2}^{K} \mathop{\mathbb{E}}_{\mathcal{D}_{\mathcal{X}}}\left[\psi_k \circ d_1(f_k^\star(x), f_k^{\star,\alpha}(x))\right] + \frac{\epsilon}{4}.$$

Next, using the L-Lipschitzness of ψ_k and the fact that $\psi_k(0) = 0$ implies $\psi_k \circ d_1(f_k^\star(x), f_k^{\star,\alpha}(x)) \leq L\, d_1(f_k^\star(x), f_k^{\star,\alpha}(x)) \leq L\alpha$ for all $k \geq 2$. Thus, picking $\alpha = \frac{\epsilon}{4LK}$ and using the definition of $f_1^\star$, we obtain

$$\mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] \leq \inf_{f_1 \in \mathcal{F}_1} \mathop{\mathbb{E}}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{\epsilon}{2}.$$

Therefore, we have shown the existence of one predictor $g \in C(S)$ such that its restriction to the first component, g_1, obtains the agnostic bound. Note that since ψ_1 is L-Lipschitz and satisfies $\psi_1(0) = 0$, we obtain that $\psi_1 \circ d_1(\cdot, \cdot) \leq L$. The upper bound also uses the fact that $|f_1(x) - y^1| \leq 1$. Now recall that by Hoeffding's Inequality and union bound, with probability at least $1 - \delta/2$, the empirical risk of every hypothesis in $C_1(S)$ on a sample of size $\geq \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $|\widetilde{S}| \geq \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta}$, then with probability at least $1 - \delta/2$, the empirical risk of the predictor g_1 is

$$\frac{1}{|\widetilde{S}|} \sum_{(x,y^1) \in \widetilde{S}} \psi_1 \circ d_1(g_1(x), y^1) \leq \mathop{\mathbb{E}}_{\mathcal{D}_1}\left[\psi_1 \circ d_1(g_1(x), y^1)\right] + \frac{\epsilon}{4} \leq \inf_{f_1 \in \mathcal{F}_1} \mathop{\mathbb{E}}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{3\epsilon}{4}.$$

Since $\hat{g}_1$, the output of Algorithm 2 is the ERM on $\widetilde{S}$ over $C_1(S)$, its empirical risk can be at most the empirical risk of g_1, which is at most $\inf_{f_1 \in \mathcal{F}_1} \mathbb{E}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \frac{3\epsilon}{4}$. Given that the population risk of $\hat{g}_1$ is at most $\epsilon/4$ away from its empirical risk, we can conclude that the population risk of $\hat{g}_1$ is

$$\mathop{\mathbb{E}}_{\mathcal{D}_1}[\psi_1 \circ d_1(\hat{g}_1(x), y^1)] \leq \inf_{f_1 \in \mathcal{F}_1} \mathop{\mathbb{E}}_{\mathcal{D}_1}[\psi_1 \circ d_1(f_1(x), y^1)] + \epsilon.$$

Applying union bounds, the entire process, algorithm $\mathcal{A}$ on the dataset augmented by $f_{2:K}^{\star,\alpha}$ and ERM in step 4, succeeds with probability $1-\delta$. The sample complexity of Algorithm 2 is the sample complexity of Algorithm $\mathcal{A}$ and the sample complexity of ERM in step 4, which is

$$\begin{aligned}
&\leq m_{\mathcal{A}}(\epsilon/4, \delta/2, K) + \frac{8L^2}{\epsilon^2} \log \frac{4|C_1(S)|}{\delta} \\
&\leq m_{\mathcal{A}}(\epsilon/4, \delta/2, K) + \frac{8KL^2}{\epsilon^2} \left(m_{\mathcal{A}}(\epsilon/4, \delta/2, K) \, \log \left(\frac{4K\,L}{\epsilon} \right) + \log \frac{4}{\delta} \right),
\end{aligned}$$

where the second inequality follows due to $|C_1(S)| \leq (2/\alpha)^{m_{\mathcal{A}}(\epsilon/4,\delta/2,K)\,K}$ is the required size of $C_1(S)$ and our choice of $\alpha = \epsilon/(4KL)$. This completes the proof as we have shown that Algorithm 2 is an agnostic learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$. ■

2.3.2.2 A More General Characterization of Learnability for Decomposable Losses

Unlike Theorems 1 and 2 in classification setting where we connected the learnability of $\mathcal{F}$ to the learnability of $\mathcal{F}_k$'s with respect to 0-1 loss, Theorem 3 relates the learnability of $\mathcal{F}$ to the learnability of $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ instead of the more canonical loss d_1. In this section, we complete that final step to characterizing learnability in terms of d_1 with an additional assumption on ψ_k.

Assumption 2. *For all $k \in [K]$, the function $\psi_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is monotonic.*

Under these assumptions, Theorem 4 provides a more general characterization than Theorem 3.

Theorem 4. *Let ℓ be any loss function that satisfies Assumptions 1 and 2. Then, a multioutput function class $\mathcal{F} \subseteq ([0,1]^K)^{\mathcal{X}}$ is agnostic learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq [0,1]^{\mathcal{X}}$ is agnostic learnable with respect to d_1.*

Since the fat-shattering dimension of a real-valued function class characterizes its learnability with respect to d_1 loss, Theorem 4 implies that a multioutput function class $\mathcal{F}$ is learnable with respect to ℓ satisfying Assumptions 1 and 2 if and only if $\text{fat}_\gamma(\mathcal{F}) < \infty$ for every fixed scale $\gamma > 0$. Theorem 4 is an immediate consequence of Theorem 3 and the following lemma, the proof of which is deferred to Appendix A.3.2.

Lemma 3. *Let $\mathcal{Y} = [0,1]$ be the label space and $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be any Lipschitz and monotonic function that satisfies $\psi(0) = 0$. A scalar-valued function class $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ is agnostic learnable with respect to $\psi \circ d_1$ if and only if $\mathcal{G}$ is agnostic learnable with respect to d_1.*

The part of the lemma showing that d_1 learnability implies $\psi \circ d_1$ is trivial using the Rademacher-based argument and Talagrand's contraction lemma. However, proving $\psi \circ d_1$ learnability implies d_1 learnability is non-trivial. The case $\psi(z) = |z|^2$ is considered in [Anthony and Bartlett, 1999, Theorem 19.5], but their proof requires a mismatch between the label space and the prediction space, namely $\mathcal{Y} = [-1, 2]$ but $\mathcal{G} \subseteq [0, 1]^{\mathcal{X}}$. In this work, we improve their result by showing the equivalence between $\psi \circ d_1$ learnability and d_1 learnability without requiring extended label space.

An application of Lemma 3 is that the learnability of a real-valued function class $\mathcal{G}$ with respect to losses d_1 and d_p are equivalent for any $p > 1$.

2.3.2.3 Characterizing Learnability for Non-Decomposable Losses

Next, we study the learnability of function class $\mathcal{F}$ with respect to non-decomposable losses. In the regression setting, the natural non-decomposable loss to consider is ℓ_p norm, which is defined as $\ell_p(f(x), y) := \left(\sum_{k=1} |f_k(x) - y^k|^p\right)^{1/p}$ for $1 \leq p < \infty$. For $p = \infty$, the p-norm is defined as $\ell_\infty(f(x), y) := \max_k |f_k(x) - y^k|$. One might be interested in ℓ_p norms instead of their decomposable counterparts ℓ_p^p losses discussed in the previous section mainly for robustness to outliers. The following result characterizes the agnostic learnability of $\mathcal{F}$ with respect to ℓ_p norms.

Theorem 5. *Fix $p \geq 1$. The function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic learnable with respect to ℓ_p norm if and only if each of $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is agnostic learnable with respect to the absolute value loss, d_1.*

Using $\psi_k(z) = |z|$ in Theorem 3 implies that $\mathcal{F}$ is learnable with respect to ℓ_1 if and only if each $\mathcal{F}_k$ is learnable with respect to d_1. Thus, to prove Theorem 5, it suffices to show that for all $p > 1$, $\mathcal{F}$ is learnable with respect to ℓ_p if and only if $\mathcal{F}$ is learnable with respect to ℓ_1 norm.

Proof. We begin by proving the sufficiency direction. As discussed above, the learnability of each $\mathcal{F}_k$ with respect to d_1 implies that $\mathcal{F}$ is learnable with respect to ℓ_1. Then, the high-level idea of the proof is to use an agnostic learner for $\mathcal{F}$ with respect to ℓ_1 to construct a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Using the fact $\ell_1(y_1, y_2) = 0$ if and only if $\ell_p(y_1, y_2) = 0$ for any $y_1, y_2 \in \mathcal{Y}$, the realizable learner for ℓ_1 is also a realizable learner for ℓ_p. Finally, as $|\text{im}(\mathcal{F}^\alpha)| < \infty$ for every $\alpha > 0$, Lemma 1 guarantees the existence of an agnostic learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Then, a simple application of the triangle inequality shows that an agnostic learner for $\mathcal{F}^\alpha$ is also an agnostic learner for $\mathcal{F}$ with respect to ℓ_p. We now proceed with the formal proof.

Part 1: Sufficiency. Fix $p > 1$. Recall that the learnability of each $\mathcal{F}_k$ with respect to d_1 implies that $\mathcal{F}$ is learnable with respect to ℓ_1. Let $\mathcal{D}$ be arbitrary distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{A}$ be an agnostic PAC learner for $\mathcal{F}$ with respect to ℓ_1 with sample complexity $m(\epsilon, \delta)$. For any $\epsilon, \delta > 0$, with n sufficiently larger than $m(\epsilon/2, \delta, K)$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\ell_1(g(x), y)] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}[\ell_1(f(x), y)] + \frac{\epsilon}{2},$$

where $g = \mathcal{A}(S)$. Define $\mathcal{F}^\alpha$ to be a discretized function class obtained by discretizing $\mathcal{F}$ component-wise using the scheme (2.1). Consider $\mathcal{D}$ to be a realizable distribution with respect to $\mathcal{F}^\alpha$. Note that the triangle inequality implies $\ell_1(f(x), y) \leq \ell_1(f^\alpha(x), y) + \ell_1(f(x), f^\alpha(x)) \leq \ell_1(f^\alpha(x), y) + K\alpha$. Taking $\alpha = \frac{\epsilon}{2K}$ and using the fact that $\inf_{f \in \mathcal{F}} \mathbb{E}[\ell_1(f^\alpha(x), y)] = 0$, we obtain $\mathbb{E}_{\mathcal{D}}[\ell_1(g(x), y)] \leq \epsilon$. Next, using the inequality $\ell_p(g(x), y) \leq \ell_1(g(x), y)$ pointwise yields

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\ell_p(g(x), y)] \leq \epsilon.$$

Therefore, $\mathcal{A}$ is a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_p with sample complexity $m(\epsilon/2, \delta, K)$. Since $|\mathrm{im}(\mathcal{F}^\alpha)| < \infty$ and $\ell_p(\cdot, \cdot)$ are 1-subadditive, Lemma 1 implies that $\mathcal{F}^\alpha$ is agnostic learnable with respect to ℓ_p via Algorithm 1, referred to as algorithm $\mathcal{B}$ henceforth. Thus, for any $\epsilon, \delta > 0$, there exists a $n \geq m_{\mathcal{B}}(\epsilon/2, \delta/2)$, for any distribution $\widetilde{\mathcal{D}}$ on $\mathcal{X} \times \mathcal{Y}$, running $\mathcal{B}$ on $S \sim \mathcal{D}^n$ outputs a predictor $\tilde{g} \in \mathcal{Y}^{\mathcal{X}}$ such that with probability at least $1 - \delta$ over $S \sim \widetilde{\mathcal{D}}^n$, we have

$$\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}[\ell_p(\tilde{g}(x), y)] \leq \inf_{f^\alpha \in \mathcal{F}^\alpha} \mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}[\ell_p(f^\alpha(x), y)] + \frac{\epsilon}{2} = \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}[\ell_p(f^\alpha(x), y)] + \frac{\epsilon}{2}.$$

Using triangle inequality, we have $\ell_p(f^\alpha(x), y) \leq \ell_p(f(x), y) + \ell_p(f^\alpha(x), f(x)) \leq \ell_p(f(x), y) + \alpha K$ pointwise. Again taking $\alpha = \frac{\epsilon}{2K}$, we obtain

$$\mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}[\ell_p(\tilde{g}(x), y)] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{\widetilde{\mathcal{D}}}[\ell_p(f(x), y)] + \epsilon.$$

Therefore, we have shown that $\mathcal{F}$ is agnostic PAC learnable with respect to ℓ_p.

The sample complexity of agnostic learner $\mathcal{B}$ can be made precise using the sample complexity of Algorithm 1. In particular, the sample complexity of $\mathcal{B}$ is the sample complexity of the realizable learner $\mathcal{A}$ and the sample complexity of the ERM in step 2 of Algorithm 1. Proof of Lemma 1 shows that the sample complexity of $\mathcal{B}$ must be

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) + \frac{8b^2}{\epsilon^2}\left(m_{\mathcal{A}}\left(\frac{\epsilon}{2c}, \frac{\delta}{2}, K\right) \log |(\mathrm{im}(\mathcal{F}^\alpha)|) + \log \frac{4}{\delta}\right),$$

where b is the upperbound on the loss and c is the subadditivity constant. Since $b \leq K$, and $c = 1$ for all ℓ_p norms with $p \geq 1$, we obtain

$$m_{\mathcal{B}}(\epsilon, \delta, K) \leq m_{\mathcal{A}}(\epsilon/2, \delta/2, K) + \frac{8K^3}{\epsilon^2}\left(m_{\mathcal{A}}(\epsilon/2, \delta/2, K)\,\log\left(\frac{4K}{\epsilon}\right) + \log\frac{4}{\delta}\right),$$

where we also use the fact that $|\mathrm{im}(\mathcal{F}^\alpha)| \leq (2/\alpha)^K$ for our choice of $\alpha = \frac{\epsilon}{2K}$.

Part 2: Necessity. Fix $p > 1$. We now prove that $\mathcal{F}$ being learnable with respect to ℓ_p implies $\mathcal{F}$ is learnable with respect to ℓ_1. The proof is identical to the proof of sufficiency, so we only provide a sketch of the argument here. Our proof strategy follows a similar route through realizable learnability of the discretized class $\mathcal{F}^\alpha$ and then the use of Lemma 1.

Recall that any agnostic learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_p is a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Using the inequality $\ell_1(\cdot,\cdot) \leq K\ell_p(\cdot,\cdot)$ pointwise, we can deduce that $\mathcal{A}$ is also a realizable learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Since $|\mathrm{im}(\mathcal{F}^\alpha)| \leq (2/\alpha)^K < \infty$, Lemma 1 guarantees existence of an agnostic learner for $\mathcal{F}^\alpha$ with respect to ℓ_1. Using triangle inequality, we obtain $\ell_1(f^\alpha(x), y) \leq \ell_1(f(x), y) + \ell_1(f^\alpha(x), f(x)) \leq \ell_1(f(x), y) + \alpha K$ pointwise, and choosing appropriate discretization scale allows us to turn agnostic bound for $\mathcal{F}^\alpha$ into an agnostic bound for $\mathcal{F}$. ■

Remark. We note that Theorem 5 holds for any norm on $\mathbb{R}^K$, but we only focus on ℓ_p norms here due to their practical significance. As the fat-shattering dimension of a real-valued function class characterizes its learnability with respect to d_1 loss [Bartlett et al., 1996], Theorem 5 implies that a multioutput function class $\mathcal{F}$ is learnable with respect to ℓ_p for $1 \leq p \leq \infty$ if and only if $\mathrm{fat}_\gamma(\mathcal{F}_k) < \infty$ for all $k \in [K]$ at every fixed scale $\gamma > 0$.

Since we are only concerned with the question of learnability in this work, our focus is not on optimal sample complexity rates. However, we point out that for any $p \geq 1$, if each $\mathcal{F}_k$ is learnable with respect to d_1, then $\mathcal{F}$ is learnable with respect to ℓ_p via ERM with a better sample complexity than Algorithm 1. The proof of this claim is based on Rademacher complexity and is provided in Appendix A.4.

2.4 Online Multioutput Learnability

Here, we study the online learnability of multioutput function classes. Throughout this section, we give regret bounds assuming an *oblivious* adversary. A standard reduction (Chapter 4 in Cesa-Bianchi and Lugosi [2006]) allows us to convert oblivious regret bounds to adaptive regret bounds in the full-information setting. A key requirement allowing an oblivious regret bound to generalize to an adaptive regret bound is that the learner's predictions on round t

should not depend on any of its past predictions from previous rounds. This is true for all of the online learning algorithms in this section.

2.4.1 Online Agnostic-to-Realizable Reduction

Our strategy for constructively characterizing the learnability of general losses in both the batch classification and regression setting required the ability to convert a realizable learner to an agnostic learner in a black-box fashion. In this section, we provide an analog of this conversion for the *online* setting. More specifically, we focus on the setting where $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a multioutput function class but $|\text{im}(\mathcal{F})| < \infty$ is finite. Then, for any c-subadditive loss function ℓ, we constructively convert a (potentially randomized) realizable online learner for $\mathcal{F}$ with respect to ℓ into an agnostic online learner for $\mathcal{F}$ with respect to ℓ. Theorem 6 formalizes the main result of this subsection.

Theorem 6. *Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multioutput function class such that $|im(\mathcal{F})| < \infty$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any c-subadditive loss function such that $\ell(\cdot, \cdot) \leq M$. If $\mathcal{A}$ is a realizable online learner for $\mathcal{F}$ with respect to ℓ with sub-linear expected regret $R(T, |im(\mathcal{F})|)$, then for every $\beta \in (0, 1)$, there exists an agnostic online learner for $\mathcal{F}$ with respect to ℓ with expected regret*

$$\frac{cT}{T^{\beta}} \overline{R}(T^{\beta}, |im(\mathcal{F})|) + M\sqrt{2T^{1+\beta} \ln(|im(\mathcal{F})|)},$$

where $\overline{R}(T, |im(\mathcal{F})|)$ is any concave, sublinear upperbound on $R(T, |im(\mathcal{F})|)$.

Note that if $\overline{R}(T^{\beta}, |\text{im}(\mathcal{F})|)$ is sublinear in its first argument, then $\frac{cT}{T^{\beta}}\overline{R}(T^{\beta}, |\text{im}(\mathcal{F})|) + M\sqrt{2T^{1+\beta} \ln(|\text{im}(\mathcal{F})|)}$ is sublinear in T for any $\beta \in (0, 1)$. By Lemma 4, we are guaranteed the existence of $\overline{R}(T, |\text{im}(\mathcal{F})|)$.

Lemma 4. *[Ceccherini-Silberstein et al., 2017, Lemma 5.17] Let g be a positive sublinear function. Then, g is bounded from above by a concave sublinear function.*

Therefore, Theorem 6 and Lemma 4 show that for any function class $\mathcal{F}$ with finite image space and any c-subadditive loss function, realizable and agnostic online learnability are equivalent. We now begin the proof of Theorem 6.

Proof. Let $\mathcal{A}$ be a (potentially randomized) online realizable learner for $\mathcal{F}$ with respect to ℓ. By definition, this means that for any (realizable) sequence $(x_1, f(x_1)), ..., (x_T, f(x_T))$ labeled by a function $f \in \mathcal{F}$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), f(x_t))\right] \leq R(T, |\text{im}(\mathcal{F})|),$$

where $R(T, |\text{im}(\mathcal{F})|)$ is a sub-linear function of T. We now use $\mathcal{A}$ to construct an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ. Since we are assuming an oblivious adversary, let $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times \mathcal{Y})^T$ denote the stream of points to be observed by the online learner and $f^\star = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)$ to be the optimal function in hindsight.

Our high-level strategy is to construct a large set of Experts that approximately cover all possible labelings of the instances $x_1, ..., x_T$ by functions in $\mathcal{F}$. In particular, each Expert uses an independent copy of $\mathcal{A}$ to make predictions, but update $\mathcal{A}$ using *different* sequences of labeled instances. Together, our set of Experts update $\mathcal{A}$ using all possible sequences of labeled instances. In order to ensure that the number of Experts is not too large, we construct such a set of Experts over a sufficiently small *sub-sample* of the stream. Finally, we run the celebrated Randomized Exponential Weights Algorithm (REWA) [Cesa-Bianchi and Lugosi, 2006] using our set of experts and the scaled loss function $\frac{\ell}{M}$ over the original stream of points $(x_1, y_1), ..., (x_T, y_T)$. We now formalize this idea below.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \text{im}(\mathcal{F})$ denote a function mapping time points where $b_t = 1$ to elements in the image space $\text{im}(\mathcal{F})$. Let $\Phi_b \subseteq (\text{im}(\mathcal{F}))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $f \in \mathcal{F}$, let $\phi_b^f \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi_b^f(t) = f(x_t)$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 3, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to compute a labeled instance $(x_t, \phi(t))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and plays $\hat{y}_t = \mathcal{A}(x_t)$ for all $t \in [T]$. Note that $1 \leq |\mathcal{E}_b| \leq (|\text{im}(\mathcal{F})|)^{|b|}$.

Algorithm 3 Expert(b, ϕ)

Require: Independent copy of realizable online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ

```
for t = 1, ..., T do
    Receive example x_t
    Predict ŷ_t = A(x_t)
    if b_t = 1 then
        Update A by passing (x_t, φ(t))
    end if
end for
```

With this notation in hand, we are now ready to present Algorithm 4, our main agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ. Our goal is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in

Algorithm 4 Agnostic online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ

Require: Parameter $0 < \beta < 1$

1: Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{T^\beta}{T}\right)$
2: Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 3
3: Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the loss function $\frac{\ell}{M}$ over the stream $(x_1, y_1), \ldots, (x_T, y_T)$

sampling B, the internal randomness of $\mathcal{A}$, and the internal randomness of REWA. Let B, A and P denote the random variables associated with each source of randomness respectively. By construction, B, A, and P are independent.

Using Theorem 21.11 in Shalev-Shwartz and Ben-David [2014] and the fact that B, A and P are independent, REWA guarantees almost surely that

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell(\mathcal{P}(x_t), y_t) | B, A\right] \leq \inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t) + M\sqrt{2T \ln(|\mathcal{E}_B|)}.$$

Taking an outer expectation gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{P}(x_t), y_t)\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{Q}(x_t), y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{P}(x_t), y_t)\right] \\
&\leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] + \mathbb{E}\left[M\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] + M\mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].
\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $f \in \mathcal{F}$, we have $E_{b,\phi_b^f} \in \mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T} \ell(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right]$. To do so, we need some additional notation. Given the realizable online learner $\mathcal{A}$, an instance $x \in \mathcal{X}$, and an ordered finite sequence of labeled examples $L \in (\mathcal{X} \times \mathcal{Y})^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $f \in \mathcal{F}$, and $t \in [T]$, let $L_{b_{<t}}^f = \{(x_i, f(x_i)) : i < t \text{ and } b_i = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_i, f(x_i))\}_{i=1}^{t-1}$ where $b_i = 1$. Using this notation, we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\ell(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\frac{\mathbb{P}\left[B_t=1\right]}{\mathbb{P}\left[B_t=1\right]}\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\mathbb{P}\left[B_t=1\right]\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\mathbb{1}\{B_t=1\}\right].
\end{aligned}$$

To see the last equality, note that the prediction $\mathcal{A}(x_t|L_{B_{<t}}^{f^\star})$ only depends on bitstring $(B_1,\ldots,B_{t-1})$ and the internal randomness of A, both of which are independent of B_t. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\,\mathbb{1}\{B_t=1\}\right] &= \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\right]\,\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right] \\
&= \mathbb{E}\left[\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\right]\mathbb{P}[B_t=1]
\end{aligned}$$

as needed. Continuing onwards,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\ell(E_{B,\phi_B^{f^\star}}(x_t),y_t)\right] &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\mathbb{1}\{B_t=1\}\right] \\
&= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),y_t)\right] \\
&\le \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),f^\star(x_t))\right] + \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t),y_t)\right] \\
&= \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),f^\star(x_t))\right] + \sum_{t=1}^{T}\ell(f^\star(x_t),y_t) \\
&= \frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),f^\star(x_t))\right] + \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)
\end{aligned}$$

The inequality follows from the fact that ℓ is a c-subadditive and the last equality follows from the definition of $f^\star$. We now need to bound $\frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L_{B_{<t}}^{f^\star}),f^\star(x_t))\right]$. Using the fact that $\mathcal{A}^\star$ is a realizable online learner and gets updated on a stream of instances

labeled by $f^\star$ only on rounds where $B_t = 1$, we get

$$\begin{aligned}\frac{cT}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), f^\star(x_t))\right] &= \frac{cT}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), f^\star(x_t))\Big|B\right]\right] \\ &\leq \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right].\end{aligned}$$

Putting things together, we find that,

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T}\ell(E_{B,\phi^{f^\star}_B}(x_t), y_t)\right] + M\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\ &\leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\ &\leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\mathbb{E}\left[\sqrt{2T|B|\ln(|\mathrm{im}(\mathcal{F})|)}\right],\end{aligned}$$

where the last inequality follows from the fact that that $|\mathcal{E}_B| \leq (|\mathrm{im}(\mathcal{F})|)^{|B|}$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2T|B|\ln(|\mathrm{im}(\mathcal{F})|)}\right] \leq \sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\mathbb{E}\left[R(|B|, |\mathrm{im}(\mathcal{F})|)\right] + M\sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}.$$

Next, by Lemma 4, there exists a concave sublinear function $\overline{R}(|B|, |\mathrm{im}(\mathcal{F})|)$ that upperbounds $R(|B|, |\mathrm{im}(\mathcal{F})|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}(|B|, |\mathrm{im}(\mathcal{F})|)] \leq \overline{R}(T^\beta, |\mathrm{im}(\mathcal{F})|)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t), y_t) + \frac{cT}{T^\beta}\,\overline{R}(T^\beta, |\mathrm{im}(\mathcal{F})|) + M\sqrt{2T^{\beta+1}\ln(|\mathrm{im}(\mathcal{F})|)}.$$

This completes the proof as we have shown that $\mathcal{Q}$ is an agnostic online learner for $\mathcal{F}$ with respect to ℓ with the stated regret bound. ■

2.4.2 Online Multilabel Classification

Let $\mathcal{Y} = \{-1, 1\}^K$. We provide analogs of Theorem 1 and 2 in the online setting. We begin by characterizing the learnability of the Hamming loss and then move to give a characterization of learnability for all losses satisfying the identity of indiscernibles. Similar to the batch setting, we can show that the MCLdim of $\mathcal{F}$ characterizes online multilabel learnability (see Appendix A.6), but here, we give a characterization that better exploits the multilabel structure of the problem.

2.4.2.1 Characterizing Online Learnability for the Hamming Loss

Theorem 7 characterizes the online learnability of a multilabel function class $\mathcal{F}$ with respect to ℓ_H.

Theorem 7. *A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to the Hamming loss if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to the 0-1 loss.*

The proof of Theorem 7 is similar to that of Theorem 1, so we defer the full proof to Appendix A.5 and only provide a sketch here. The proof of sufficiency direction is based on a reduction: given oracle access to online learners $\{\mathcal{A}_k\}_{k=1}^K$ for $\{\mathcal{F}_k\}_{k=1}^K$ with respect to $\ell_{0\text{-}1}$, we construct an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. In fact, similar to the batch setting, the online multilabel learning algorithm $\mathcal{A}$ is simple: in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. Using some algebra, one can show that this prediction rule achieves sublinear regret for $\mathcal{F}$.

For the necessity direction, given oracle access to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H, we construct an online learner $\mathcal{B}$ for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. A similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. Similar to the batch setting, the online learning algorithm $\mathcal{B}$ is simple: in each round $t \in [T]$, receive x_t, query $\hat{y}_t = \mathcal{A}(x_t)$ and predict $\hat{y}_t^1 = \mathcal{A}_1(x_t)$. Once the true label y_t^1 is revealed, update $\mathcal{A}$ by passing (x_t, y_t) where $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$ and $\{\sigma_t^i\}_{i=2}^K$ is an i.i.d sequence of Rademacher random variables. A straightforward analysis shows that such a prediction rule achieves sublinear regret for $\mathcal{F}_1$.

2.4.2.2 Characterizing Online Learnability for General Losses

Using Theorem 6 and Theorem 7, we now characterize the learnability of arbitrary multilabel loss functions ℓ as long as they satisfy the identity of indiscernibles. The key idea is that since there are only finite number of possible inputs to ℓ, for any ℓ satisfying the identity of

indiscernibles, there must exist universal constants a and b such that $a\ell_H(y_1, y_2) \leq \ell(y_1, y_2) \leq b\ell_H(y_1, y_2)$. Then, we can characterize the learnability of ℓ by relating it to the learnability of ℓ_H. In fact, we prove a slightly more general result, showing an equivalence between the learnability of any two arbitrary losses satisfying the identity of indiscernibles.

Lemma 5. *Let ℓ and ℓ' be any two loss functions satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if $\mathcal{F}$ is online learnable with respect to ℓ'.*

The proof of Lemma 5 is similar to that of Lemma 2 with the main difference being the use of Theorem 6 instead of Lemma 1. Since Lemma 5 is our first application of Theorem 6, we provide the full proof here.

Proof. Since ℓ and ℓ' are arbitrary, it suffices to prove only one direction. To that end, suppose $\mathcal{F}$ is online learnable with respect to ℓ. We now show that $\mathcal{F}$ is online learnable with respect to ℓ' as well.

Let a and b be the universal constants such that for all $y_1, y_2 \in \mathcal{Y}$, $a\ell(y_1, y_2) \leq \ell'(y_1, y_2) \leq b\ell(y_1, y_2)$. Let $c = \frac{\max_{r \neq t} \ell'(r,t)}{\min_{r \neq t} \ell'(r,t)}$. Since $|\text{im}(\mathcal{F})| = 2^K < \infty$ and ℓ' is a c-subadditive, by Theorem 6, it suffices to give a realizable online learner for $\mathcal{F}$ with respect to ℓ'. Since $\mathcal{F}$ is online learnable with respect to ℓ, there exists an algorithm $\mathcal{A}$ such that for any sequence $(x_1, y_1), ..., (x_T, y_T)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \leq R(T, 2^K)$$

where $R(T, 2^K)$ is a sublinear function of T. In the realizable setting, we are guaranteed that for any sequence $(x_1, y_1), ..., (x_T, y_T)$ that the online learner may observe, there exists a $f \in \mathcal{F}$ s.t $f(x_t) = y_t$ for all $t \in [T]$. Since ℓ satisfies the identity of indiscernibles, we have that for any realizable sequence $(x_1, y_1), ..., (x_T, y_T)$, $\inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t) = 0$. Thus, we have that $\mathbb{E}\left[\sum_{t=1}^{T} \ell(\mathcal{A}(x_t), y_t)\right] \leq R(T, 2^K)$. Noting that $\ell(\mathcal{A}(x_t), y_t) \geq \frac{\ell'(\mathcal{A}(x_t), y_t)}{b}$ implies that $\mathbb{E}\left[\sum_{t=1}^{T} \ell'(\mathcal{A}(x_t), y_t)\right] \leq bR(T, 2^K)$, showing that $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}$ with respect to ℓ'. For any $\beta \in (0, 1)$, the construction in Theorem 6 can then be used to convert $\mathcal{A}$ into an agnostic online learner for $\mathcal{F}$ with respect to ℓ' with expected regret bound

$$\frac{cbT}{T^\beta}\overline{R}(T^\beta, 2^K) + M\sqrt{4KT^{1+\beta}}$$

where M is such that $\ell \leq M$ and $\overline{R}(T^\beta, 2^K)$ is any concave sublinear upperbound of $R(T^\beta, 2^K)$. This completes our proof.

■

As an immediate consequence of Lemma 5 and Theorem 7, we get the following theorem characterizing the online learnability of general multilabel losses.

Theorem 8. *Let ℓ be any multilabel loss function that satisfies the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to the 0-1 loss.*

Remark. Since the Littlestone dimension characterizes online learnability for binary classification under the 0-1 loss [Ben-David et al., 2009], Theorem 8 also implies that finiteness of $\mathrm{Ldim}(\mathcal{F}_k)$ for all $k \in [K]$ is a necessary and sufficient condition for online multilabel learnability.

Moreover, if $\mathrm{Ldim}(\mathcal{F}_k) < \infty$ for all $k \in [K]$, then we have $\mathrm{MCLdim}(\mathcal{F}) < \infty$. This follows from the fact that $\mathrm{MCLdim}(\mathcal{F}) \leq \sum_{k=1}^{K} \mathrm{Ldim}(\mathcal{F}_k)$. To see this, note that $\mathrm{MCLdim}(\mathcal{F})$ is the lowerbound on the number of mistakes of any deterministic multiclass learner in the realizable setting [Daniely et al., 2011, Theorem 17]. On the other hand, one can construct a deterministic realizable learner for $\mathcal{F}$ using K different Standard Optimal Algorithms (SOA) for binary function classes $\mathcal{F}_k$'s. Namely, define an algorithm $\mathcal{A}$ such that $\mathcal{A}(x) := (\mathrm{SOA}(\mathcal{F}_1)(x), \ldots, \mathrm{SOA}(\mathcal{F}_K)(x)) \in \{-1, 1\}^K$. Since each $\mathrm{SOA}(\mathcal{F}_k)$ makes at most $\mathrm{Ldim}(\mathcal{F}_k)$ number of mistakes, $\mathcal{A}$ makes no more than $\sum_{k=1}^{K} \mathrm{Ldim}(\mathcal{F}_k)$ mistakes. We can use this fact to give an improved version of Theorem 6 for classes $\mathcal{F}$ with $\mathrm{MCLdim}(\mathcal{F}) < \infty$. In particular, when $\mathcal{Y} = \{-1, 1\}^K$, any $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ that is learnable in the realizable setting with respect to ℓ is also learnable in the agnostic setting with regret $O\left(B\sqrt{T\mathrm{MCLdim}(\mathcal{F})\ln(T)}\right) \leq O\left(B\sqrt{T\sum_{k=1}^{K}\mathrm{Ldim}(\mathcal{F}_k)\ln(T)}\right)$. Here, B is the maximum value ℓ can take. The improved regret bound can be found in the sufficiency proof of Theorem 37 in Appendix A.6.

2.4.3 Bandit Online Multilabel Classification

We extend the results in the previous subsection to the online setting where the learner only observes *bandit* feedback in each round. Theorem 9 gives a characterization of bandit online learnability of a function class $\mathcal{F}$ in terms of the online learnability of each restriction.

Theorem 9. *Let ℓ be any loss function that satisfies the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is bandit online learnable with respect to ℓ if and only if each restriction $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to the 0-1 loss.*

Similar to the full-feedback setting, Theorem 9 also gives that the finiteness of $\mathrm{Ldim}(\mathcal{F}_k)$ for all $k \in [K]$ is a necessary and sufficiency condition for bandit online multilabel learnability. The proof of Theorem 9 uses the realizable-to-agnostic conversion for *bandit* feedback setting when the label space $\mathcal{Y}$ is finite. The following Theorem makes this argument precise.

Theorem 10. *Let $\mathcal{Y}$ be a finite label space, $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ a multioutput function class, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be any c-subadditive loss function such that $\ell(\cdot, \cdot) \leq M$. If $\mathcal{A}$ is a realizable online learner for $\mathcal{F}$ with respect to ℓ under full-feedback with sub-linear expected regret $R(T, |\mathcal{Y}|)$, then for every $\beta \in (0, 1)$, there exists an online learner for $\mathcal{F}$ with respect to ℓ with expected regret*

$$\frac{cT}{T^{\beta}}\overline{R}(T^{\beta}, |\mathcal{Y}|) + eM\sqrt{2T^{1+\beta}|\mathcal{Y}|\ln(|\mathcal{Y}|)},$$

under bandit feedback, where $\overline{R}(T, |\mathcal{Y}|)$ is any concave, sublinear upperbound on $R(T, |\mathcal{Y}|)$.

The proofs for Theorem 9 and Theorem 10 are provided in Appendix A.7.

2.4.4 Online Multioutput Regression

In this section, we characterize the online learnability of multioutput function classes. Similar to the batch setting, we consider, without loss of generality, the case when $\mathcal{Y} = [0, 1]^K \subset \mathbb{R}^K$ for $K \in \mathbb{N}$. In addition, we consider the same set of decomposable and non-decomposable loss functions as in the batch setting. Namely, our decomposable loss functions satisfy Assumptions 1 and 2, and our non-decomposable loss functions are ℓ_p norms. Informally, our main result asserts that a multioutput function class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ is online learnable if and only if each restriction $\mathcal{F}_k$ is online learnable.

2.4.4.1 Characterizing Learnability for Decomposable Losses

In this subsection, we characterize the online learnability of multioutput function classes with respect to decomposable losses satisfying Assumption 1. Our main theorem is presented below.

Theorem 11. *Let ℓ be a decomposable loss function satisfying Assumption 1. A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to $\psi_k \circ d_1$.*

Proof. As usual, we prove Theorem 11 in two parts: first sufficiency and then necessity. The sufficiency proof is similar to that of the proof for Hamming loss in Theorems 1 and 7. The necessity direction is more involved, but the main idea is to combine the augmentation technique used in the proof of Theorem 3 with the algorithmic conversion technique developed in the proof of Theorem 6.

Part 1: Sufficiency. We first prove that online learnability of each restriction $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ is sufficient for online learnability of $\mathcal{F}$ with respect to ℓ. Since

$\ell(f(x), y) = \sum_{k=1}^{K} \psi_k \circ d_1(f_k(x), y_k)$ is decomposable, we can use the exact same strategy as in Section 2.4.2.1 to convert online learners $\mathcal{A}_1$, ..., $\mathcal{A}_K$ for $\mathcal{F}_1, ..., \mathcal{F}_K$ with respect to $\psi_k \circ d_1$ to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ. More specifically, in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. Using the exact same proof as in Section 2.4.2.1, it follows that the expected regret of $\mathcal{A}$ is $\sum_{k=1}^{K} R_k(T)$ where $R_k(T)$ is the regret of online algorithm $\mathcal{A}_k$. Since K is finite, the regret of $\mathcal{A}$ is sublinear in T when evaluated using ℓ.

Part 2: Necessity. Similar to the batch setting, we prove the necessity direction of Theorem 11 constructively. That is, given oracle access to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ, we construct an online learner $\mathcal{Q}$ for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$. By symmetry, a similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. As mentioned before, we assume an oblivious adversary, and therefore the stream of points to be observed by the online learner, denoted $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times [0, 1])^T$, is fixed beforehand. Let $f_1^\star = \arg\min_{f_1 \in \mathcal{F}_1} \sum_{t=1}^{T} \psi_k \circ d_1(f_1(x_t), y_t)$ denote the optimal function in hindsight and $f^\star \in \mathcal{F}$ its completion.

Since we are trying to construct an online learner for $\mathcal{F}_1$, the targets $y_1, ..., y_T$ are *scalar-valued*. However, $\mathcal{A}$ is an online learner for $\mathcal{F}$ and therefore can only processes *vector-valued* targets. Thus, we need to figure out how to augment the scalar-valued targets $y_1, ..., y_T$ in a way that allows us to use $\mathcal{A}$ to construct an online learner $\mathcal{Q}$ for $\mathcal{F}_1$. Following a similar strategy as in the proof of Theorem 6, we can construct a set of Experts that simulate online games with $\mathcal{A}$ by augmenting, in all possible ways, the scalar-valued targets of a *sub-sample* of the stream into vector-valued targets using vectors in $\mathcal{Y}_{2:k}^{\alpha}$, the discretized label space for components 2 through K. In particular, our high-level strategy is to:

1. Randomly *sub-sample* points from the stream
2. Construct a set of Experts, each of which:
   (a) Uses an independent copy of $\mathcal{A}$ to make predictions $\hat{y}_t = \mathcal{A}_1(x_t)$
   (b) Augments the scalar-valued targets of each labeled instance in the *sub-sampled* stream to vector-valued targets using vectors in the discretized image space $\text{im}(\mathcal{F}_{2:K}^{\alpha})$
   (c) Simulates an online game with its independent copy of $\mathcal{A}$ over only the augmented *sub-sampled* stream with vector-valued targets

3. Run REWA using the set of experts in Step 2 and the $\psi_1 \circ d_1$ loss function over the *original* stream of points.

We now formalize this idea. For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \text{im}(\mathcal{F}^{\alpha}_{2:K})$ denote a function mapping time points where $b_t = 1$ to vectors in the discretized image space $\text{im}(\mathcal{F}^{\alpha}_{2:K})$. Let $\Phi_b \subseteq (\text{im}(\mathcal{F}^{\alpha}_{2:K}))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $f \in \mathcal{F}$, let $\phi^f_b \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi^f_b(t) = f^{\alpha}_{2:K}(x_t)$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 5, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to augment the scalar-valued labeled instance (x_t, y_t) to the vector-valued labeled instance $(x_t, (y_t, \phi(t)))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$ and plays $\mathcal{A}_1(x_t)$ for all $t \in [T]$. Note that $1 \leq |\mathcal{E}_b| \leq (\frac{2}{\alpha})^{K|b|}$.

Algorithm 5 Expert(b, ϕ)

Require: Independent copy of Online Learner $\mathcal{A}$ for ℓ

for $t = 1, \ldots, T$ **do**
  Receive example x_t
  Predict $\tilde{y}_t = \mathcal{A}_1(x_t)$
  Receive y_t
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(x_t, (y_t, \phi(t)))$
  end if
end for

With this notation in hand, we are now ready to present Algorithm 6, our main online learner $\mathcal{Q}$ for $\mathcal{F}_1$.

Algorithm 6 Online learner $\mathcal{Q}$ for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$

Require: Parameters $0 < \beta < 1$ and $0 < \alpha < 1$

Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{T^{\beta}}{T}\right)$
Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 5
Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the loss function $\psi_1 \circ d_1$ over the stream $(x_1, y_1), \ldots, (x_T, y_T)$

Our goal now is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in sampling B, the internal randomness of each independent copy of the online learner $\mathcal{A}$, and the internal randomness of REWA.

Let B, A and P denote the random variable associated with these sources of randomness respectively. By construction, B, A, and P are independent.

Using Theorem 21.11 in Shalev-Shwartz and Ben-David [2014] and the fact that A, P, and B are independent, REWA guarantees

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{P}(x_t), y_t)\right] \leq \mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\psi_1 \circ d_1(E(x_t), y_t)\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right].$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{P}(x_t), y_t)\right] \\
&\leq \mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\psi_1 \circ d_1(E(x_t), y_t)\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] + \mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right].
\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $f \in \mathcal{F}$, $E_{b,\phi_b^f} \in \mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T}\psi \circ d_1(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right]$. We use the same notation used to prove Theorem 6, but for the sake of completeness, we restate it here. Given an online learner $\mathcal{A}$ for ℓ, an instance $x \in \mathcal{X}$, and an ordered sequence of labeled examples $L \in (\mathcal{X} \times [0,1]^K)^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $f_{2:K}^\alpha \in \mathcal{F}_{2:K}^\alpha$, and $t \in [T]$, let $L_{b_{<t}}^f = \{(x_i, (y_i, f_{2:K}^\alpha(x_i))) : i < t \text{ and } b_i = i\}$ denote the *subsequence* of the sequence of labeled instances $\{(x_i, (y_i, f_{2:K}^\alpha(x_i)))\}_{i=1}^{t-1}$ where $b_i = 1$. Using this notation, we can write

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(E_{B,\phi_B^{f^\star}}(x_t), y_t)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{f^\star}), y_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{f^\star}), y_t)\frac{\mathbb{P}\left[B_t = 1\right]}{\mathbb{P}\left[B_t = 1\right]}\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\psi_1 \circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{f^\star}), y_t)\mathbb{P}\left[B_t = 1\right]\right] \\
&= \frac{T}{T^\beta}\sum_{t=1}^{T}\mathbb{E}\left[\psi_1 \circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{f^\star}), y_t)\mathbb{1}\{B_t = 1\}\right].
\end{aligned}$$

To see the last equality, note that the prediction $\mathcal{A}(x_t|L_{B_{<t}}^{f^\star})$ (and therefore $\mathcal{A}_1(x_t|L_{B_{<t}}^{f^\star})$)

only depends on bitstring $(B_1, \ldots, B_{t-1})$ and the internal randomness of A, both of which are independent of B_t. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\psi_1 \circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}), y_t)\mathbb{1}\{B_t = 1\}\right] &= \mathbb{E}\left[\psi_1 \circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}), y_t)\right]\mathbb{E}\left[\mathbb{1}\{B_t = 1\}\right] \\ &= \mathbb{E}\left[\psi_1 \circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}), y_t)\right]\mathbb{P}[B_t = 1]\end{aligned}$$

as needed. Continuing onwards,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(E_{B,\phi^{f^\star}_B}(x_t), y_t)\right] &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}), y_t)\mathbb{1}\{B_t = 1\}\right] \\ &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi_1 \circ d_1(\mathcal{A}_1(x_t|L^{f^\star}_{B_{<t}}), y_t)\right] \\ &\leq \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), (y_t, f^{\star,\alpha}_{2:K}(x_t)))\right] \\ &= \frac{T}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\ell(\mathcal{A}(x_t|L^{f^\star}_{B_{<t}}), (y_t, f^{\star,\alpha}_{2:K}(x_t)))\middle| B\right]\right] \\ &\leq \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t), (y_t, f^{\star,\alpha}_{2:K}(x_t))) + R_{\mathcal{A}}(|B|)\right] \\ &= \frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t), (y_t, f^{\star,\alpha}_{2:K}(x_t)))\right] + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\end{aligned}$$

The first inequality follows from the definition of ℓ. The second inequality follows from the fact that $\mathcal{A}$ is an online learner for ℓ with regret bound $R_{\mathcal{A}}(T)$ and is updated on the stream labeled by $f^{\star,\alpha}_{2:K}$ only when $B_t = 1$. Now, we can upperbound the first term as follows:

$$
\begin{aligned}
&\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\ell(f^\star(x_t),(y_t,f^{\star,\alpha}_{2:K}(x_t)))\right]\\
&=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\left(\psi_1\circ d_1(f^\star_1(x_t),y_t)+\sum_{k=2}^{K}\psi_k\circ d_1(f^\star_k(x_t),f^{\star,\alpha}_k(x_t))\right)\right]\\
&\leq\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi_1\circ d_1(f^\star_1(x_t),y_t)+\sum_{t:B_t=1}KL\alpha\right]\\
&\leq\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)\mathbb{1}\{B_t=1\}\right]+\frac{T}{T^\beta}\mathbb{E}\left[|B|KL\alpha\right]\\
&=\frac{T}{T^\beta}\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)\frac{T^\beta}{T}+\frac{T}{T^\beta}T^\beta KL\alpha\\
&=\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)+KTL\alpha.
\end{aligned}
$$

The first inequality follows from the fact that ψ_k is L-Lipschitz and $d_1(f^\star_k(x_t),f^{\star,\alpha}_k(x_t))\leq\alpha$. Putting things together, we find that,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(\mathcal{Q}(x_t),y_t)\right]\\
&\qquad\leq\mathbb{E}\left[\sum_{t=1}^{T}\psi_1\circ d_1(E_{B,\phi^{f^\star}_B}(x_t),y_t)\right]+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]\\
&\qquad\leq\sum_{t=1}^{T}\psi_1\circ d_1(f^\star_1(x_t),y_t)+KTL\alpha+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]\\
&\qquad\leq\inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1\circ d_1(f_1(x_t),y_t)+KTL\alpha+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]+\mathbb{E}\left[\sqrt{2TK|B|\ln(\frac{2}{\alpha})}\right].
\end{aligned}
$$

where the last inequality follows from the fact that that $|\mathcal{E}_B|\leq(\frac{2}{\alpha})^{K|B|}$ and the definition of $f^\star$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2TK|B|\ln(\frac{2}{\alpha})}\right]\leq\sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right]$$
$$\leq \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right] + \sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}$$

Next, by Lemma 4, there exists a concave sublinear function $\overline{R}_{\mathcal{A}}(|B|)$ of $|B|$ that upper-bounds $R_{\mathcal{A}}(|B|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}_{\mathcal{A}}(|B|)] \leq \overline{R}_{\mathcal{A}}(T^\beta)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right]$$
$$\leq \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t) + KTL\alpha + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + \sqrt{2T^{\beta+1}K\ln(\frac{2}{\alpha})}.$$

Picking $\alpha = \frac{1}{KTL}$ and $\beta \in (0,1)$, gives that $\mathcal{Q}$ enjoys sublinear expected regret:

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi_1 \circ d_1(\mathcal{Q}(x_t), y_t)\right] - \inf_{f_1\in\mathcal{F}_1}\sum_{t=1}^{T}\psi_1 \circ d_1(f_1(x_t), y_t)$$
$$\leq 1 + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + \sqrt{4T^{\beta+1}K\ln(KTL)}.$$

This completes the proof as we have shown that $\mathcal{Q}$ is an online learner for $\mathcal{F}_1$ with respect to $\psi_1 \circ d_1$.

■

2.4.4.2 A More General Characterization of Learnability for Decomposable Losses

Theorem 11 characterizes the learnability of multioutput function classes $\mathcal{F}$ with respect to decomposable loss functions ℓ in terms of the learnability of $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$. Similar to the batch setting, we can remove ψ_k, and characterize the learnability of $\mathcal{F}$ with respect to ℓ in terms of the learnability of $\mathcal{F}_k$'s with respect to d_1. However, to do so, we need to an place additional assumption on the decomposable loss function ℓ. Theorem 12 below summarizes the main result of this section.

Theorem 12. *Let ℓ be any decomposable loss function satisfying Assumptions 1 and 2. A multioutput function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to d_1.*

The main tool needed to prove Theorem 12 is Lemma 6, which relates the online learnability of a scalar-output function class $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ with respect to $\psi \circ d_1$ to its online learnability with respect to d_1, where ψ is any monotonic, Lipschitz function such that $\psi(0) = 0$. The proof of Lemma 6 can be found in Appendix A.8.

Lemma 6. *Let $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be any monotonic and Lipschitz function such that $\psi(0) = 0$. A scalar-valued function class $\mathcal{G} \subset [0,1]^{\mathcal{X}}$ is online learnable with respect to $\psi \circ d_1$, if and only if $\mathcal{G}$ is online learnable with respect to d_1.*

Note that for every $1 \leq p < \infty$ and $x \geq 0$, $\psi(x) = x^p$ is a monotonic increasing, Lipschitz function. Therefore, Lemma 6 shows that online learnability with respect to d_p is equivalent to online learnability with respect to d_1. Combining Assumption 2, Theorem 11, and Lemma 6 immediately gives Theorem 12. Since the Sequential Fat Shattering dimension characterizes online learnability with respect to the absolute loss, Theorem 12 further implies that for any decomposable loss satisfying Assumptions 1 and 2, the finiteness of $\mathrm{fat}_\gamma^{\mathrm{seq}}(\mathcal{F}_k)$ for all $k \in [K]$ and $\gamma > 0$ is a sufficient and necessary condition for online multioutput learnability.

2.4.4.3 Characterizing Learnability of Non-Decomposable Losses

In this section, we characterize the online learnability of multioutput function classes $\mathcal{F}$ for a natural family of non-decomposable losses, ℓ_p norms for $1 \leq p \leq \infty$. We prove an analogous theorem to Theorem 5, by relating the online learnability of $\mathcal{F}$ with respect to ℓ_p to the online learnability of each $\mathcal{F}_k$ with respect to d_1. We note that the proof of only uses the fact that ℓ_p norms are equivalent (up to a K dependent constant) to the ℓ_1 norm. Since any two norms in a finite dimensional space are equivalent, Theorem 13 actually holds true for *any* norm in $\mathbb{R}^K$. But we only consider ℓ_p norms here due to their practical importance.

Theorem 13. *Let $1 \leq p \leq \infty$. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ_p if and only if each $\mathcal{F}_k \subseteq \mathcal{Y}_k^{\mathcal{X}}$ is online learnable with respect to d_1.*

By Theorem 11, $\mathcal{F}$ is online learnable with respect to ℓ_1 if and only if each restriction $\mathcal{F}_k$ is online learnable with respect to d_1. Thus, to prove Theorem 13 it suffices to show that $\mathcal{F}$ is online learnable with respect to ℓ_p if and only if $\mathcal{F}$ is online learnable with respect to ℓ_1 for $p > 1$. At a high-level, the proof Theorem 13 follows a similar route as the proof of Theorem 5: convert an agnostic learner for $\mathcal{F}$ into a realizable learner for $\mathcal{F}^\alpha$ and then use realizable-to-agnostic conversion for $\mathcal{F}^\alpha$.

Proof. Fix $p > 1$. By the argument above, it suffices to show that $\mathcal{F}$ is online learnable with respect to ℓ_p if and only if $\mathcal{F}$ is online learnable with respect to ℓ_1. We begin by proving

sufficiency - if $\mathcal{F}$ is online learnable with respect to ℓ_1 then $\mathcal{F}$ is online learnable with respect to ℓ_p.

Let $\mathcal{A}$ be an online learner for $\mathcal{F}$ with respect to ℓ_1. Our goal is to construct an online learner $\mathcal{Q}$ for $\mathcal{F}$ with respect to ℓ_p. We assume an oblivious adversary, and therefore the stream of points to be observed by the online learner $\mathcal{Q}$, denoted $(x_1, y_1), ..., (x_T, y_T) \in (\mathcal{X} \times [0,1]^K)^T$, is fixed beforehand. Let $f^\star = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^T \ell_p(f(x_t), y_t)$ also denote the optimal function in hindsight with respect to the ℓ_p loss.

Our strategy follows three steps. First, we show that $\mathcal{A}$ is a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_p. Then, since $|\text{im}(\mathcal{F}^\alpha)| \leq (\frac{2}{\alpha})^K < \infty$ is finite and ℓ_p is a 1-subadditive (by triangle inequality), Theorem 6 allows to convert the realizable online learner $\mathcal{A}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p into an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p. Finally, for an appropriately selected discretization parameter α, we show that $\mathcal{Q}$ is also an agnostic online for $\mathcal{F}$ with respect to ℓ_p, which completes the proof. To that end, let $(x_1, f^{\star,\alpha}(x_1)), ..., (x_T, f^{\star,\alpha}(x_T)$ denote a (realizable) sequence of instances labeled by some function $f^{\star,\alpha} \in \mathcal{F}^\alpha$. Since $||\cdot||_q \leq ||\cdot||_p$ for $q > p$, we have that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \mathbb{E}\left[\sum_{t=1}^T \ell_1(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right]$$

Since $\mathcal{A}$ is an online learner for $\mathcal{F}$ with respect to ℓ_1, we get,

$$\mathbb{E}\left[\sum_{t=1}^T \ell_1(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell_1(f(x_t), f^{\star,\alpha}(x_t)) + R_{\mathcal{A}}(T) \leq \alpha K T + R_{\mathcal{A}}(T)$$

where $R_{\mathcal{A}}(T)$ is the regret of online learner $\mathcal{A}$. Combining things together, we have that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \alpha K T + R_{\mathcal{A}}(T),$$

showing that $\mathcal{A}$ is a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_p for a small enough α. Now, since $|\text{im}(\mathcal{F}^\alpha)| \leq (\frac{2}{\alpha})^K < \infty$ is a finite, ℓ_p is a 1- subadditive loss function, and $\mathcal{A}$ is a realizable online learner, for any $\beta \in (0,1)$, Theorem 6 gives an agnostic online learner $\mathcal{Q}$ for $\mathcal{F}^\alpha$ with respect to ℓ_p with the following regret guarantee over the original stream $(x_1, y_1), ..., (x_T, y_T)$:

$$\mathbb{E}\left[\sum_{t=1}^T \ell_p(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f^\alpha \in \mathcal{F}^\alpha} \sum_{t=1}^T \ell_p(f^\alpha(x_t), y_t) + \frac{T}{T^\beta}(\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)) + K\sqrt{2T^{\beta+1} K \ln(\frac{2}{\alpha})},$$

where $\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)$ is any concave sublinear upperbound of $\alpha K T^\beta + R_{\mathcal{A}}(T^\beta)$. We also use the fact that the function $T \mapsto \alpha K T^\beta$ is a concave sublinear function of T and the sum of two concave functions is itself a concave function. Noting that $\ell_p(f^\alpha(x_t), y_t) \leq \ell_p(f(x_t), y_t) + \ell_p(f^\alpha(x_t), f(x_t)) \leq \ell_p(f(x_t), y_t) + \alpha K$, gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \ell_p(\mathcal{Q}(x_t), y_t)\right] \\
&\leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_p(f(x_t), y_t) + \alpha K T + \frac{T}{T^\beta}(\alpha K T^\beta + \overline{R}_{\mathcal{A}}(T^\beta)) + K\sqrt{2T^{\beta+1} K \ln(\frac{2}{\alpha})}.
\end{aligned}$$

Combining like terms together, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_p(\mathcal{Q}(x_t), y_t)\right] \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_p(f(x_t), y_t) + 2\alpha K T + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1} K \ln(\frac{2}{\alpha})}.$$

Finally, picking $\alpha = \frac{1}{2KT}$ gives that

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_p(\mathcal{Q}(x_t), y_t)\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_p(f(x_t), y_t) \leq 1 + \frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1} K \ln(4KT)}.$$

Since $\overline{R}_{\mathcal{A}}(T^\beta)$ is sublinear in T^β and $\beta \in (0, 1)$, $\mathcal{Q}$ enjoys sublinear expected regret. Thus, we have shown that $\mathcal{Q}$ is also an agnostic online learner for $\mathcal{F}$ with respect to ℓ_p.

The reverse direction follows identically and uses the fact that for any $p > 1$, $||\cdot||_p \leq ||\cdot||_1 \leq K||\cdot||_p$. In particular, using the exact same argument, we can show that if $\mathcal{A}$ is an online learner for $\mathcal{F}$ with respect to ℓ_p, then $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}^\alpha$ with respect to ℓ_1 with expected regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_1(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq K\mathbb{E}\left[\sum_{t=1}^{T} \ell_p(\mathcal{A}(x_t), f^{\star,\alpha}(x_t))\right] \leq \alpha K^2 T + K R_{\mathcal{A}}(T).$$

Using $\mathcal{A}$ as the realizable learner in Theorem 6, for any $\beta \in (0, 1)$, picking $\alpha = \frac{1}{(K+K^2)T}$ gives a regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_1(\mathcal{Q}(x_t), y_t)\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_1(f(x_t), y_t) \leq 1 + \frac{KT}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta) + K\sqrt{2T^{\beta+1} K \ln(4K^2T)},$$

where $\overline{R}_{\mathcal{A}}(T^\beta)$ is any concave sublinear upperbound of $R_{\mathcal{A}}(T^\beta)$. Since $\beta \in (0,1)$, $\mathcal{Q}$ is an online learner for $\mathcal{F}$ with respect to ℓ_1 as needed. This completes the proof of Theorem 13. ■

Remark. As with decomposable losses, Theorem 13 also implies that for any ℓ_p norm loss, the finiteness of $\text{fat}_\gamma^{\text{seq}}(\mathcal{F}_k)$, for all $k \in [K]$ and fixed $\gamma > 0$, is a sufficient and necessary condition for online multioutput learnability.

2.5 Discussion

In this chapter, we give a characterization of multioutput learnability in four settings: batch classification, online classification, batch regression, and online regression. In all four settings, we show that a multioutput function class is learnable if and only if each restriction is learnable. All of our bounds in this paper scale with K, preventing our current techniques from extending to the case when K is infinite. Chapter 5 discusses partial results for the regime $K \to \infty$ for linear function classes, while Chapter 4 shows how similar characterizations can be extended to general target spaces and loss functions in the online setting. Extending such characterizations to the case $K \to \infty$ in the batch setting remains an open problem.

CHAPTER 3

Online Learning with Set-Valued Feedback

In this chapter[1], we study a variant of online multiclass classification where the learner predicts a single label but receives a *set of labels* as feedback. Recall that, in the standard online multiclass classification setting, a learner plays a repeated game against an adversary. In each round $t \in [T]$, the adversary picks a labeled example $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals the unlabeled example x_t to the learner. The learner observes x_t and then makes a prediction $\hat{y}_t \in \mathcal{Y}$. Finally, the adversary reveals the true label y_t and the learner suffers the loss $\mathbb{1}\{\hat{y}_t \neq y_t\}$ [Littlestone, 1987, Daniely et al., 2011].

In practice, however, there may not be a single correct label $y \in \mathcal{Y}$, but rather, a *collection* of correct labels $S \subseteq \mathcal{Y}$. For example, in online multilabel ranking, the learner is tasked with ranking a set of labels in terms of their relevance to an instance. However, as feedback, the learner only receives a bitstring indicating which of the labels were relevant. This feedback model is standard in multilabel ranking since obtaining the full ranking is generally costly [Liu et al., 2009]. Since, for any given bitstring, there can be multiple rankings that correctly place relevant labels above non-relevant labels, the learner effectively only observes a *set* of correct rankings. Beyond ranking, other notable examples of set-valued feedback include multilabel classification with a thresholded Hamming loss, where the learner is only penalized after misclassifying a certain number of labels, and real-valued prediction where the response is an interval on the real line [Diamond, 1990, Gil et al., 2002, Huber et al., 2009]. Even more generally, one can equivalently represent the ground truth label as a collection of elements from the prediction space for any learning problem with the 0-1 loss where there is an asymmetry between the prediction and label space.

Motivated by online multilabel ranking and other natural learning problems, we study a variant of online multiclass classification where in each round $t \in [T]$, the learner still predicts a single label $\hat{y}_t \in \mathcal{Y}$, but the adversary reveals a set of correct labels $S_t \in \mathcal{S}(\mathcal{Y})$,

[1]This chapter is based on: Vinod Raman*, Unique Subedi*, and Ambuj Tewari (2024). *Online Learning with Set-Valued Feedback*. Conference on Learning Theory (COLT).

where $\mathcal{S}(\mathcal{Y}) \subseteq 2^{\mathcal{Y}}$ is an arbitrary set system. The learner suffers a loss if and only if $\hat{y}_t \notin S_t$. Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions such that its regret, the difference between its cumulative loss and the cumulative loss of the best-fixed hypothesis in hindsight, is small. The class $\mathcal{H}$ is said to be online learnable if there exists an online learning algorithm whose regret is a sublinear function of the time horizon T.

Given a learning problem $(\mathcal{X}, \mathcal{Y}, \mathcal{S}(\mathcal{Y}), \mathcal{H})$, what are necessary and sufficient conditions for $\mathcal{H}$ to be online learnable? For example, under single-label feedback (multiclass classification), the online learnability of a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is characterized by the finiteness of a combinatorial parameter called the Littlestone dimension [Littlestone, 1987, Ben-David et al., 2009, Daniely et al., 2011]. Analogously, is there a combinatorial parameter that characterizes online learnability under set-valued feedback? Motivated by these questions, we make the following contributions.

(1) We show that under set-valued feedback, deterministic and randomized learnability are *not equivalent* even in the realizable setting. This is in contrast to online learning with single-label feedback, where there is no separation between deterministic and randomized realizable learnability [Littlestone, 1987, Daniely et al., 2011]. Additionally, we show deterministic and randomized realizable learnability are equivalent if the *Helly number*, a parameter that arises in combinatorial geometry, of $\mathcal{S}(\mathcal{Y})$ is finite.

(2) In light of this separation, we give two new combinatorial dimensions, the Set Littlestone and Measure shattering dimension, and show that they characterize deterministic and randomized realizable learnability respectively.

(3) Moving beyond the realizable setting, we show that the Measure Shattering dimension continues to characterize *agnostic* learnability. This implies an equivalence between randomized realizable learnability and agnostic learnability.

(4) Finally, as applications, we use our results to bound the minimax expected regret for three practical learning settings: online multilabel ranking, online multilabel classification, and real-valued prediction with interval-valued response.

To prove the separation in (1), we identify a learning problem where every deterministic learner fails, but there exists a simple randomized learner. As for our combinatorial dimensions in (2), the Set Littlestone and Measure shattering dimensions are defined using complete trees with *infinite-width*. This is in contrast to much of the existing combinatorial dimensions in online learning. To prove that the Set Littlestone dimension is sufficient for deterministic realizable learnability, we extend the Standard Optimal Algorithm for single-label to set-valued feedback. On the other hand, to prove that the Measure shattering dimension

is sufficient for randomized realizable learnability, we adapt the recent algorithmic chaining technique from Daskalakis and Golowich [2022]. Lastly, our construction of an agnostic learner in (3) uses a non-trivial extension of the adaptive covering technique introduced in Hanneke et al. [2023].

3.1 Related Works

There is a rich history of characterizing online learnability in terms of combinatorial dimensions. For example, Littlestone [1987], Ben-David et al. [2009] proved that the Littlestone dimension characterizes online learnability in binary classification. Studying optimal randomized learnability, Filmus et al. [2023] proposed the Randomized Littlestone and showed that it characterizes optimal regret bounds for randomized learners in the realizable setting. Daniely et al. [2011], Hanneke et al. [2023] show that the Littlestone dimension continues to characterize online learnability in the multiclass classification setting. Recent work by Moran, Sharon, Tsubari, and Yosebashvili [2023] showed that a modification of the Littlestone dimension characterizes *list online classification*, the "flip" of our setting where the learner outputs a set of labels, but the adversary reveals a single label. In addition, Daniely and Helbertal [2013] showed that the Bandit Littlestone dimension characterizes online learnability when the adversary can output a set of correct labels, however, the learner only observes the indication of whether their predicted label was in the set or not. Moreover, there is a growing literature on online multiclass learning with feedback graphs [van der Hoeven et al., 2021, Alon et al., 2015]. In this setting, the learner predicts a single label but observes the losses of a specific set of labels determined by an arbitrary directed feedback graph. Finally, the Helly number Helly [1923] has previously been used to characterize proper learning in both online and PAC settings [Hanneke et al., 2021, Bousquet et al., 2020] and has also appeared in the literature on distributed learning [Kane et al., 2019].

3.1.1 Relation to List Online Classification

List online classification, studied by Moran et al. [2023], is intimately related to online classification with set-valued feedback. Indeed, online classification with set-valued feedback is equivalent to a modified list online classification game, where in each round $t \in [T]$: (1) the learner picks a label $\hat{y}_t \in \mathcal{Y}$ and constructs a list $\hat{L}_t \subset \mathcal{S}(\mathcal{Y})$ such that $\hat{y}_t \in S$ for every $S \in \hat{L}_t$, (2) Nature reveals the true set $S_t \in S(\mathcal{Y})$, and (3) the learner suffers the loss $\mathbb{1}\{S_t \notin \hat{L}_t\} \geq \mathbb{1}\{\hat{y}_t \notin S_t\}$. However, there are important differences between this "modified" list online classification game and the "original" list online classification game proposed by

Moran et al. [2023] when taking $S(\mathcal{Y})$ to be the label space. First, in the "original" list online classification game, the learner is allowed to output *any* finite list of elements in $S(\mathcal{Y})$. This is not the case with the "modified" list online classification game. Indeed, the "modified" list online learner is required to pick any sequence of elements in $S(\mathcal{Y})$ whose sequence-wise intersection is not empty. This means that the "modified" list online classification game can be harder than the "original" list online classification game, for example, when $S(\mathcal{Y})$ contains all disjoint sets. On the other hand, the "original" list online classification game can also be harder than the "modified" list online classification game, for example, when $\bigcap_{S \in S(\mathcal{Y})} S \neq \emptyset$. These statements are true even when the sets $S_t \in S(\mathcal{Y})$ are all finite. Therefore, the "modified" and "original" list online classification game with label space $S(\mathcal{Y})$ are incomparable.

3.2 Preliminaries

3.2.1 Notation

Let $\mathcal{X}$ denote the instance space and $(\mathcal{Y}, \sigma(\mathcal{Y}))$ be a measurable label space. Let $\Pi(\mathcal{Y})$ denote the set of all probability measures on $(\mathcal{Y}, \sigma(\mathcal{Y}))$. In this paper, we consider the case where $\mathcal{Y}$ can be unbounded (e.g. $\mathcal{Y} = \mathbb{N}$). Given a measurable label space $(\mathcal{Y}, \sigma(\mathcal{Y}))$, let $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ denote an arbitrary, measurable collection of subsets of $\mathcal{Y}$. For any set $S \in \mathcal{S}(\mathcal{Y})$, we let $S^c = \mathcal{Y} \setminus S$ denote its complement. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote an arbitrary hypothesis class consisting of predictors $h : \mathcal{X} \to \mathcal{Y}$. Finally, we let $[N] := \{1, 2, \ldots, N\}$.

3.2.2 Online Learning

In the online setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, an adversary selects a labeled instance $(x_t, S_t) \in \mathcal{X} \times \mathcal{S}(\mathcal{Y})$ and reveals x_t to the learner. The learner makes a potentially randomized prediction $\hat{y}_t \in \mathcal{Y}$. Finally, the adversary reveals the set S_t, and the learner suffers the loss $\mathbb{1}\{\hat{y}_t \notin S_t\}$. Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the goal of the learner is to output predictions $\hat{y}_t$ such that its cumulative loss is close to the best possible cumulative loss over hypotheses in $\mathcal{H}$. Before we define online learnability, we provide formal definitions of deterministic and randomized online learning algorithms.

Definition 5 (Deterministic Online Learner)**.** *A deterministic online learner is a deterministic mapping $\mathcal{A} : (\mathcal{X} \times \mathcal{S}(\mathcal{Y}))^\star \times \mathcal{X} \to \mathcal{Y}$ that maps past examples and the newly revealed instance $x \in \mathcal{X}$ to a label $y \in \mathcal{Y}$.*

Definition 6 (Randomized Online Learner). *A randomized online learner is a deterministic mapping* $\mathcal{A} : (\mathcal{X} \times \mathcal{S}(\mathcal{Y}))^\star \times \mathcal{X} \to \Pi(\mathcal{Y})$ *that maps past examples and the newly revealed instance* $x \in \mathcal{X}$ *to a probability distribution* $\hat{\mu} \in \Pi(\mathcal{Y})$. *The learner then randomly samples a label* $\hat{y} \sim \hat{\mu}$ *to make a prediction.*

We typically use $\mathcal{A}(x)$ to denote the prediction of $\mathcal{A}$ on x. When $\mathcal{A}$ is randomized, we use $\mathcal{A}(x)$ to denote the random sample $\hat{y}$ drawn from the distribution that $\mathcal{A}$ outputs.

A hypothesis class is said to be online learnable if there exists an online learning algorithm, either deterministic or randomized, whose (expected) cumulative loss, on any sequence of labeled examples, $(x_1, S_1), ..., (x_T, S_T)$, is not too far from that of best-fixed hypothesis in hindsight.

Definition 7 (Online Agnostic Learnability). *A hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ *is online learnable in the agnostic setting if there exists a (potentially randomized) algorithm* $\mathcal{A}$ *such that its expected regret*

$$\mathrm{R}_\mathcal{A}(T, \mathcal{H}) := \sup_{(x_1,S_1),...,(x_T,S_T)} \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}(x_t) \notin S_t\} \right] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \notin S_t\} \right)$$

is a non-decreasing, sub-linear function of T.

A sequence of labeled examples $\{(x_t, S_t)\}_{t=1}^T$ is said to be *realizable* by $\mathcal{H}$ if there exists a hypothesis $h^\star \in \mathcal{H}$ such that $h^\star(x_t) \in S_t$ for all $t \in [T]$. In such case, we have $\inf_{h \in \mathcal{H}} \sum_{t=1}^T \mathbb{1}\{h(x_t) \notin S_t\} = 0$.

Definition 8 (Online Realizable Learnability). *A hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ *is online learnable in the realizable setting if there exists a (potentially randomized) algorithm* $\mathcal{A}$ *such that its expected number of mistakes*

$$\mathrm{M}_\mathcal{A}(T, \mathcal{H}) := \sup_{\substack{(x_1,S_1),...,(x_T,S_T) \\ \exists h^\star \in \mathcal{H} \text{ such that } h^\star(x_t) \in S_t}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}(x_t) \notin S_t\} \right]$$

is a non-decreasing, sub-linear function of T.

One may analogously define a slightly restricted notion of deterministic realizable learnability by restricting the algorithm $\mathcal{A}$ to be deterministic.

3.3 Combinatorial Dimensions

In online learning theory, combinatorial dimensions are often defined in terms of *trees*, a basic unit that captures temporal dependence. Accordingly, we start this section by formally

defining the notion of a tree.

Given an instance space $\mathcal{X}$ and a (potentially uncountable) set of objects $\mathcal{M}$, an $\mathcal{X}$-valued, $\mathcal{M}$-ary tree $\mathcal{T}$ of depth T is a complete rooted tree such that each internal node v is labeled by an instance $x \in \mathcal{X}$ and for every internal node v and object $m \in \mathcal{M}$, there is an outgoing edge e_v^m indexed by m. We can mathematically represent this tree by a sequence $(\mathcal{T}_1, ..., \mathcal{T}_T)$ of labeling functions $\mathcal{T}_t : \mathcal{M}^{t-1} \to \mathcal{X}$ which provide the labels for each internal node. A path of length T down the tree is given be a sequence of objects $m = (m_1, ..., m_T) \in \mathcal{M}^T$. Then, $\mathcal{T}_t(m_1, ..., m_{t-1})$ gives the label of the node by following the path $(m_1, ..., m_{t-1})$ starting from the root node, going down the edges indexed by the m_t's. We let $\mathcal{T}_1 \in \mathcal{X}$ denote the instance labeling the root node. For brevity, we define $m_{<t} = (m_1, ..., m_{t-1})$ and therefore write $\mathcal{T}_t(m_1, ..., m_{t-1}) = \mathcal{T}_t(m_{<t})$. Analogously, we let $m_{\leq t} = (m_1, ..., m_t)$.

Often, it is useful to label the edges of a tree with some *auxiliary* information. Given an $\mathcal{X}$-valued, $\mathcal{M}$-ary tree $\mathcal{T}$ of depth T and a (potentially uncountable) set of objects $\mathcal{N}$, we can formally label the edges of $\mathcal{T}$ using objects in $\mathcal{N}$ by considering a sequence $(f_1, ..., f_T)$ of edge-labeling functions $f_t : \mathcal{M}^t \to \mathcal{N}$. For each depth $t \in [T]$, the function f_t takes as input a path $m_{\leq t}$ of length t and outputs an object in $\mathcal{N}$. Accordingly, we can think of the object $f_t(m_{\leq t})$ as labeling the edge indexed by m_t after following the path $m_{<t}$ down the tree. We now use this notation to rigorously define existing combinatorial dimensions in online learning.

We begin with the Littlestone dimension, which is known to characterize binary/multiclass online classification, where $\mathcal{S}(\mathcal{Y}) = \{\{y\} : y \in \mathcal{Y}\}$.

Definition 9 (Littlestone dimension [Littlestone, 1987, Daniely et al., 2011])**.** *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued , $\{\pm 1\}$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a sequence $(f_1, ..., f_d)$ of edge-labeling functions $f_t : \{\pm 1\}^t \to \mathcal{Y}$ such that for every path $\sigma = (\sigma_1, ..., \sigma_d) \in \{\pm 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$, $h_\sigma(\mathcal{T}_t(\sigma_{<t})) = f_t(\sigma_{\leq t})$ and $f_t((\sigma_{<t}, -1)) \neq f_t((\sigma_{<t}, +1))$. The Littlestone dimension of $\mathcal{H}$, denoted $\mathrm{L}(\mathcal{H})$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exists shattered trees of arbitrarily large depth, we say $\mathrm{L}(\mathcal{H}) = \infty$.*

A natural extension of the Littlestone dimension to set-valued feedback is to (1) replace the two differing labels on the edges of the Littlestone tree with two disjoint sets in $\mathcal{S}(\mathcal{Y})$ and (2) require that for every path down the tree, there is a hypothesis whose outputs on the sequence of instances lie inside the sets labeling the sequence of edges. In fact, one can even consider trees with more than two outgoing edges. Such combinatorial structures have been previously studied to characterize online learnability under bandit feedback [Daniely and Helbertal, 2013] and list classification [Moran et al., 2023].

Along this direction, Definition 10 considers complete trees where each internal node has p outgoing edges. Each outgoing edge is labeled by a set in $\mathcal{S}(\mathcal{Y})$ with the additional constraint that the mutual intersection of the p sets labeling the p edges has to be empty. Finally, such a $[p]$-ary is shattered if for every root-to-leaf path down the tree, there exists a hypothesis whose outputs on the sequence of instances lie in the sets labeling the edges along the sequence.

Definition 10 (p-Set Littlestone dimension)**.** *Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued, $[p]$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a sequence $(f_1, ..., f_d)$ of edge-labeling set-valued functions $f_t : [p]^t \to \mathcal{S}(\mathcal{Y})$ such that for every path $q = (q_1, ..., q_d) \in [p]^d$, we have $\bigcap_{i\in[p]} f_t((q_{<t}, i)) = \emptyset$ and there exists a hypothesis $h_q \in \mathcal{H}$ such that $h_q(\mathcal{T}_t(q_{<t})) \in f_t(q_{\leq t})$ for all $t \in [d]$,. The p-Set Littlestone dimension of $\mathcal{H}$ denoted $\mathrm{SL}_p(\mathcal{H}, \mathcal{S}(\mathcal{Y}))$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exists shattered trees of arbitrarily large depth, we say $\mathrm{SL}_p(\mathcal{H}, \mathcal{S}(\mathcal{Y})) = \infty$.*

When it is clear from context, we drop the dependence of $\mathcal{S}(\mathcal{Y})$ and only write $\mathrm{SL}_p(\mathcal{H})$. Note that if $p_1 > p_2$, then $\mathrm{SL}_{p_1}(\mathcal{H}) \geq \mathrm{SL}_{p_2}(\mathcal{H})$. It is not too hard to see that the finiteness of $\mathrm{SL}_p(\mathcal{H})$ for every $p \geq 2$ is a necessary condition for online learnability. For many natural problems (see Theorem 14 and Section 3.6), the finiteness of $\mathrm{SL}_p(\mathcal{H})$ for every $p \geq 2$ is also sufficient for online learnability. However, Example 1 shows that the finiteness of $\mathrm{SL}_p(\mathcal{H})$ for every $p \geq 2$ is actually not sufficient.

Example 1. *Let $\mathcal{Y} = \mathbb{N}$, $\mathcal{S}(\mathcal{Y}) = \{A^c : A \subset \mathbb{N}, |A| < \infty\}$, and suppose $\mathcal{H} = \{x \mapsto y : y \in \mathcal{Y}\}$ is the class of constant functions. First, we claim that $\mathrm{SL}_p(\mathcal{H}) = 0$ for all $p \geq 2$. Fix $p \geq 2$ and let $S_1, ..., S_p \in \mathcal{S}(\mathcal{Y})$ denote an arbitrary sequence of p sets. For each $i \in [p]$, let A_i be the finite set such that $S_i = A_i^c$. Then, $\bigcap_{i=1}^p S_i = \bigcap_{i=1}^p A_i^c = (\bigcup_{i=1}^p A_i)^c \neq \emptyset$ since $|\bigcup_{i=1}^p A_i| < \infty$. Thus, $\mathrm{SL}_p(\mathcal{H}) = 0$ because it is not possible to find p sets in $\mathcal{S}(\mathcal{Y})$ whose mutual intersection is empty. Since p is arbitrary, this is true for every $p \geq 2$. Next, we claim that $\mathcal{H}$ is not online learnable. This follows from the fact that for every $\varepsilon \in [0, 1]$ and measure $\mu \in \Pi(\mathcal{Y})$, there exists a finite set $A_\mu \subset \mathbb{N}$ such that $\mu(A_\mu) \geq \varepsilon$. Suppose for the sake of contradiction this is not true. That is, there exists an $\varepsilon \in [0, 1]$ and a measure $\mu_\varepsilon \in \Pi(\mathcal{Y})$ such that for all finite sets $A \subset \mathbb{N}$, we have $\mu_\varepsilon(A) < \varepsilon$. For every $i \in \mathbb{N}$, let $N_i = \{1, 2, ..., i\}$ denote the first i natural numbers. Note that $\mu_\varepsilon(N_i) < \varepsilon$ and that $\{N_i\}_{i\in\mathbb{N}}$ is a monotone increasing sequence of finite sets such that $\lim_{i\to\infty} N_i = \mathbb{N}$. Therefore, we have that $1 = \mu_\varepsilon(\mathbb{N}) = \mu_\varepsilon(\lim_{i\to\infty} N_i) = \lim_{i\to\infty} \mu_\varepsilon(N_i) < \varepsilon$, a contradiction. Accordingly, for any $\varepsilon \in [0, 1]$, no matter what measure $\hat{\mu}_t$ the algorithm picks to make its prediction in round t, there always exists a finite set $A_{\hat{\mu}_t}$ such that $\hat{\mu}_t(A_{\hat{\mu}_t}) \geq \varepsilon$. Since $|A_{\hat{\mu}_t}| < \infty$, we know that $A_{\hat{\mu}_t}^c \in \mathcal{S}(\mathcal{Y})$. Thus, there is always a strategy for the adversary to force the learner's expected*

loss to be at least ε in each round $t \in [T]$. On the other hand, since for any sequence of sets $S_1, ..., S_T \in \mathcal{S}(\mathcal{Y})$, we have that $\cap_{t=1}^{T} S_t \neq \emptyset$, there exists a hypothesis $h_y \in \mathcal{H}$ such that $h_y(x) \in S_t$ for all $x \in \mathcal{X}$ and $t \in [T]$. Thus, every stream is realizable by $\mathcal{H}$. Accordingly, for every $\varepsilon \in [0, 1]$, the expected regret of any online learner in the realizable setting is at least εT.

Example 1 shows that, in full generality, one might need to go beyond trees with finite width in order to characterize online learnability with set-valued feedback. Using this observation, we define two new combinatorial dimensions, the Set Littlestone and Measure shattering dimension, whose associated trees can have infinite-width. In Section 3.4, we show that the Set Littlestone dimension (SLdim) tightly characterizes the online learnability of $\mathcal{H}$ by any *deterministic* online learner in the *realizable* setting.

Definition 11 (Set Littlestone dimension)**.** *Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued, $\mathcal{Y}$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a sequence $(f_1, ..., f_d)$ of edge-labeling set-valued functions $f_t : \mathcal{Y}^t \to \mathcal{S}(\mathcal{Y})$ such that for every path $y = (y_1, ..., y_d) \in \mathcal{Y}^d$, we have $y_t \notin f_t(y_{\leq t})$ and there exists a hypothesis $h_y \in \mathcal{H}$ such that $h_y(\mathcal{T}_t(y_{<t})) \in f_t(y_{\leq t})$ for all $t \in [d]$. The Set Littlestone dimension of $\mathcal{H}$, denoted $\mathrm{SL}(\mathcal{H}, \mathcal{S}(\mathcal{Y}))$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exists shattered trees of arbitrarily large depth, we say $\mathrm{SL}(\mathcal{H}, \mathcal{S}(\mathcal{Y})) = \infty$.*

On the other hand, we show that the Measure Shattering dimension (MSdim) characterizes the online learnability of $\mathcal{H}$ by any *randomized* online learner in both the realizable and agnostic settings under set-valued feedback. We note that the Measure Shattering dimension is similar to the sequential fat-shattering dimension in the sense that it is a *scale-sensitive*, and therefore defined at every $\gamma > 0$.

Definition 12 (Measure Shattering dimension)**.** *Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued, $\Pi(\mathcal{Y})$-ary tree of depth d, and fix $\gamma \in (0, 1]$. The tree $\mathcal{T}$ is γ-shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a sequence $(f_1, ..., f_d)$ of edge-labeling set-valued functions $f_t : \Pi(\mathcal{Y})^t \to \mathcal{S}(\mathcal{Y})$ such that for every path $\mu = (\mu_1, ..., \mu_d) \in \Pi(\mathcal{Y})^d$, we have $\mu_t(f_t(\mu_{\leq t})) \leq 1 - \gamma$ and there exists a hypothesis $h_\mu \in \mathcal{H}$ such that $h_\mu(\mathcal{T}_t(\mu_{<t})) \in f_t(\mu_{\leq t})$ for all $t \in [d]$. The Measure Shattering dimension of $\mathcal{H}$ at scale γ, denoted $\mathrm{MS}_\gamma(\mathcal{H}, \mathcal{S}(\mathcal{Y}))$, is the maximal depth of a tree $\mathcal{T}$ that is γ-shattered by $\mathcal{H}$. If there exists γ-shattered trees of arbitrarily large depth, we say $\mathrm{MS}_\gamma(\mathcal{H}, \mathcal{S}(\mathcal{Y})) = \infty$. Analogously, we can define $\mathrm{MS}_0(\mathcal{H}, \mathcal{S}(\mathcal{Y}))$ by requiring strict inequality, $\mu_t(f_t(\mu_{\leq t})) < 1$.*

As with most scale-sensitive dimensions, MSdim has a monotonicity property, namely, $\mathrm{MS}_{\gamma_1}(\mathcal{H}) \leq \mathrm{MS}_{\gamma_2}(\mathcal{H})$ for any $\gamma_2 \leq \gamma_1$. This follows immediately upon noting that for any $A \in \mathcal{S}(\mathcal{Y})$, we have $\mu(A) \leq 1 - \gamma_1 \leq 1 - \gamma_2$. Thus, a tree shattered at scale γ_1 is also shattered at scale γ_2.

3.3.1 Relations Between Combinatorial Dimensions

In this section, we show how the p-SLdim, SLdim, and MSdim are related under various conditions on the problem setting. One natural case is when the set system $\mathcal{S}(\mathcal{Y})$ has finite *Helly number*, a quantification of the following property: every collection-wise disjoint sequence of sets in $\mathcal{S}(\mathcal{Y})$ contains a *small* collection-wise disjoint subsequence of sets.

Definition 13 (Helly Number of $\mathcal{S}(\mathcal{Y})$)**.** *The Helly number of $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$, denoted $\mathrm{H}(\mathcal{S}(\mathcal{Y}))$, is the smallest number $p \in \mathbb{N}$ such that for any collection of sets $\mathcal{C} \subseteq \mathcal{S}(\mathcal{Y})$ where $\bigcap_{S \in \mathcal{C}} S = \emptyset$, there is a subset $\mathcal{C}' \subset \mathcal{C}$ of size at most p where $\bigcap_{S \in \mathcal{C}'} S = \emptyset$.*

We say that $\mathcal{S}(\mathcal{Y})$ is a Helly space if and only if $\mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$. The Helly property captures many practical learning settings. For example, when $\mathcal{Y}$ is finite, any collection $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ is a Helly space. However, Helly spaces are more general and capture situations where $\mathcal{Y}$ can be uncountably large. For example, if $\mathcal{Y} = [0, 1]$ and $\mathcal{S}(\mathcal{Y}) = \{[a, b] : 0 \leq a < b \leq 1\}$ is the set of all intervals in $\mathcal{Y}$, then the celebrated Helly's theorem states that $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$ [Radon, 1921]. In Section 3.6, we give even more examples of natural settings where $\mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$. In this work, we use the Helly number of $\mathcal{S}(\mathcal{Y})$ to establish a relationship between the combinatorial dimensions defined above.

Theorem 14 (Structural Properties)**.** *For $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, we have*

(i) $\mathrm{SL}_p(\mathcal{H}) \leq \mathrm{MS}_\gamma(\mathcal{H}) \leq \mathrm{SL}(\mathcal{H})$ *for all* $p \geq 2$ *and* $\gamma \in [0, \frac{1}{p}]$.

(ii) If $p = \mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$, *then* $\mathrm{SL}_p(\mathcal{H}) = \mathrm{MS}_\gamma(\mathcal{H}) = \mathrm{SL}(\mathcal{H})$ *for all* $\gamma \in [0, \frac{1}{p}]$.

The proof of Theorem 14 is found in Appendix B.1. The key idea in the proof of (ii) is that when $\mathcal{S}(\mathcal{Y})$ is a Helly space, we can "compress" the infinite-width trees in the definition of SLdim and MSdim to finite-width trees used in the definition of p-SLdim. Perhaps the most important implication of these relations is that when $\mathcal{S}(\mathcal{Y})$ is a Helly family, deterministic and randomized realizable learnability are equivalent and characterized by the same dimension. Thus, as we show in Section 3.4.1, the separation between randomized and deterministic realizable learnability only occurs when $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = \infty$. We leave it as an open question whether the finiteness of $\mathrm{H}(\mathcal{S}(\mathcal{Y}))$ is necessary for this equivalence.

3.4 Realizable Setting

3.4.1 A Separation Between Deterministic and Randomized Learnability

We first show that unlike in online multiclass learning with single-label feedback, deterministic and randomized learnability are not equivalent under set-valued feedback. We note that Hanneke and Yang [2023], Hanneke et al. [2021] show a similar separation in the context of bandit learnability and proper online learnability.

Theorem 15 (Deterministic Learnability $\not\equiv$ Randomized Learnability)**.** *There exists a $\mathcal{Y}$, $\mathcal{S}(\mathcal{Y})$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that in the realizable setting (i) $\mathcal{H}$ is online learnable, however (ii) no deterministic algorithm is an online learner for $\mathcal{H}$.*

Proof. Let $\mathcal{Y} = \mathbb{N}$ and $\mathcal{S}(\mathcal{Y}) = \{A_y\}_{y \in \mathcal{Y}}$ where $A_y = \mathbb{N} \setminus y$. Let $\mathcal{H} = \{h_y : y \in \mathbb{N}\}$ be the set of constant functions. That is, $h_y(x) = y$ for all $x \in \mathcal{X}$.

Let $\mathcal{A}$ be any deterministic online learner for $\mathcal{H}$ and $T \in \mathbb{N}$ be the time horizon. We construct a realizable stream of length T such that $\mathcal{A}$ makes a mistake on each round. Without loss of generality, we let the adversary play after $\mathcal{A}$ since $\mathcal{A}$ is deterministic. To that end, pick any sequence of instances $\{x_t\}_{t=1}^T \in \mathcal{X}^T$ and consider the labeled stream $\{(x_t, A_{\mathcal{A}(x_t)})\}_{t=1}^T$, where $\mathcal{A}(x_t)$ denotes the prediction of $\mathcal{A}$ in the t'th round. By definition of A_y, we have $\sum_{t=1}^T \mathbb{1}\{\mathcal{A}(x_t) \notin A_{\mathcal{A}(x_t)}\} = T$. Moreover, since T is finite, it also holds that $\bigcap_{t=1}^T A_{\mathcal{A}(x_t)} \neq \emptyset$. Thus, there exists $h_y \in \mathcal{H}$ such that for all $t \in [T]$, $h_y(x_t) \in A_{\mathcal{A}(x_t)}$, showing that the stream $\{(x_t, A_{\mathcal{A}(x_t)})\}_{t=1}^T$ is indeed realizable. Since $\mathcal{A}$ is arbitrary, every deterministic algorithm fails to learn $\mathcal{H}$ under set-valued feedback from $\mathcal{S}(\mathcal{Y})$.

We now give a randomized online learner for $\mathcal{H}$ that achieves sub-linear regret for any sequence of instances labeled by sets from $\mathcal{S}(\mathcal{Y})$. Let $\{(x_t, S_t)\}_{t=1}^T \in (\mathcal{X} \times \mathcal{S}(\mathcal{Y}))^T$ denote the stream of instances to be observed by the randomized online learner. Consider a randomized learner $\mathcal{A}$ that in each round samples uniformly from $\{1, ..., T\}$. Then, $\mathcal{A}$'s expected cumulative loss satisfies

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{A}(x_t) \notin S_t\}\right] = \sum_{t=1}^T \mathbb{P}\left[\mathcal{A}(x_t) \notin S_t\right] = \sum_{t=1}^T \mathbb{P}\left[S_t = A_{\mathcal{A}(x_t)}\right] \leq \sum_{t=1}^T \frac{1}{T} = 1,$$

where we have used the fact that $\mathcal{A}(x_t) \notin S_t$ iff the adversary exactly picks the set $S_t = A_{\mathcal{A}(x_t)}$. Thus, $\mathcal{A}$ achieves a constant regret bound, showcasing that it is an online learner for $\mathcal{H}$ under set-valued feedback from $\mathcal{S}(\mathcal{Y})$. This completes the overall proof as we have given a learning setting that is online learnable, but not by any deterministic learner. ■

3.4.2 Deterministic Learnability

Given that deterministic and randomized online learnability are not generally equivalent, we show that the SLdim tightly characterizes *deterministic* online learnability in the realizable setting.

Theorem 16 (Deterministic Realizable Learnability)**.** *For any* $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ *and* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$*, we have* $\inf_{\text{Deterministic}\mathcal{A}} \mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) = \mathrm{SL}(\mathcal{H})$.

Our proof of the upperbound on the optimal $\mathrm{M}_{\mathcal{A}}(T, \mathcal{H})$ is constructive. We show that Algorithm 7 makes at most $\mathrm{SL}(\mathcal{H})$ mistakes in any realizable stream by generalizing the arguments by Littlestone [1987]. To prove the lowerbound on $\mathrm{M}_{\mathcal{A}}(T, \mathcal{H})$ for any deterministic algorithm $\mathcal{A}$, we construct a difficult stream by traversing the shattered tree of depth $\mathrm{SL}(\mathcal{H})$ adapting to $\mathcal{A}$'s predictions. Both proofs can be found in Appendix B.2.

Algorithm 7 Deterministic Standard Optimal Algorithm

Initialize $V_0 = \mathcal{H}$
for $t = 1, \ldots, T$ **do**
  Receive unlabeled example $x_t \in \mathcal{X}$.
  For each $A \in \mathcal{S}(\mathcal{Y})$, define $V_{t-1}(A) := \{h \in V_{t-1} \mid h(x_t) \in A\}$.
  Let $\mathcal{S}_t(\mathcal{Y}) := \{A \in \mathcal{S}(\mathcal{Y}) : A \cap \{h(x_t) \mid h \in V_{t-1}\} \neq \emptyset\}$.
  if $\mathrm{SL}(V_{t-1}) > 0$ **then**
    Predict $\hat{y}_t = \arg\min_{y \in \mathcal{Y}} \max_{\substack{A \in \mathcal{S}(\mathcal{Y}) \\ y \notin A}} \mathrm{SL}(V_{t-1}(A))$.
  else
    Predict $\hat{y}_t \in \bigcap_{A \in \mathcal{S}_t(\mathcal{Y})} A$.
  end if
  Receive feedback $S_t \in \mathcal{S}_t(\mathcal{Y})$ and update $V_t = V_{t-1}(S_t)$.
end for

Remark. We highlight that Algorithm 7 generalizes the classical Standard Optimal Algorithm. In fact, if $\mathcal{S}(\mathcal{Y}) = \{\{y\} : y \in \mathcal{Y}\}$ then Algorithm 7 reduces exactly to the classical Standard Optimal Algorithm from Littlestone [1987] and SLdim reduces to the Ldim. Moreover, when $\mathcal{S}(\mathcal{Y}) = \{\mathcal{Y} \setminus \{y\} : y \in \mathcal{Y}\}$, Algorithm 7 reduces to the Bandit Standard Optimal Algorithm from Daniely et al. [2011] and SLdim reduces to the Bandit Littlestone dimension.

3.4.3 Randomized Learnability

Next, we characterize randomized online learnability in the realizable setting. The proof of Theorem 17 can be found in Appendix B.3.

Theorem 17 (Randomized Realizable Learnability). *For any $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$,*

$$\sup_{\gamma \in (0,1]} \gamma \, \mathrm{MS}_\gamma(\mathcal{H}) \leq \inf_{\mathcal{A}} \mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) \leq C \inf_{\gamma \in (0,1]} \left\{ \gamma T + \int_\gamma^1 \mathrm{MS}_\eta(\mathcal{H}) d\eta \right\}$$

where $C > 0$ is some universal constant. Moreover, both the upper and lowerbounds can be tight in general up to constant factors.

Using Theorem 14, it follows that $\mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) = \Theta(\mathrm{SL}(\mathcal{H}))$ whenever $\mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$. We highlight that the upperbound can be tight up to logarithmic factors in T. If $\mathcal{S}(\mathcal{Y})$ is a set of singletons, then we have $\mathrm{MS}_0(\mathcal{H}) = \mathrm{L}(\mathcal{H})$. Thus, the upperbound reduces to $\mathrm{L}(\mathcal{H})$, which matches the known lowerbound of $\mathrm{L}(\mathcal{H})/2$ in the realizable multiclass classification [Daniely et al., 2011]. Example 2 shows that the lowerbound of $\sup_{\gamma>0} \gamma \, \mathrm{MS}_\gamma(\mathcal{H})$ can be tight in the realizable setting.

To achieve our upperbound, we first construct a randomized online learner running at a fixed scale $\gamma \in (0, 1)$, whose expected cumulative loss, in the realizable setting, is at most $\gamma T + \mathrm{MS}_\gamma(\mathcal{H})$. Then, we upgrade this result by adapting the algorithmic chaining technique from Daskalakis and Golowich [2022] to give a randomized, *multi-scale* online learner in the realizable setting. Our lowerbound is obtained by traversing the tree of depth $\mathrm{MS}_\gamma(\mathcal{H})$ adapting to the distributions that the algorithm produces to make its randomized predictions.

We conclude this section by showing that the Helly number of $\mathcal{S}(\mathcal{Y})$ is a sufficient condition for deterministic and randomized learnability to be equivalent in the realizable setting. Corollary 1 follows directly upon using Theorems 14(ii), 16, and 17.

Corollary 1 (Deterministic Learnability $\equiv$ Randomized Learnability for Helly Families). *Let $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ such that $\mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$. Then, in the realizable setting, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable via a randomized algorithm if and only if $\mathcal{H}$ is online learnable via a deterministic algorithm.*

3.5 Agnostic Setting

In this section, we move beyond the realizable setting, and consider the more general agnostic setting, where we are not guaranteed that there exists a consistent hypothesis. Our main theorem shows that the finiteness of MSdim at every scale $\gamma > 0$ is both a necessary and sufficient condition for agnostic online learnability with set-valued feedback.

Theorem 18 (Agnostic Learnability). *For any $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ and hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ where $\sup_{\gamma\in(0,1]} \mathrm{MS}_\gamma(\mathcal{H}) > 0$, we have*

$$\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T,\mathcal{H}) \geq \max\left\{\sqrt{\frac{\mathrm{SL}_2(\mathcal{H})\,T}{8}}, \sup_{\gamma\in(0,1]} \gamma\ \mathrm{MS}_\gamma(\mathcal{H})\right\}$$
$$\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T,\mathcal{H}) \leq \inf_{\gamma\in(0,1]} \left\{\mathrm{MS}_\gamma(\mathcal{H}) + \gamma T + \sqrt{2\ \mathrm{MS}_\gamma(\mathcal{H})\,T\ln(T)}\right\}.$$

Here, the upper and lowerbounds can be tight in general up to constant factors. Moreover, when $\sup_{\gamma\in(0,1]} \mathrm{MS}_\gamma(\mathcal{H}) = 0$, there is no non-negative lowerbound.

Using Theorem 14, it follows that $\mathrm{R}_{\mathcal{A}}(T,\mathcal{H}) = \tilde{\Theta}(\sqrt{T})$ whenever $\mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$ and $\mathrm{SL}(\mathcal{H}) < \infty$. We highlight that the upper bound can be tight up to logarithmic factors in T. If $\mathcal{S}(\mathcal{Y})$ is a set of singletons, then we have $\mathrm{MS}_0(\mathcal{H}) = \mathrm{L}(\mathcal{H})$. Thus, the upper bound reduces to $\mathrm{L}(\mathcal{H}) + \sqrt{2\ \mathrm{L}(\mathcal{H})\,T\ln(T)}$, which matches the known lower bound of $\sqrt{\mathrm{L}(\mathcal{H})\,T/8}$ in the agnostic multiclass classification [Daniely et al., 2011]. The following example shows that the lower bound cannot be improved in general.

Example 2. *Let $\mathcal{Y} = \{1,2,3,4,5,6\}$, $\mathcal{S}(\mathcal{Y}) = \{\{1,4,5\},\{2,5,6\},\{3,4,6\}\}$, and $\mathcal{H} = \{h_1,h_2,h_3\}$, where again h_i is the hypothesis that always outputs i. Let $d = \mathrm{SL}_2(\mathcal{H})$ and $d_\gamma = \mathrm{MS}_\gamma(\mathcal{H})$. Since there are no disjoint sets in $\mathcal{S}(\mathcal{Y})$, we trivially have $d = 0$, reducing the lowerbound to $\gamma\, d_\gamma$. First, we prove that $\sup_\gamma \gamma d_\gamma = \frac{1}{3}$. This follows from the fact that $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 3$, and therefore, by Theorem 14, for all $\gamma \in [0,\frac{1}{3}]$ we have $d_\gamma = \mathrm{SL}(\mathcal{H}) = 1$. Moreover, by the monotonicty property of MSdim, $d_\gamma \leq d_{\frac{1}{3}} = 1$ for all $\gamma > \frac{1}{3}$. Thus, it must be the case $\sup_{\gamma>0} \gamma\, d_\gamma = \frac{1}{3}$.*

Now, we give a randomized online learner whose expected regret is at most $\sup_{\gamma>0} \gamma d_\gamma = \frac{1}{3}$ on the worst-case sequence, matching the lowerbound. Consider an online learner $\mathcal{A}$, which on the round $t = 1$ predicts by uniformly sampling from $\{4,5,6\}$, and on all other rounds predicts by uniformly sampling from $\{4,5,6\} \cap S_{t-1}$, where S_{t-1} is the set revealed by the adversary on round $t-1$. Our goal will be to show that $\mathcal{A}$'s expected regret on any sequence is at most $\frac{1}{3}$. Let $\{(x_t,S_t)\}_{t=1}^T$ denote the stream chosen by the adversary. Then, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\mathcal{A}(x_t) \notin S_t\}\right] &= \frac{1}{3} + \sum_{t=2}^T \mathbb{E}\left[\mathbb{1}\{\mathcal{A}(x_t) \notin S_t\} | S_t \neq S_{t-1}\right] \mathbb{1}\{S_t \neq S_{t-1}\} \\ &= \frac{1}{3} + \frac{1}{2}\sum_{t=2}^T \mathbb{1}\{S_t \neq S_{t-1}\},\end{aligned}$$

where the first equality follows from the fact that $\mathbb{E}\left[\mathbb{1}\{\mathcal{A}(x_1) \notin S_1\}\right] = \frac{1}{3}$ and $\mathbb{E}\left[\mathbb{1}\{\mathcal{A}(x_t) \notin S_t\} \,|\, S_t = S_{t-1}\right] = 0$. Moreover, we can lowerbound the expected cumulative

loss of the best fixed hypothesis as

$$\min_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\{h(x_t)\notin S_t\}=\min_{i\in[3]}\sum_{t=1}^{T}\mathbb{1}\{i\notin S_t\}\geq\frac{1}{2}\sum_{t=2}^{T}\mathbb{1}\{S_t\neq S_{t-1}\}$$

Combining the upper- and lowerbound gives that $\mathrm{R}_{\mathcal{A}}(T,\mathcal{H})\leq\frac{1}{3}$.

Remark. An important implication of Theorem 18 is that when $\mathrm{H}(\mathcal{S}(\mathcal{Y}))=2$, a lowerbound scaling with T is always possible. However, Example 2 above witnessing the tightness of the lowerbounds in Theorem 18 shows that this is not the case when $\mathrm{H}(\mathcal{S}(\mathcal{Y}))\geq 3$. Thus, a sharp phase transition occurs when $\mathrm{H}(\mathcal{S}(\mathcal{Y}))$ increases from 2 to 3.

3.6 Applications

In this section, we show how online multilabel ranking with relevance-score feedback and online multilabel classification are special instances of our model of online learning with set-valued feedback. In Appendix B.5, we also consider real-valued prediction with interval-valued response.

3.6.1 Online Multilabel Ranking

In online multilabel ranking, we let $\mathcal{X}$ denote the instance space, $\mathcal{Y}$ denote the set of permutations over labels $[K]:=\{1,...,K\}$, and $\mathcal{R}=\{0,1\}^K$ denote the target space for some $K\in\mathbb{N}$. We refer to an element $r\in\mathcal{R}$ as a *binary relevance-score vector* that indicates the relevance of each of the K labels. A permutation $\pi\in\mathcal{Y}$ induces a *ranking* of the K labels in decreasing order of relevance. For an index $i\in[K]$, we let $\pi^i\in[K]$ denote the *rank* of label i. Likewise, given an index $i\in[K]$, we let r^i denote the relevance of label i. A ranking hypothesis $h\in\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$ maps instances in $\mathcal{X}$ to a permutation (ranking) in $\mathcal{Y}$. Given an instance $x\in\mathcal{X}$, one can think of $h(x)$ as h's ranking of the K different labels in decreasing order of relevance.

Unlike classification, a distinguishing property of multilabel ranking is the *mismatch* between the predictions the learner makes and the feedback it receives. Because of this mismatch, there is no canonical loss in multilabel ranking like the 0-1 loss in classification. Nevertheless, a natural analog of the 0-1 loss in multilabel ranking is $\ell_{0\text{-}1}(\pi,r)=\sup_{i,j\in[K]}\mathbb{1}\{r^i<r^j\}\,\mathbb{1}\{\pi^i<\pi^j\}$. At a high-level, the 0-1 ranking loss penalizes a permutation π if it ranks a less relevant item above a more relevant item.

Under the 0-1 loss, online multilabel ranking with binary relevance-score feedback is a specific instance of our general online learning model with set-valued feedback. To see this,

note that given a relevance score vector $r \in \mathcal{R}$, there can be many permutations $\pi \in \mathcal{Y}$ such that $\ell_{0\text{-}1}(\pi, r) = 0$. Indeed, suppose $r = (0, 1, 1)$. Then, both the permutations $\pi_1 = (3, 1, 2)$ and $\pi_2 = (3, 2, 1)$ achieve 0 loss. Thus, an *equivalent* way of representing $r = (0, 1, 1)$ is to consider the set of permutations in $\mathcal{Y}$ for which $\ell_{0\text{-}1}(\pi, r) = 0$. To this end, given any $r \in \mathcal{R}$, let $\mathcal{Y}(r) = \{\pi \in \mathcal{Y} : \ell_{0\text{-}1}(\pi, r) = 0\}$. Then, note that for every $\pi \in \mathcal{Y}$ and $r \in \mathcal{R}$, we have $\ell_{0\text{-}1}(\pi, r) = \mathbb{1}\{\pi \notin \mathcal{Y}(r)\}$. From this perspective, we can equivalently define the online multilabel ranking setting by having the adversary in each round $t \in [T]$, reveal a *set* $\mathcal{Y}(r_t) \in \{\mathcal{Y}(r) : r \in \mathcal{R}\} = \mathcal{S}(\mathcal{Y})$ instead of the binary relevance score vector $r_t \in \mathcal{R}$, and penalizing the learner according to the 0-1 *set loss* $\mathbb{1}\{\pi_t \notin \mathcal{Y}(r_t)\}$, instead of $\ell_{0\text{-}1}(\pi, r)$.

Since online multilabel ranking is a specific instance of our general online learning with set-valued feedback, our qualitative characterization in terms of the SLdim and MSdim carry over. Thus, in this section, we instead focus on establishing a sharp quantitative characterization of online learnability. To do so, we first show that $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$. The proof of Lemma 7 is deferred to Appendix B.5.1.

Lemma 7 (Helly Number of Permutation Sets)**.** *Let* $\mathcal{S}(\mathcal{Y}) = \{\mathcal{Y}(r) : r \in \mathcal{R}\}$ *where* $\mathcal{Y}(r) = \{\pi \in \mathcal{Y} : \ell_{0\text{-}1}(\pi, r) = 0\}$*. Then,* $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$*.*

Since $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$, by Theorem 14, we know that for all $\gamma \in [0, \frac{1}{2}]$, $\mathrm{SL}_2(\mathcal{H}) = \mathrm{MS}_\gamma(\mathcal{H}) = \mathrm{SL}(\mathcal{H})$. Therefore, the $\mathrm{SL}_2(\mathcal{H})$ characterizes both deterministic and randomized online multilabel ranking learnability. Moreover, we can use Theorems 16, 17, and 18 to give Corollary 2, a sharp quantitative characterization of online multilabel ranking learnability in both the realizable and agnostic settings.

Corollary 2 (Online Learnability of Multilabel Ranking)**.** *Let* $\mathcal{Y}$*,* $\mathcal{R}$*, and* $\mathcal{S}(\mathcal{Y})$ *be defined as above. For any ranking hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *we have*

(i) $\frac{\mathrm{SL}_2(\mathcal{H})}{2} \leq \inf_{\mathcal{A}} \mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) \leq \mathrm{SL}_2(\mathcal{H})$*.*

(ii) $\sqrt{\frac{\mathrm{SL}_2(\mathcal{H})\, T}{8}} \leq \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}) \leq \mathrm{SL}_2(\mathcal{H}) + \sqrt{2\, \mathrm{SL}_2(\mathcal{H})\, T \ln(T)}$*.*

We note that the infimum in Corollary 2(i) is over all algorithms, not just deterministic ones. Also, observe that the upper- and lowerbounds in Corollary 2 do not depend on $|\mathcal{Y}|$ or $|\mathcal{R}|$.

3.6.2 Online Multilabel Classification

In online multilabel *classification*, we let $\mathcal{X}$ denote the instance space, and $\mathcal{Y} = \{0, 1\}^K$ is the set of all bit strings of length $K \in \mathbb{N}$. Unlike multilabel ranking, instead of predicting a permutation over $[K]$, the goal is to predict $\hat{y} \in \mathcal{Y}$, which indicates which of the labels are

relevant. As feedback, the learner also receives a bit string $y \in \mathcal{Y}$ which gives the ground truth on which of the K labels are relevant. A multilabel hypothesis $h \in \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ maps instances in $\mathcal{X}$ to a bit string in $\mathcal{Y}$.

The most natural loss in multilabel classification is the Hamming loss, defined by $\ell_H(\hat{y}, y) = \sum_{i=1}^{K} \mathbb{1}\{\hat{y}^i \neq y^i\}$. However, when K is very large, evaluating performance using the Hamming loss might be too stringent. Instead, it might be more natural to consider a thresholded version of the Hamming loss, defined as $\ell_{H,q}(\hat{y}, y) = \mathbb{1}\{\ell_H(\hat{y}, y) > q\} = \mathbb{1}\{\hat{y} \notin \mathcal{B}(y, q)\}$, where $q \in [K-1]$ and $\mathcal{B}(y, q) = \{\hat{y} \in \mathcal{Y} : \ell_H(\hat{y}, y) \leq q\}$ denotes the Hamming ball of radius q centered at y. The loss $\ell_{H,q}$ allows the learner's prediction $\hat{y}$ to be incorrect in at most q different spots before penalizing the learner. By taking $\mathcal{Y} = \{0,1\}^K$ and $\mathcal{S}_q(\mathcal{Y}) = \{\mathcal{B}(y, q) : y \in \mathcal{Y}\}$, it is not hard to see that online multilabel classification with $\ell_{H,q}$ is a specific instance of our general online learning model with set-valued feedback. Thus, a quantitative characterization of online multilabel classification in terms of $\mathrm{SL}(\mathcal{H})$ and $\mathrm{MS}_\gamma(\mathcal{H})$ follows immediately from Theorems 16 and 18. The precise statement is provided in Appendix B.5.3.

In multilabel ranking, we showed that the 2-SLdim, provides a tight quantitative characterization of online learnability without any dependence on K. Such a characterization in terms of the 2-SLdim, as opposed to SLdim or MSdim, is desirable because it satisfies the Finite Character Property [Ben-David et al., 2019, Definition 4]. A crucial step in doing so was showing that the Helly number of the permutation set system is 2, and more importantly, does not scale with K. Along this direction, it is natural to ask whether there exists a $p \in \mathbb{N}$ such that the p-SLdim gives a K-free quantitative characterization of online multilabel classification under $\ell_{H,q}$. To resolve this question positively it suffices to show that $\mathrm{H}(\mathcal{S}_q(\mathcal{Y}))$ does not scale with K, as conjectured below.

Conjecture 1 (Helly Number of Hamming Balls)**.** *For any $K \in \mathbb{N}$ and $q \in [K-1]$, we have that $\mathrm{H}(\mathcal{S}_q(\mathcal{Y})) = 2^{q+1}$.*

In Appendix B.5.3, we partially resolve this conjecture by showing $2^{q+1} \leq \mathrm{H}(\mathcal{S}_q(\mathcal{Y})) \leq \sum_{r=0}^{q} \binom{K}{r} + 1$. We leave it as an open question to prove a matching upperbound.

Remark. This conjecture was resolved by Alon, Jin, and Sudakov [2024], who proved a matching upper bound of 2^{q+1} after the publication of our work [Raman, Subedi, and Tewari, 2024a], where these results originally appeared. We keep the conjecture here in its originally stated form for historical reference.

CHAPTER 4

A Unified Theory of Supervised Online Learnability

In this chapter[1], we study the online learnability of hypothesis classes with respect to arbitrary, but bounded loss functions. The results in this chapter, in some sense, unify and generalize four decades of work in online learning theory dating back to Littlestone [1987]. To formally discuss our results, let us first setup the relevant problem of supervised online learning. In this setting, a learner plays a repeated game against an adversary over $T \in \mathbb{N}$ rounds. In each round $t \in [T]$, an adversary picks a labeled example $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner observes x_t, picks a probability measure μ_t over the prediction space $\mathcal{Z}$, and then makes a randomized prediction $z_t \sim \mu_t$. Finally, the adversary reveals the true label y_t and the learner suffers the loss $\ell(y_t, z_t)$, where $\ell : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ is some pre-specified, bounded loss function. For a hypothesis class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ known apriori to the learner, the goal of the learner is to make predictions such that its expected regret, defined as the difference between the expected cumulative loss of the learner's predictions and that of the best-fixed hypothesis in $\mathcal{H}$, is small. We say that a tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ is *online learnable* if there exists an online learner such that its expected regret is a sublinear function of T, for any strategy of the adversary.

Given a tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ one is often interested in answering the following question:
What are necessary and sufficient conditions for $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ to be online learnable?

For instance, when $\mathcal{Z} = \mathcal{Y}$ and $\ell(y, z) = \mathbb{1}\{y \neq z\}$, online learnability has been characterized in terms of the Littlestone dimension of $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, henceforth denoted as $\mathrm{L}(\mathcal{H})$. That is, $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ is online learnable if and only if $\mathrm{L}(\mathcal{H}) < \infty$ [Littlestone, 1987, Daniely et al., 2011, Hanneke et al., 2023]. Similarly, when $\mathcal{Z} = \mathcal{Y} = [-1, 1]$ and $\ell(y, z) = |y - z|$, the *sequential* fat-shattering dimension of $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, denoted $\mathrm{sfat}_\gamma(\mathcal{H})$, characterizes the online learnability of $\mathcal{H}$. A class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ is online learnable if and only if $\mathrm{sfat}_\gamma(\mathcal{H}) < \infty$ at every scale $\gamma > 0$ [Rakhlin et al., 2015a]. Analogous dimensions for ranking and list learning have also been

[1]This chapter is based on: Vinod Raman*, Unique Subedi*, and Ambuj Tewari (2025). *A Unified Theory of Supervised Online Learnability.* Conference on Algorithmic Learning Theory (ALT).

established and shown to characterize online learnability in their respective settings [Raman et al., 2024a, Moran et al., 2023].

Existing characterizations of online learnability follow three steps. First, one identifies a combinatorial parameter, like the Littlestone or sequential fat-shattering dimension, whose finiteness provides an obvious *necessary* condition. Then, one shows that the finiteness of such a dimension is sufficient for online learnability under a suitable notion of *realizability*, where one places an assumption on the label-generating process. This step involves constructing a learning algorithm that computes the combinatorial dimension as a subroutine. These two steps were first outlined in the seminal work by Littlestone [1987]. Finally, to complete the proof of sufficiency, the realizable learner is converted into an agnostic learner using the conversion introduced by Ben-David et al. [2009]. By the end, the finiteness of the combinatorial dimension is established as both a necessary and sufficient condition for online learnability.

While this technique has been used to characterize online learnability for specific tuples, a general characterization for an *arbitrary* tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ is missing from the literature. In fact, the only known sequential complexity measure for an arbitrary learning problem is the sequential Rademacher complexity of the loss class $\ell \circ \mathcal{H} := \{(x, y) \mapsto \ell(y, h(x)) : h \in \mathcal{H}\}$. In particular, Rakhlin et al. [2015a] show that if the sequential Rademacher complexity of the loss class $\ell \circ \mathcal{H}$ is a sublinear function of T, then $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ is online learnable. However, even for natural problems like online multiclass classification [Hanneke et al., 2023] and linear regression [Raman et al., 2024b], sublinear sequential Rademacher complexity is not *necessary* for online learnability.

Our Contributions. In this chapter, we show that the previously outlined procedure for characterizing online learnability is *universal* - it works for any learning tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ as long as ℓ is bounded. In particular, we identify a new scale-sensitive combinatorial dimension, termed the Sequential Minimax dimension (SMdim), whose finiteness at every scale is an obvious necessary condition for online learnability. Then, by identifying the right notion of realizability and providing a new realizable-to-agnostic conversion, we establish that the finiteness of the SMdim is also sufficient for online learnability. Finally, and perhaps most surprisingly, we show that the SMdim reduces exactly to existing combinatorial dimensions in their respective setting. This includes the case where $\mathcal{Z} = \mathcal{Y}$, like the Littlestone and sequential fat-shattering dimensions, as well as the case where $\mathcal{Z} \neq \mathcal{Y}$, like the $(k+1)$-Littlestone dimension from Moran et al. [2023] and Measure shattering dimension from Raman et al. [2024a].

At the highest level of generality, the SMdim may not be insightful as it is an abstract combinatorial object that cannot be efficiently computed. However, given a specific learning

problem $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$, one can use this object to define more concrete combinatorial objects that provide better insight into the hardness of learning and the minimax rates. In fact, in the proof of Theorem 19, our techniques illustrate how one can use tools from discrete geometry to show that existing combinatorial dimensions are just special instances of the SMdim. Thus, beyond providing a unification of existing results in online learnability, the SMdim provides a good starting point for understanding the true complexity of a learning problem $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$.

4.1 Related Works

Characterizing learnability in terms of complexity measures has a long rich history in statistical learning theory, originating from the seminal work of Vapnik and Chervonenkis [1971]. In online learning, Littlestone [1987] showed that a combinatorial parameter, later named the Littlestone dimension, provides a quantitative characterization of online binary classification in the realizable setting. Twenty-two years later, Ben-David et al. [2009] proved that the Littlestone dimension also provides a tight quantitative characterization of online binary classification in the agnostic setting. Daniely et al. [2011] generalized the Littlestone dimension to multiclass classification and showed that it fully characterizes online learnability when the label space is finite. Recently, Hanneke et al. [2023] proved that the multiclass extension of the Littlestone dimension characterizes multiclass learnability under the 0-1 loss even when the label space is unbounded. In a parallel line of work, Rakhlin et al. [2015a,b] defined the sequential fat-shattering dimension and showed that it tightly characterizes the online learnability of scalar-valued regression with respect to the absolute value loss. In addition, they defined a general complexity measure called the sequential Rademacher complexity and proved that it upper bounds the minimax expected regret of any supervised online learning game. In a similar spirit, we define a *combinatorial dimension* that upper and lower bounds the minimax expected regret of any supervised online learning game.

The proof techniques in online learning are generally constructive and result in beautiful algorithms such as Follow The (Regularized) Leader, Hedge, Multiplicative Weights, Online Gradient Descent, and so forth. In online binary classification, Littlestone [1987] proposed the Standard Optimal Algorithm and proved its optimality in the realizable setting. Daniely et al. [2011] and Rakhlin et al. [2015a] generalize this algorithm to multiclass classification and scalar-valued regression respectively. The idea of the Standard Optimal Algorithm is foundational in online learning and still appears in more recent works by Moran et al. [2023], Filmus et al. [2023], and Raman et al. [2024a]. A common theme in these variants of the Standard Optimal Algorithm is their use of combinatorial dimensions to make predictions.

Similarly, Rakhlin et al. [2012a] use the sequential Rademacher complexity to directly construct a generic online learner in the agnostic setting. However, their online learner requires the sequential Rademacher complexity of the loss class to be sublinear in T, and thus does not work for arbitrary tuples $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$. Closing this gap, we define a new scale-sensitive dimension, named the Sequential Minimax dimension, and use it to give a generic online learner for any tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$.

Finally, we compare our work to the recent work by Blanchard [2022] on universal online learning for bounded losses. We highlight three main differences. First, in our setup, there exists a function class $\mathcal{F}$, and the goal of the learner is to drive the expected regret with respect to $\mathcal{F}$ to 0. In contrast, there is no function class in the work by Blanchard [2022]. Instead, the stream is labeled by some unknown measurable function and the goal is to drive the average cumulative loss to after placing some restrictions on the sequence of instances that can be chosen by the adversary. Second, we place no restrictions on the sequence of instances the adversary may reveal to the learner (the restriction is instead placed on how the stream is labeled). In contrast, Blanchard [2022] considers a collection of stochastic processes and restrict the adversary to play a sequence of instances sampled according to a process from this set. Finally, in our setup, the prediction space and label space may be different. In contrast, Blanchard [2022] only studies the case where the prediction and label space are the same.

4.2 Preliminaries

4.2.1 Notation

Let $\mathcal{X}$ denote the instance space, $\mathcal{Y}$ denote the label space, and $\mathcal{Z}$ denote the prediction space. For a sigma algebra $\sigma(\mathcal{Z})$ on the prediction space $\mathcal{Z}$, define $\Pi(\mathcal{Z})$ to be the set of all distributions on $(\mathcal{Z}, \sigma(\mathcal{Z}))$. For any set $S \in \sigma(\mathcal{Z})$, let S^c denote its complement. Let $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ denote an arbitrary hypothesis class consisting of predictors $h : \mathcal{X} \to \mathcal{Z}$ that maps an instance to a prediction. Given any prediction $z \in \mathcal{Z}$ and a label $y \in \mathcal{Y}$, we consider a loss function $\ell : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$. We put no restrictions on the loss function ℓ, except that it is bounded, $\sup_{y,z} \ell(y, z) \leq c$ for some $c \in \mathbb{R}_{>0}$. In particular, the loss can asymmetric, and therefore we reserve the first argument for the label and the second argument for the prediction. Finally, $[N] := \{1, 2, \ldots, N\}$.

4.2.2 Supervised Online Learning

In the supervised online learning setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, the adversary selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner picks a probability measure $\mu_t \in \Pi(\mathcal{Z})$ and then makes a randomized prediction $z_t \sim \mu_t$. Finally, the adversary reveals the feedback y_t, and the learner suffers the loss $\ell(y_t, z_t)$. Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, the goal of the learner is to output randomized predictions z_t such that its expected cumulative loss is close to the smallest possible cumulative loss over hypotheses in $\mathcal{H}$.

We follow the convention in online learning literature (see, e.g., Cesa-Bianchi and Lugosi [2006, Chapter 4] by defining a randomized learner as a sequence of deterministic mappings to probability distributions.

Definition 14 (Supervised Online Learning Algorithm)**.** *A supervised online learning algorithm is a deterministic mapping $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^\star \times \mathcal{X} \to \Pi(\mathcal{Z})$ that maps past examples and the newly revealed instance $x \in \mathcal{X}$ to a probability measure $\mu \in \Pi(\mathcal{Z})$. The learner then randomly samples $z \sim \mu$ to make a prediction.*

Remark 1. *Our definition of supervised online learning algorithm prevents an algorithm from using the realizations of its past predictions to make future predictions. While this may seem as a restriction at first, our upper bounds are achievable using online learning algorithms of exactly this type. Moreover, in Appendix C.1, we show that our lower bounds can be generalized to algorithms which can use past realizations of their predictions to make future plays.*

Although $\mathcal{A}$ is a deterministic mapping, the prediction $z \sim \mu$ is random. Restricting the range of $\mathcal{A}$ to be the set of Dirac measures on $\mathcal{Z}$ yields a deterministic online learner. When the context is clear, with a slight abuse of notation, we use $\mathcal{A}(x)$ to denote the random sample z drawn from the distribution that $\mathcal{A}$ outputs. We say that $\mathcal{H}$ is online learnable with respect to ℓ if there exists an online learning algorithm $\mathcal{A}$ with "small" *expected regret*:

$$\mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) := \sup_{(x_1,y_1),\ldots,(x_T,y_T)} \left(\sum_{t=1}^{T} \mathbb{E}\Big[\ell(y_t, \mathcal{A}(x_t))\Big] - \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(x_t)) \right).$$

Definition 15 (Supervised Online Learnability)**.** *A hypothesis class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) = o(T)$.*

Implicit in our definition of expected regret and online learnability is the fact that the adversary is *oblivious* – it must pick the entire sequence of examples before the game begins.

In this paper, we will always assume an oblivious adversary. That said, all our results also apply to adaptive adversaries given our definition of an online learning algorithm and the standard conversion of oblivious to adaptive regret bounds (see Exercise 4.1 in Cesa-Bianchi and Lugosi [2006]).

4.2.3 Combinatorial dimensions

In online learning theory, combinatorial dimensions play an important role in providing crisp quantitative characterizations of learnability. Formally, we define a combinatorial dimension as a function D that maps $(\mathcal{H}, \ell)$ to $\mathbb{N} \cup \{0, \infty\}$ and satisfies the following two properties: (1) $\mathcal{H}$ is online learnable with respect to ℓ if and only if $\mathrm{D}(\mathcal{H}, \ell) < \infty$ and (2) the minimax expected regret $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell)$ depends only on $\mathrm{D}(\mathcal{H}, \ell)$ and T. In particular, $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell)$ should not depend on any other property of the tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ such as $|\mathcal{Y}|$ or $|\mathcal{Z}|$. We also allow a combinatorial dimension to take a scale parameter as an input. That is, a scale-sensitive combinatorial dimension is a function D that maps $(\mathcal{H}, \ell)$ and a scale $\gamma > 0$ to $\mathbb{N} \cup \{0, \infty\}$ with the following two properties: (1) $\mathcal{H}$ is online learnable with respect to ℓ if and only if $\mathrm{D}(\mathcal{H}, \ell, \gamma) < \infty$ for every $\gamma > 0$ and (2) the minimax expected regret $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell)$ can be lower- and upper bounded in terms of T and $\mathrm{D}(\mathcal{H}, \ell, \cdot)$. Our definition of a combinatorial dimension is similar to the definition given by Ben-David et al. [2019] with two key differences. In particular, the notion of dimension given by Ben-David et al. [2019] requires $\mathrm{D}(\mathcal{H}, \ell)$ to satisfy the finite-character property (see Section 4.5), but does not require it to provide a quantitative characterization of learnability.

Nevertheless, our definition of dimension also captures all existing combinatorial dimensions in online learning theory, such as the Littlestone and sequential fat-shattering dimension. These dimensions are typically defined in terms of trees, a basic combinatorial object that captures the temporal dependence inherent in online learning. Given an instance space $\mathcal{X}$ and a (potentially uncountable) set of objects $\mathcal{M}$, a $\mathcal{X}$-valued, $\mathcal{M}$-ary tree $\mathcal{T}$ of depth T is a complete rooted tree such that (1) each internal node is labeled by an instance $x \in \mathcal{X}$ and (2) for every internal node and object $m \in \mathcal{M}$, there is an outgoing edge indexed by m. Such a tree can be identified by a sequence $(\mathcal{T}_1, \dots, \mathcal{T}_T)$ of labeling functions $\mathcal{T}_t : \mathcal{M}^{t-1} \to \mathcal{X}$ which provide the labels for each internal node. A path of length T is given by a sequence of objects $m = (m_1, \dots, m_T) \in \mathcal{M}^T$. Then, $\mathcal{T}_t(m_1, \dots, m_{t-1})$ gives the label of the node by following the path $(m_1, \dots, m_{t-1})$ starting from the root node, going down the edges indexed by the m_t's. We let $\mathcal{T}_1 \in \mathcal{X}$ denote the instance labeling the root node. For brevity, we define $m_{<t} = (m_1, \dots, m_{t-1})$ and therefore write $\mathcal{T}_t(m_1, \dots, m_{t-1}) = \mathcal{T}_t(m_{<t})$. Analogously, we let $m_{\leq t} = (m_1, \dots, m_t)$.

Often, it is useful to label the edges of a tree with some *auxiliary* information. Given a $\mathcal{X}$-valued, $\mathcal{M}$-ary tree $\mathcal{T}$ of depth T and a (potentially uncountable) set of objects $\mathcal{N}$, we can formally label the edges of $\mathcal{T}$ using objects in $\mathcal{N}$ by considering a sequence $(f_1, \dots, f_T)$ of edge-labeling functions $f_t : \mathcal{M}^t \to \mathcal{N}$. For each depth $t \in [T]$, the function f_t takes as input a path $m_{\le t}$ of length t and outputs an object in $\mathcal{N}$. Accordingly, we can think of the object $f_t(m_{\le t})$ as labeling the edge indexed by m_t after following the path $m_{<t}$ down the tree. We now use this notation to rigorously define existing combinatorial dimensions in online learning.

We start with the Littlestone dimension, which is known to characterize binary/multiclass online classification. In this setting, we take $\mathcal{Y} = \mathcal{Z}$ and $\ell(y, z) = \mathbb{1}\{y \neq z\}$.

Definition 16 (Littlestone dimension [Littlestone, 1987, Daniely et al., 2011])**.** *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued, $\{\pm 1\}$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ if there exists a sequence $(f_1, \dots, f_d)$ of edge-labeling functions $f_t : \{\pm 1\}^t \to \mathcal{Y}$ such that for every path $\sigma = (\sigma_1, \dots, \sigma_d) \in \{\pm 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$, $h_\sigma(\mathcal{T}_t(\sigma_{<t})) = f_t(\sigma_{\le t})$ and $f_t((\sigma_{<t}, -1)) \neq f_t((\sigma_{<t}, +1))$. The Littlestone dimension of $\mathcal{H}$, denoted $\mathrm{L}(\mathcal{H})$, is the maximal depth of a tree $\mathcal{T}$ that is shattered by $\mathcal{H}$. If there exists shattered trees of arbitrarily large depth, we say $\mathrm{L}(\mathcal{H}) = \infty$.*

For online regression, where we take $\mathcal{Z} = \mathcal{Y} = [-1, 1]$ and $\ell(y, z) = |y - z|$, online learnability is characterized by the sequential-fat shattering (seq-fat) dimension.

Definition 17 (Sequential fat-shattering dimension [Rakhlin et al., 2015a])**.** *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued, $\{\pm 1\}$-ary tree of depth d and fix $\gamma \in (0, 1]$. The tree $\mathcal{T}$ is γ-shattered by $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ if there exists a sequence $(f_1, \dots, f_d)$ of edge-labeling functions $f_t : \{\pm 1\}^t \to \mathcal{Y}$ such that for every path $\sigma = (\sigma_1, \dots, \sigma_d) \in \{\pm 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$, $\sigma_t(h_\sigma(\mathcal{T}_t(\sigma_{<t})) - f_t(\sigma_{\le t})) \ge \gamma$ and $f_t((\sigma_{<t}, -1)) = f_t((\sigma_{<t}, +1))$. The sequential fat-shattering dimension of $\mathcal{H}$ at scale γ, denoted $\mathrm{sfat}_\gamma(\mathcal{H})$, is the maximal depth of a tree $\mathcal{T}$ that is γ-shattered by $\mathcal{H}$. If there exists γ-shattered trees of arbitrarily large depth, we say that $\mathrm{sfat}_\gamma(\mathcal{H}) = \infty$.*

Recently, Moran et al. [2023] study list online classification, where we take $\mathcal{Z} = \{S : S \subset \mathcal{Y}, |S| \le k\}$ and $\ell(y, z) = \mathbb{1}\{y \notin z\}$. Here, they show that the $(k+1)$- Littlestone dimension, characterizes online learnability of a hypothesis class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$.

Definition 18 ($(k+1)$-Littlestone dimension [Moran et al., 2023])**.** *Let $\mathcal{T}$ be a complete, $\mathcal{X}$-valued, $[k+1]$-ary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ if there exists a sequence $(f_1, \dots, f_d)$ of edge-labeling functions $f_t : [k+1]^t \to \mathcal{Y}$ such that for every path $p = (p_1, \dots, p_d) \in [k+1]^d$, there exists a hypothesis $h_p \in \mathcal{H}$ such that for all $t \in [d]$,*

$f_t(p_{\leq t})) \in h_\sigma(\mathcal{T}_t(\sigma_{<t}))$ *and for all distinct* $i, j \in [k+1]$, $f_t((p_{<t}, i)) \neq f_t((p_{<t}, j))$. *The* $(k+1)$*-Littlestone dimension of* $\mathcal{H}$ *denoted* $\mathrm{L}_{k+1}(\mathcal{H})$, *is the maximal depth of a tree* $\mathcal{T}$ *that is shattered by* $\mathcal{H}$. *If there exists shattered trees of arbitrarily large depth, we say that* $\mathrm{L}_{k+1}(\mathcal{H}) = \infty$.

Finally, in the "flip" of list online classification, where $\mathcal{Y} \subset \sigma(\mathcal{Z})$ is some collection of measurable subsets of $\mathcal{Z}$ and $\ell(y, z) = \mathbb{1}\{z \notin y\}$, Raman et al. [2024a] show that the Measure shattering dimension characterizes online learnability of a hypothesis class $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$.

Definition 19 (Measure shattering dimension [Raman et al., 2024a])**.** *Let* $\mathcal{T}$ *be a complete* $\mathcal{X}$*-valued,* $\Pi(\mathcal{Z})$*-ary tree of depth* d, *and fix* $\gamma \in (0, 1]$. *The tree* $\mathcal{T}$ *is* γ*-shattered by* $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ *if there exists a sequence* $(f_1, \ldots, f_d)$ *of edge-labeling set-valued functions* $f_t : \Pi(\mathcal{Z})^t \to \mathcal{Y}$ *such that for every path* $\mu = (\mu_1, \ldots, \mu_d) \in \Pi(\mathcal{Z})^d$, *there exists a hypothesis* $h_\mu \in \mathcal{H}$ *such that for all* $t \in [d]$, $h_\mu(\mathcal{T}_t(\mu_{<t})) \in f_t(\mu_{\leq t})$ *and* $\mu_t(f_t(\mu_{\leq t})) \leq 1 - \gamma$. *The Measure Shattering dimension (MSdim) of* $\mathcal{H}$ *at scale* γ, *denoted* $\mathrm{MS}_\gamma(\mathcal{H}, \mathcal{Y})$, *is the maximal depth of a tree* $\mathcal{T}$ *that is* γ*-shattered by* $\mathcal{H}$. *If there exists* γ*-shattered trees of arbitrarily large depth, we say* $\mathrm{MS}_\gamma(\mathcal{H}, \mathcal{Y}) = \infty$.

4.3 A Unifying Combinatorial Dimension

Following the procedure outlined in the introduction, we begin our characterization of online learnability by defining a dimension that provides an "obvious" necessary condition. In the context of online learning, this means giving the adversary a strategy against every possible move of the learner. Since the learner plays measures in $\Pi(\mathcal{Z})$, it suffices to consider a tree where each internal node has an outgoing edge labeled by an element of $\mathcal{Y}$ for every measure in $\Pi(\mathcal{Z})$. For any prediction $\mu \in \Pi(\mathcal{Z})$ by the learner, the label on the edge associated to μ gives the element $y \in \mathcal{Y}$ that the adversary should play to force the learner to suffer a large expected loss.

Definition 20 (Sequential Minimax dimension)**.** *Let* $\mathcal{T}$ *be a complete* $\mathcal{X}$*-valued,* $\Pi(\mathcal{Z})$*-ary tree of depth* d, *and fix* $\gamma > 0$. *The tree* $\mathcal{T}$ *is* γ*-shattered by* $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ *with respect to* $\ell : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ *if there exists a sequence* $(f_1, \ldots, f_d)$ *of edge-labeling functions* $f_t : \Pi(\mathcal{Z})^t \to \mathcal{Y}$ *such that for every path* $\mu = (\mu_1, \ldots, \mu_d) \in \Pi(\mathcal{Z})^d$, *there exists a hypothesis* $h_\mu \in \mathcal{H}$ *such that for all* $t \in [d]$, $\mathbb{E}_{z \sim \mu_t}[\ell(f_t(\mu_{\leq t}), z)] \geq \ell(f_t(\mu_{\leq t}), h_\mu(\mathcal{T}_t(\mu_{<t}))) + \gamma$. *The sequential minimax dimension (SMdim) of* $\mathcal{H}$ *at scale* γ, *denoted* $\mathrm{SM}_\gamma(\mathcal{H}, \ell)$, *is the maximal depth of a tree* $\mathcal{T}$ *that is* γ*-shattered by* $\mathcal{H}$. *If there exists* γ*-shattered trees of arbitrarily large depth, we say* $\mathrm{SM}_\gamma(\mathcal{H}, \ell) = \infty$. *Analogously, we can define* $\mathrm{SM}_0(\mathcal{H}, \ell)$ *by requiring strict inequality,* $\mathbb{E}_{z \sim \mu}[\ell(f_t(\mu_{\leq t}), z)] > \ell(f_t(\mu_{\leq t}), h_\mu(\mathcal{T}_t(\mu_{<t})))$.

Remark 2. *The astute reader might notice the strong similarity between the MSdim and the SMdim. This similarity is not coincidental – the SMdim is a generalization of the MSdim designed to capture general loss functions and go beyond realizability.*

Observe that the SMdim is a function of both the hypothesis class $\mathcal{H}$ and the loss function ℓ. However, when it is clear from context, we drop the dependence of ℓ and only write $\mathrm{SM}_\gamma(\mathcal{H})$. As with most scale-sensitive dimensions, SMdim has a monotonicity property, namely, $\mathrm{SM}_{\gamma_1}(\mathcal{H}) \leq \mathrm{SM}_{\gamma_2}(\mathcal{H})$ for any $\gamma_2 \leq \gamma_1$.

In Section 4.4, we show that the finiteness of the SMdim is both necessary and sufficient for online learnability. Theorem 19 then shows that the SMdim unifies several existing results in online supervised learning.

Theorem 19 (Unifying Learnability)**.** *The following statements are true.*

(i) If $\mathcal{Y} = \mathcal{Z}$ and $\ell(y, z) = \mathbb{1}\{y \neq z\}$, then $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{L}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{2}]$.

(ii) If $\mathcal{Y} = \mathcal{Z} = [-1, 1]$ and $\ell(y, z) = |y - z|$, then $\mathrm{sfat}_\gamma(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{sfat}_{\gamma'}(\mathcal{H})$ for every $0 < \gamma' < \gamma < 1$.

(ii) If $\mathcal{Z} = \{S : S \subset \mathcal{Y}, |S| \leq k\}$ and $\ell(y, z) = \mathbb{1}\{y \notin z\}$, then $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{L}_{k+1}(\mathcal{H})$ for every $\gamma \in [0, \frac{1}{k+1}]$.

(iv) If $\mathcal{Y} \subseteq \sigma(\mathcal{Z})$ and $\ell(y, z) = \mathbb{1}\{z \notin y\}$, then $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{MS}_\gamma(\mathcal{H}, \mathcal{Y})$ for all $y \in [0, 1]$.

As an immediate consequence, Theorem 19 shows that the SMdim provides a tight quantitative characterization of online learnability for these problems. Our proof of Theorem 19, found in Appendix C.2, uses combinatorial arguments. In all four cases, our proof uses the following strategy. To show that the SMdim upper bounds the existing dimension, we take the shattered tree guaranteed by the existing dimension and for every node, use the labels on its outgoing edges to add new, labeled edges indexed by measures in $\Pi(\mathcal{Z})$. We then remove all the old edges. To show that the SMdim lower bounds the existing dimension, we take a shattered SMdim tree, and for every node, use the labels on its outgoing edges to add new, labeled edges that match the requirements of the existing dimension. Finally, we remove all the old edges indexed by measures in $\Pi(\mathcal{Z})$. In either direction, the addition of new, labeled edges requires tools from discrete geometry. For example, the proof of (ii) uses the celebrated Helly's theorem [Radon, 1921]. Thus, despite being an abstract object, the techniques used in the proof of Theorem 19 show how one can use tools from discrete geometry to derive more concrete dimensions for particular choices of $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$.

4.4 Bounding Minimax Expected Regret

Our main result in this section shows that the finiteness of SMdim at every scale $\gamma > 0$ is both necessary and sufficient for online learnability.

Theorem 20 (Minimax Expected Regret)**.** *For any* $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ *with* $\sup_{\gamma>0} \mathrm{SM}_\gamma(\mathcal{H}) > 0$*,*

$$\sup_{\gamma>0} \gamma\ \mathrm{SM}_\gamma(\mathcal{H}) \leq \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) \leq \inf_{\gamma>0} \left\{\gamma T + c\ \mathrm{SM}_\gamma(\mathcal{H}) + 4c\sqrt{\mathrm{SM}_\gamma(\mathcal{H})\, T \ln(T)}\right\}.$$

Moreover, the upper bound and lower bound can be tight up to logarithmic factors in T*.*

The upper bound in Theorem 9 is $o(T)$ as long as $\mathrm{SM}_\gamma(\mathcal{H}) < \infty$ for every $\gamma > 0$. Indeed, the average regret satisfies $\limsup_{T\to\infty} \inf_{\gamma>0} \left\{\gamma + c\ \mathrm{SM}_\gamma(\mathcal{H})/T + 4c\sqrt{\mathrm{SM}_\gamma(\mathcal{H})\ \ln(T)/T}\right\} \leq \inf_{\gamma>0} \limsup_{T\to\infty} \left\{\gamma + c\ \mathrm{SM}_\gamma(\mathcal{H})/T + 4c\sqrt{\mathrm{SM}_\gamma(\mathcal{H})\ \ln(T)/T}\right\} = \inf_{\gamma>0}\{\gamma\} = 0$, where the first equality follows because $\mathrm{SM}_\gamma(\mathcal{H}) < \infty$, $\forall \gamma > 0$.

The condition $\sup_{\gamma>0} \mathrm{SM}_\gamma(\mathcal{H}) > 0$ is necessary to ensure a non-negative lower bound. Raman et al. [2024a] provide an example of a tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ with $\sup_{\gamma>0} \mathrm{MS}_\gamma(\mathcal{H}) = 0$ where the corresponding minimax expected regret is negative. Moreover, Raman et al. [2024a, Example 5.1] provide a tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$, where there exists an algorithm $\mathcal{A}$ such that $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) \leq \sup_{\gamma>0} \gamma\ \mathrm{MS}_\gamma(\mathcal{H})$. Since $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{MS}_\gamma(\mathcal{H})$ by Theorem 19, the lower bound in Theorem 20 cannot be improved in full generality. For the tightness of the upper bound, consider scalar-valued regression where $\mathcal{Y} = \mathcal{Z} = [-1, 1]$, $\ell(y, z) = |y - z|$. Since, by Theorem 19, we have that $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{sfat}_{\gamma'}(\mathcal{H})$ for all $\gamma' < \gamma$, Theorem 20 implies that $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) \leq \inf_{\gamma>0}\{2\gamma T + 2\,\mathrm{sfat}_\gamma(\mathcal{H}) + 4\sqrt{\mathrm{sfat}_\gamma(\mathcal{H})\, T \ln(T)}\}$. However, for scalar-valued regression, Rakhlin et al. [2015a] show that $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) \geq \sup_{\gamma>0} \frac{\gamma}{8}\sqrt{\mathrm{sfat}_\gamma(\mathcal{H})T}$. Thus, the upper bound in Theorem 20 is tight up to $O(\sqrt{\ln(T)})$.

The proof of Theorem 20 will follow the procedure outlined in the introduction. Namely, the lower bound will follow just from the definition of the SMdim. As for the upper bound, in Section 4.4.2, we will first define a notion of realizability we term ϵ_t-realizability. Then, in Lemma 8, we will constructively show that the finiteness of the SMdim at every scale is sufficient for online learnability under ϵ_t-realizability. Finally, in Section 4.4.3, we will provide a conversion of our ϵ_t-realizable learner into a fully agnostic online learner with the stated upper bound in Theorem 20.

4.4.1 Proof sketch of lower bound

Our proof of the lower bound in Theorem 20 is constructive. Given an algorithm and a scale $\gamma > 0$, we construct a stream by traversing the sequential minimax tree of depth

$\mathrm{SM}_\gamma(\mathcal{H})$, adapting to the deterministic sequence of measures the algorithm uses to make its randomized prediction. Then, our claimed lower bound follows immediately from the definition of a shattered sequential minimax tree. Since the proof of the lower bound is relatively straightforward, we defer it to Appendix C.4.

4.4.2 The ε_t-realizable setting

In the ε_t-realizable setting, an adversary plays a sequential game with the learner over T rounds. In each round $t \in [T]$, the adversary selects a *thresholded* labeled instance $(x_t, (y_t, \varepsilon_t)) \in \mathcal{X} \times (\mathcal{Y} \times [0, c])$ and reveals x_t to the learner. The learner selects a measure $\mu_t \in \Pi(\mathcal{Z})$ and makes a randomized prediction $z_t \sim \mu_t$. Finally, the adversary reveals both the true label y_t and the threshold ε_t and the learner suffers the loss $\ell(y_t, z_t)$. A sequence of thresholded labeled examples $\{(x_t, (y_t, \varepsilon_t))\}_{t=1}^T$ is called ε_t-realizable if there exists a hypothesis $h^\star \in \mathcal{H}$ such that $\ell(y_t, h^\star(x_t)) \leq \varepsilon_t$ for all $t \in [T]$. Given any ε_t-realizable stream, the goal of the learner is to output predictions such that $\sum_{t=1}^T \mathbb{1}\{\mathbb{E}_{z\sim\mu_t}[\ell(y_t, z)] \geq \gamma + \varepsilon_t\}$ is sublinear in T. We can think of the thresholds ε_t as the adversary additionally revealing the loss that the best fixed hypothesis in hindsight suffers on the labeled instance (x_t, y_t). This intuition is critical to our construction of an agnostic learner in Section 4.4.3. Note that if it is guaranteed ahead of time that $\varepsilon_t = 0$ for all $t \in [T]$, then this setting boils down to the standard realizable setting. Lemma 8 shows that the finiteness of $\mathrm{SM}_\gamma(\mathcal{H})$ at every scale $\gamma > 0$ is sufficient for $\mathcal{H}$ to be online learnable in ε_t-realizable setting.

The ε-additive noise setting is a widely used model in regression. The ε in these works typically represents stochastic noise, which may be unbounded. In contrast, ε_t is always bounded in our model. To the best of our knowledge, the ε_t-realizable model has not been previously studied in the learning theory literature. However, this model may provide a more realistic framework for certain practical learning scenarios and thus be of independent theoretical interest.

Lemma 8 (ε_t-Realizable Learner)**.** *For any tuple $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$, Algorithm 8 running on any ε_t-realizable stream $\{(x_t, (y_t, \varepsilon_t))\}_{t=1}^T$ outputs $\{\mu_t\}_{t=1}^T$ such that*

$$\sum_{t=1}^{T} \mathbb{1}\Big\{\mathbb{E}_{z\sim\mu_t}[\ell(y_t, z)] \geq \gamma + \varepsilon_t\Big\} \leq \mathrm{SM}_\gamma(\mathcal{H}). \tag{4.1}$$

To prove Lemma 8, we show that (i) on any round where $\mathbb{E}_{z_t\sim\mu_t}[\ell(y_t, z_t)] \geq \gamma + \varepsilon_t$ and $\mathrm{SM}_\gamma(V_{t-1}) > 0$, we have $\mathrm{SM}_\gamma(V_t) \leq \mathrm{SM}_\gamma(V_{t-1}) - 1$, and (ii) if $\mathrm{SM}_\gamma(V_{t-1}) = 0$ there exists a distribution $\mu_t \in \Pi(\mathcal{Z})$ such that $\mathbb{E}_{z_t\sim\mu_t}[\ell(y_t, z_t)] < \gamma + \varepsilon_t$. We defer the proof to Appendix C.3.

Algorithm 8 Minimax Randomized Standard Optimal Algorithm (MRSOA)

Require: $\mathcal{H}$, Target accuracy $\gamma > 0$

Initialize $V_0 = \mathcal{H}$

for $t = 1, \ldots, T$ **do**

  Receive unlabeled example $x_t \in \mathcal{X}$.

  For all $(y, \varepsilon) \in \mathcal{Y} \times [0, c]$, define $V_{t-1}(y, \varepsilon) := \{h \in V_{t-1} \mid \ell(y, h(x_t)) \leq \varepsilon\}$.

  Define $\mathcal{C}_t := \{(y, \varepsilon) \in \mathcal{Y} \times [0, c] : |V_{t-1}(y, \varepsilon)| > 0\}$.

  if $\mathrm{SM}_\gamma(V_{t-1}) = 0$ **then**

    Pick $\mu_t \in \Pi(\mathcal{Z})$ such that $\mathbb{E}_{z \sim \mu_t}[\ell(y, z)] < \varepsilon + \gamma$ for all $(y, \varepsilon) \in \mathcal{C}_t$.

  else

    Set

$$\mu_t = \arg\min_{\mu \in \Pi(\mathcal{Z})} \max_{\substack{(y,\varepsilon) \in \mathcal{Y} \times [0,c] \\ \mathbb{E}_{z \sim \mu}[\ell(y,z)] \geq \varepsilon + \gamma}} \mathrm{SM}_\gamma(V_{t-1}(y, \varepsilon)).$$

  end if

  Predict $z_t \sim \mu_t$.

  Receive feedback (y_t, ε_t) and update $V_t = V_{t-1}(y_t, \varepsilon_t)$.

end for

Algorithm 8 can be viewed as a generalization of RSOA introduced by Raman et al. [2024a]. When $\varepsilon_t = 0$ for all t, then MRSOA reduces exactly to Algorithm 2 in Raman et al. [2024a]. At its core, Algorithm 8 is a version space algorithm based on principles similar to that of the standard optimal algorithm (SOA) of Littlestone [1987]. Recently, other variants of SOA have also been introduced in various settings. These include the Bandit SOA by Daniely and Helbertal [2013], List SOA by Moran et al. [2023], and randomized SOA by Filmus et al. [2023].

4.4.3 Realizable-to-Agnostic conversion

Now, we show how to convert Algorithm 8 into an agnostic learner satisfying the guarantee in Theorem 20. A primary approach to proving online agnostic upper bounds involves defining a set of experts that exactly covers the hypothesis class and then running multiplicative weights [Cesa-Bianchi and Lugosi, 2006] using these experts. This technique originated in work [Ben-David et al., 2009] on binary classification and was later generalized by Daniely et al. [2011] to multiclass classification. Daniely et al. [2011]'s generalization involves simulating all possible labels in $\mathcal{Y}$ to update the experts, thus making their upper bound vacuous when $|\mathcal{Y}|$ is unbounded. Recently, Hanneke et al. [2023] removed $|\mathcal{Y}|$ from the upper bound by (1) constructing an approximate cover of the hypothesis class instead of an exact cover and (2) using the feedback in the stream to update experts rather than simulating all possible labels. Our proof of the upper bound in Theorem 20 combines

the ideas of both Daniely et al. [2011] and Hanneke et al. [2023]. In particular, following Hanneke et al. [2023], we construct an approximate cover of the hypothesis class but follow Daniely et al. [2011] in simulating all possible *loss values*.

Proof. (of upper bound in Theorem 20) Let $(x_1, y_1), \ldots, (x_T, y_T)$ be the data stream and $h^\star \in \arg\min_{h\in\mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(x_t))$ be an optimal function in hind-sight. For a target accuracy $\gamma > 0$, let $d_\gamma = \mathrm{SM}_\gamma(\mathcal{H})$.

Defining Experts. Given time horizon T, let $L_T = \{L \subset [T]; |L| \leq d_\gamma\}$ denote the set of all possible subsets of $[T]$ with size at most d_γ. For $\alpha \in [0, 1]$, let $\{0, \alpha, \ldots, \lceil \frac{c}{\alpha} \rceil \alpha\}$ be an α-cover of the loss space $[0, c]$. For every $L \in L_T$, define $\Phi_L = \{0, \alpha, \ldots, \lceil \frac{c}{\alpha} \rceil \alpha\}^L$ to be the set of all functions from L to the α-cover of $[0, c]$. Given $L \in L_T$ and $\phi_L \in \Phi_L$, define an expert $E_L^{\phi_L}$ such that

$$E_L^{\phi_L}(x_t) := \mathrm{MRSOA}_\gamma\Big(x_t \mid \{i, \phi_L(i)\}_{i\in L\cap[t-1]}\Big),$$

where $\mathrm{MRSOA}_\gamma\Big(x_t \mid \{i, \phi_L(i)\}_{i\in L\cap[t-1]}\Big)$ is the prediction of the Minimax Randomized Standard Optimal Algorithm (MRSOA) running at scale γ that has updated on thresholded labeled examples $\{(x_i, (y_i, \phi_L(i))\}_{i\in L\cap[t-1]}$. Let $\mathcal{E} = \bigcup_{L\in L_T} \bigcup_{\phi_L\in\Phi_L} \{E_L^{\phi_L}\}$ denote the set of all Experts. Note that $|\mathcal{E}| = \sum_{i=0}^{d_\gamma} \left(\frac{2c}{\alpha}\right)^i \binom{T}{i} \leq \left(\frac{2cT}{\alpha}\right)^{d_\gamma}$.

Multiplicative Weights as our Agnostic Learner. Finally, given our set of experts $\mathcal{E}$, we run the Multiplicative Weights Algorithm (MWA), denoted hereinafter as $\mathcal{A}$, over the stream

$$(x_1, y_1), \ldots, (x_T, y_T)$$

with a learning rate $\eta = \sqrt{2\ln(|\mathcal{E}|)/T}$. Let B denote the random variable denoting the randomized prediction of all experts (or their corresponding randomized algorithms) . Then, conditioned on B, Theorem 21.11 of Shalev-Shwartz and Ben-David [2014] tells us that

$$\sum_{t=1}^{T} \mathbb{E}\left[\ell(y_t, \mathcal{A}(x_t)) \mid B\right] \leq \inf_{E\in\mathcal{E}} \sum_{t=1}^{T} \ell(y_t, E(x_t)) + c\sqrt{2T\ln(|\mathcal{E}|)}\,.$$

Using $|\mathcal{E}| \leq \left(\frac{2cT}{\alpha}\right)^{d_\gamma}$, and taking expectations on both sides yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(y_t, \mathcal{A}(x_t))\right] \leq \mathbb{E}\left[\inf_{E\in\mathcal{E}} \sum_{t=1}^{T} \ell(y_t, E(x_t))\right] + c\sqrt{2d_\gamma T \ln\left(\frac{2cT}{\alpha}\right)}. \tag{4.2}$$

Next, we show that the expected loss of the optimal expert is at most the loss of $h^\star$ plus a sublinear quantity.

Tracking the Best Expert. Define $\varepsilon_t := \ell(y_t, h^\star(x_t))$ to be the loss of the optimal hypothesis in hindsight on each round t. Define $\mu_t = \mu\text{-MRSOA}_\gamma\Big(x_t \mid \{i, \phi_L(i)\}_{i \in L \cap [t-1]}\Big)$ to be the measure returned by MRSOA_γ (Algorithm 8) to make its randomized prediction given that the algorithm has updated on thresholded labeled examples $\{(x_i, (y_i, \phi_L(i)))\}_{i \in L \cap [t-1]}$. We say that μ-MRSOA_γ makes a mistake on round t if $\mathbb{E}_{z_t \sim \mu_t}[\ell(y_t, z_t)] \geq \lceil \frac{\varepsilon_t}{\alpha} \rceil \alpha + \gamma$. As $\lceil \frac{\varepsilon_t}{\alpha} \rceil \alpha \geq \varepsilon_t$, the stream

$$(x_1, (y_1, \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha)), \ldots, (x_T, (y_T, \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha))$$

is $\left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha$-realizable. Thus, with this notion of the mistake, Equation 4.1 tells us that MRSOA_γ makes at most d_γ mistakes on the stream $(x_1, (y_1, \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha)), \ldots, (x_T, (y_T, \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha))$.

Since μ-MRSOA_γ is a deterministic mapping from the past examples to a probability measure in $\Pi(\mathcal{Z})$, we can recursively define a sequence of time points where μ-MRSOA_γ, had it run exactly on this sequence of time points, would make mistakes at each time point. To that end, let

$$t_1 = \min\left\{ t \in [T] : \mathop{\mathbb{E}}_{z_t \sim \mu_t}[\ell(y_t, z_t)] \geq \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha + \gamma \text{ where } \mu_t = \mu\text{-MRSOA}_\gamma\Big(x_t | \, \{\}\Big) \right\}$$

be the earliest time point, where a fresh, unupdated copy of μ-MRSOA_γ makes a mistake if it exists. Given t_1, we recursively define t_i for $i > 1$ as

$$t_i = \min\left\{ t > t_{i-1} : \mathop{\mathbb{E}}_{z_t \sim \mu_t}[\ell(y_t, z_t)] \geq \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha + \gamma, \right.$$
$$\left. \text{where } \mu_t = \mu\text{-MRSOA}_\gamma\left(x_t \middle| \left\{ t_j, \left\lceil \frac{\varepsilon_{t_j}}{\alpha} \right\rceil \alpha \right\}_{j=1}^{i-1}\right) \right\}$$

if it exists. That is, t_i is the earliest timepoint in $[T]$ after t_{i-1} where μ-MRSOA_γ having updated only on the sequence $\{(x_{t_j}, (y_{t_j}, \left\lceil \frac{\varepsilon_{t_j}}{\alpha} \right\rceil \alpha))\}_{j=1}^{i-1}$ makes a mistake. We stop this process when we reach an iteration where no such time point in $[T]$ can be found where μ-MRSOA_γ makes a mistake.

Using the definitions above, let $t_1, t_2, \ldots,$ denote the sequence of timepoints in $[T]$ selected via this recursive procedure. Define $L^\star = \{t_1, t_2 \ldots, \}$ and $\phi_{L^\star}$ be the function such that $\phi_{L^\star}(t) = \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha$ for each $t \in L^\star$. Let $E_{L^\star}^{\phi_{L^\star}}$ be the expert parametrized by the pair $(L^\star, \phi_{L^\star})$. The expert $E_{L^\star}^{\phi_{L^\star}}$ exists because Equation (4.1) implies that $|L^\star| \leq d_\gamma$.

Bounding the Loss of the Best Expert. By definition of the expert, we have

$$E_{L^\star}^{\phi_{L^\star}}(x_t) = \text{MRSOA}_\gamma\Big(x_t \mid \{i, \phi_{L^\star}(i)\}_{i \in L^\star \cap [t-1]}\Big)$$

for all $t \in [T]$. Let us define $\mu_t^\star = \mu\text{-MRSOA}_\gamma\Big(x_t \mid \{i, \phi_{L^\star}(i)\}_{i \in L^\star \cap [t-1]}\Big)$. Using the guarantee of MRSOA (Algorithm 8), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell\Big(y_t, E_{L^\star}^{\phi_{L^\star}}(x_t)\Big)\right] &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{z_t \sim \mu_t^\star}\Big[\ell\Big(y_t, z_t\Big)\Big] \\
&\le \sum_{t=1}^{T} c\, \mathbb{1}\left\{\mathbb{E}_{z_t \sim \mu_t^\star}\Big[\ell\Big(y_t, z_t\Big)\Big] \ge \left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha + \gamma\right\} + \sum_{t=1}^{T} \left(\left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha + \gamma\right) \\
&\le c\, d_\gamma + \sum_{t=1}^{T} \varepsilon_t + \alpha T + \gamma T,
\end{aligned}$$

where the final inequality uses the fact that the indicator is 1 only on $L^\star$ whose size is $\le d_\gamma$ and $\left\lceil \frac{\varepsilon_t}{\alpha} \right\rceil \alpha \le \varepsilon_t + \alpha$.

Completing the Proof. Finally, substituting this loss bound of the expert $E_{L^\star}^{\phi_{L^\star}}$ in Equation (4.2), we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \ell(y_t, \mathcal{A}(x_t))\right] &\le \sum_{t=1}^{T} \varepsilon_t + c\, d_\gamma + \alpha T + \gamma T + c\sqrt{2 d_\gamma T \ln\left(\frac{2cT}{\alpha}\right)} \\
&= \inf_{h \in \mathcal{H}} \sum_{t=1}^{T} \ell(y_t, h(x_t)) + c\, d_\gamma + \gamma T + 2c + 2c\sqrt{d_\gamma T \ln(T)},
\end{aligned}$$

where we pick $\alpha = \frac{2c}{T}$ and use the fact that $\varepsilon_t := \ell(y_t, h^\star(x_t))$. Finally, note that $c\, d_\gamma + 2c + 2c\sqrt{d_\gamma T \ln(T)} \le c\, d_\gamma + 4c\sqrt{d_\gamma T \ln(T)}$. Since $\gamma > 0$ is arbitrary, this completes our proof. ■

4.5 SMdim and the Finite Character Property

In addition to characterizing learnability, existing combinatorial dimensions in learning theory satisfy the "Finite Character Property" (FCP) [Ben-David et al., 2019, Attias et al., 2023].

Definition 21 (Finite Character Property [Ben-David et al., 2019])**.** *A combinatorial dimension* $\mathrm{D}(\mathcal{H}, \ell, \gamma)$ *is said to satisfy the finite character property if for every* $d \in \mathbb{N}$ *and* $\gamma > 0$*, the statement* $\mathrm{D}(\mathcal{H}, \ell, \gamma) \ge d$ *can be demonstrated by a finite set of domain point* $X \subset \mathcal{X}$*, and a finite subset of hypotheses* $H \subset \mathcal{H}$*.*

In fact, according to Ben-David et al. [2019], a dimension is any function D that maps $(\mathcal{H}, \ell)$ to $\mathbb{N} \cup \{0, \infty\}$ and satisfies the following two properties: (1) $\mathcal{H}$ is learnable with respect to ℓ if and only if $\mathrm{D}(\mathcal{H}, \ell) < \infty$ and (2) D satisfies the FCP. This definition of dimension differs from ours since (1) it requires D to satisfy FCP and (2) it does not require D to provide a quantitative characterization.

Despite characterizing online learnability, the SMdim may not satisfy the FCP since it is defined using trees with *infinite* width. Naturally, this motivates the following question: *Under what conditions on* $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ *does the SMdim satisfy the FCP?*

One way that the SMdim can satisfy the FCP is if it can be equivalently represented using trees with *finite* width. For example, in Section 4.3 we showed that the SMdim reduces to the Ldim, seq-fat dimension, $(k+1)$-Ldim, all of which are defined using finite-width trees, and thus satisfy the FCP. Additionally, we showed that SMdim reduces to MSdim from Raman et al. [2024a], who established that MSdim can be written using finite width trees when the underlying set system has a finite Helly number. A unifying property in all these settings is the fact that $(\mathcal{Y}, \mathcal{Z}, \ell)$ is a *Helly space*, a generalization of "finite dimension" to abstract spaces. More formally, given any $(\mathcal{Y}, \mathcal{Z}, \ell)$, let $\mathrm{B}_\ell(y, r) := \{z \in \mathcal{Z} : \ell(y, z) \leq r\}$ denote the "ball" of radius r centered at y induced by the loss ℓ. Let $\mathrm{B}_\ell(\mathcal{Y}, \mathcal{Z}) := \{\mathrm{B}_\ell(y, r) : y \in \mathcal{Y}, r \in [0, c]\}$ to be the set of all such balls. We say $(\mathcal{Y}, \mathcal{Z}, \ell)$ is a Helly space if the *Helly number* of $\mathrm{B}_\ell(\mathcal{Y}, \mathcal{Z})$ is finite.

Definition 22 (Helly Number)**.** *Let S be a family of sets. The Helly number of S, denoted $\mathrm{H}(S)$, is the smallest number $p \in \mathbb{N}$ such that for any collection of sets $C \subseteq S$ whose intersection is empty, there is a subset $C' \subset C$ of size at most p whose intersection is empty.*

The Helly number of a set system roughly quantifies the property that every sequence of sets with empty intersection has a small sub-sequence with empty intersection. In this sense, we use the Helly number of $\mathrm{B}_\ell(\mathcal{Y}, \mathcal{Z})$ to quantify a notion of "dimension" for the space $(\mathcal{Y}, \mathcal{Z}, \ell)$.

Definition 23 (Helly Space)**.** *Let $\mathcal{Z} = \mathcal{Y}$. Then, we say $(\mathcal{Y}, \mathcal{Z}, \ell)$ is a Helly space if and only if $\mathrm{H}(\mathrm{B}_\ell(\mathcal{Y}, \mathcal{Z})) < \infty$. Define the Helly number of the space $(\mathcal{Y}, \mathcal{Z}, \ell)$ as $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) := \mathrm{H}(\mathrm{B}_\ell(\mathcal{Y}, \mathcal{Z}))$.*

All existing work in supervised online learning theory has focused on Helly spaces. For example, in classification with the 0-1 loss, one can verify that $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) = 2$. For scalar-valued regression with absolute-value loss, Helly's theorem [Radon, 1921] gives that $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) = 2$. More recently, Raman et al. [2024a] showed that for online ranking with the 0-1 ranking loss, we have that $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) = 2$. Online learning settings where $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) \geq 3$ have also been

studied. For example, in list online classification $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) = k + 1$ [Moran et al., 2023]. In online learning with set-valued feedback [Raman et al., 2024a], $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) = \mathrm{H}(\mathcal{Y})$, where $\mathcal{Y}$ denotes an arbitrary set system defined over $\mathcal{Z}$.

Remarkably, in all of these aforementioned settings, the combinatorial dimensions that characterize learnability are defined using trees whose width is *exactly* $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$. More importantly, our proofs establishing the equivalence between the SMdim and existing combinatorial dimensions crucially utilized the Helly property of $(\mathcal{Y}, \mathcal{Z}, \ell)$ to compress the infinite width trees in the definition of SMdim to finite-width trees. These facts naturally lead to the question of whether the finiteness of $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$ provides a sufficient condition under which the SMdim can be represented using finite-width trees, and more specifically, $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$-width trees.

As an initial step towards answering this question, consider the p-shattering dimension defined in Definition 24. The central combinatorial object in this dimension is an $\mathcal{X}$-valued, $[p]$-ary tree $\mathcal{T}$, where $p \in \mathbb{N}$. In such a tree, each internal node of $\mathcal{T}$ has p outgoing edges, where each edge is labeled by a tuple in $\mathcal{Y} \times [0, c]$. The tuple (y, r) induces a ball $\mathrm{B}_\ell(y, r) := \{z \in \mathcal{Z} : \ell(y, z) \leq r\}$ in the space $(\mathcal{Y}, \mathcal{Z}, \ell)$ and we further require that the collection-wise intersection of the balls induced by the tuples labeling the p edges must be empty. Such a $[p]$-ary tree is shattered by a hypothesis class if for every root-to-leaf path there exists a hypothesis whose outputs on the sequence of instances lie in the balls induced by the tuples labeling the edges along the path.

Definition 24 (p-shattering dimension)**.** *Let $\ell : \mathcal{Z} \times \mathcal{Y} \to [0, c]$ be a loss function, $p \in \mathbb{N}$, and $\gamma > 0$. Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued, $[p]$-ary tree of depth d. The tree $\mathcal{T}$ is γ-shattered by $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$ if there exists a sequence $(f_1, \ldots, f_d)$ of edge-labeling functions $f_t : [p]^t \to \mathcal{Y} \times [0, c]$ such that for every path $q = (q_1, \ldots, q_d) \in [p]^d$, we have $\bigcap_{i \in [p]} \mathrm{B}\left(f_t^1((q_{<t}, i)), f_t^2((q_{<t}, i)) + \gamma\right) = \emptyset$ and there exists a hypothesis $h_q \in \mathcal{H}$ such that for all $t \in [d]$, $h_q(\mathcal{T}_t(q_{<t})) \in \mathrm{B}\left(f_t^1(q_{\leq t}), f_t^2(q_{\leq t})\right)$. The p-shattering dimension of $\mathcal{H}$ at scale γ, denoted $p\text{-}\dim_\gamma(\mathcal{H}, \ell)$, is the maximal depth of a tree $\mathcal{T}$ that is γ-shattered by $\mathcal{H}$. If there exists γ-shattered trees of arbitrarily large depth, we say $p\text{-}\dim_\gamma(\mathcal{H}, \ell) = \infty$.*

Note that the tree in Definition 24 is parameterized by both p and γ. The number p controls the width of the tree, while the number γ is used to constrain the tuples labeling the edges. When $p = \mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$, the p-dim also reduces to all existing combinatorial dimensions in their respective setting, and thus also provides a unification of supervised online learning theory. However, unlike the SMdim, the $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$-dim is defined in terms of finite-width trees whenever $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) < \infty$.

Accordingly, it is natural to ask when can the SMdim be equivalently represented using the finite-width trees in Definition 24. Lemma 9, proved in Appendix C.5, provides a partial

answer to this question by relating the SMdim and p-dim whenever $(\mathcal{Y}, \mathcal{Z}, \ell)$ is a Helly space. The key intuition behind the proof of Lemma 9 is that Helly spaces allows us to effectively "compress" the infinite-width, $\Pi(\mathcal{Z})$-ary tree from the definition of SMdim to a finite-width, $[\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)]$-ary tree according to the definition of p-dim.

Lemma 9 ($\mathrm{SMdim} \leq p\text{-dim}$)**.** *For every $(\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}, \ell)$ such that $p^\star := \mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) < \infty$, we have $\mathrm{SM}_\gamma(\mathcal{H}) \leq p^\star\text{-dim}_{\gamma'}(\mathcal{H})$ for all $\gamma' < \gamma$.*

Lemma 9 implies that when $\mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) < \infty$, the finiteness of $p^\star\text{-dim}_\gamma(\mathcal{H})$ at every scale γ is sufficient for online learnability. The following open question asks whether it is also necessary. *Suppose that $p^\star := \mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell) < \infty$. Does online learnability of $\mathcal{H}$ imply that $p^\star\text{-dim}_\gamma(\mathcal{H}) < \infty$ for all $\gamma > 0$?* One way to resolve this question would be to show that $p^\star\text{-dim}_\gamma(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H})$ for all $\gamma > 0$. A positive resolution implies that $(\mathcal{Y}, \mathcal{Z}, \ell)$ being a Helly space is a sufficient condition for SMdim to be equivalently represented using finite-width trees and therefore satisfy the FCP.

CHAPTER 5

Online Infinite-Dimensional Regression: Learning Linear Operators

In this chapter[1], we study the problem of online learning of linear operators under squared loss between two infinite-dimensional Hilbert spaces. Our focus here is on understanding the learning-theoretic landscape of this problem and identifying some key novelties. The broader problem of operator learning, with specific focus on PDE modeling, will be examined in greater detail in Part II. In this sense, this chapter serves as a bridge between Parts I and II of this thesis.

Learning operators between infinite-dimensional spaces is of fundamental importance in many scientific and engineering applications. For instance, the classical inverse problem is often modeled as learning an inverse mapping from a function space of observed data to the function space of underlying latent parameters, both of which are infinite-dimensional spaces [Kirsch, 2011, Tarantola, 2005]. Such inverse problems have found widespread applicability in domains ranging from image processing, X-ray tomography, seismic inversion, and so forth [Neto and da Silva Neto, 2012, Uhlmann, 2003]. In addition, the solution to a partial differential equation is an operator from a space of functions specifying boundary conditions to the space of solution functions [Kovachki et al., 2023, Li et al., 2021]. Moreover, many of the traditional learning settings such as multi-task learning, matrix completion, and collaborative filtering can be modeled as learning operators between infinite-dimensional spaces [Abernethy et al., 2009]. Finally, many modern supervised learning applications involve working with datasets, where both the features and labels lie in high-dimensional spaces [Deng et al., 2009, Santhanam et al., 2017]. Thus, it is desirable to construct learning algorithms whose guarantees do not scale with the ambient dimensions of the problem.

Most of the existing work in operator learning assumes some stochastic model for the data, which can be unrealistic in many applications. For instance, the majority of appli-

[1]This chapter is based on: Vinod Raman*, Unique Subedi*, and Ambuj Tewari (2024). *Online Infinite-Dimensional Regression: Learning Linear Operators*. Conference on Algorithmic Learning Theory (ALT).

cations of operator learning are in the scientific domain where the data often comes from experiments [Lin et al., 2021]. Since experiments are costly, the data usually arrives sequentially and with a strong temporal dependence that may not be adequately captured by a stochastic model. Additionally, given the high-dimensional nature of the data, one typically uses pre-processing techniques like PCA to project the data onto a low-dimensional space [Bhattacharya et al., 2021, Lanthaler, 2023]. Even if the original data has some stochastic nature, the preprocessing step introduces non-trivial dependencies in the observations that may be difficult to model. Accordingly, it is desirable to construct learning algorithms that can handle *arbitrary* dependencies in the data. In fact, for continuous problems such as scalar-valued regression, one can often obtain guarantees similar to that of i.i.d. setting without making any assumptions on the data [Rakhlin and Sridharan, 2014].

In this paper, we study linear operator learning between two Hilbert spaces $\mathcal{V}$ and $\mathcal{W}$ in the *adversarial online setting*, where one makes no assumptions on the data generating process [Cesa-Bianchi and Lugosi, 2006]. In this model, a potentially adversarial nature plays a sequential game with the learner over T rounds. In each round $t \in [T]$, nature selects a pair of vectors $(x_t, y_t) \in \mathcal{V} \times \mathcal{W}$ and reveals x_t to the learner. The learner then makes a prediction $\hat{y}_t \in \mathcal{W}$. Finally, the adversary reveals the target y_t, and the learner suffers the loss $\|\hat{y}_t - y_t\|_{\mathcal{W}}^2$. A linear operator class $\mathcal{F} \subset \mathcal{W}^{\mathcal{V}}$ is online learnable if there exists an online learning algorithm such that for any sequence of labeled examples, the difference in cumulative loss between its predictions and the predictions of the best-fixed operator in $\mathcal{F}$ is small. In this work, we study the online learnability of linear operators and make the following contributions:

(1) We show that the class of linear operators with uniformly bounded p-Schatten norm is online learnable with regret $O(T^{\max\left\{\frac{1}{2}, 1-\frac{1}{p}\right\}})$. We also provide a lower bound of $\Omega(T^{1-\frac{1}{p}})$, which matches the upperbound for $p \geq 2$.

(2) We prove that the class of linear operators with uniformly bounded operator norm is not online learnable. Furthermore, we show that this impossibility result also holds in the batch setting.

(3) Recently, there is a growing interest in understanding when uniform convergence and learnability are not equivalent [Montasser et al., 2019, Hanneke et al., 2023]. Along this direction, we give a subset of bounded linear operators for which online learnability and uniform convergence are not equivalent.

To make contribution (1), we upperbound the sequential Rademacher complexity of the loss class to show that sequential uniform convergence holds for the p-Schatten class for $p \in [1, \infty)$. For our hardness result stated in contribution (2), we construct a class with uniformly

bounded operator norm that is not online learnable. Our construction in contribution (3) is inspired by and generalizes the example of Natarajan [1989b, Page 22], which shows a gap between uniform convergence and PAC learnability for multiclass classification. The argument showing that uniform convergence does not hold is a simple adaptation of the existing proof [Natarajan, 1989b]. However, since our loss is real-valued, showing that the class is learnable requires some novel algorithmic ideas, which can be of independent interest.

5.1 Related Works

Regression between two infinite-dimensional function spaces is a classical statistical problem often studied in functional data analysis (FDA) [Wang et al., 2016, Ferraty, 2006]. In FDA, one typically considers $\mathcal{V}$ and $\mathcal{W}$ to be $L^2[0,1]$, the space of square-integrable functions, and the hypothesis class is usually a class of kernel integral operators. We discuss the implication of our results to learning kernel integral operators in Section 5.3.1. Recently, de Hoop et al. [2023], Nelsen and Stuart [2021], Mollenhauer et al. [2022] study learning more general classes of linear operators. However, all of these works are in the i.i.d. setting and assume a data-generating process. Additionally, there is a line of work that uses deep neural networks to learn neural operators between function spaces [Kovachki et al., 2023, Li et al., 2021]. Unfortunately, there are no known learning guarantees for these neural operators. Closer to the spirit of our work is that of Tabaghi et al. [2019], who consider the agnostic PAC learnability of p-Schatten operators. They show that p-Schatten classes are agnostic PAC learnable. In this work, we complement their results by showing that p-Schatten classes are also online learnable. Going beyond the i.i.d. setting, there is a line of work that focuses on learning specific classes of operators from time series data [Brunton et al., 2016, Klus et al., 2020].

5.2 Preliminaries

5.2.1 Hilbert Space Basics

Let $\mathcal{V}$ and $\mathcal{W}$ be real, separable, and infinite-dimensional Hilbert spaces. Recall that a Hilbert space is separable if it admits a countable orthonormal basis. Throughout the paper, we let $\{e_n\}_{n=1}^{\infty}$ and $\{\psi_n\}_{n=1}^{\infty}$ denote a set of orthonormal basis for $\mathcal{V}$ and $\mathcal{W}$ respectively. Then, any element $v \in \mathcal{V}$ and $w \in \mathcal{W}$ can be written as $v = \sum_{n=1}^{\infty} \beta_n e_n$ and $w = \sum_{n=1}^{\infty} \alpha_n \psi_n$ for sequences $\{\beta_n\}_{n\in\mathbb{N}}$ and $\{\alpha_n\}_{n=1}^{\infty}$ that are ℓ_2 summable.

Consider $w_1, w_2 \in \mathcal{W}$ such that $w_1 = \sum_{n=1}^{\infty} \alpha_{n,1}\, \psi_n$ and $\sum_{n=1}^{\infty} \alpha_{n,2}\, \psi_n$. Then, the inner

product between w_1 and w_2 is defined as $\langle w_1, w_2 \rangle_{\mathcal{W}} := \sum_{n=1}^{\infty} \alpha_{n,1}\alpha_{n,2}$, and it induces the norm $\|w_1\|_{\mathcal{W}} := \sqrt{\langle w_1, w_1 \rangle_{\mathcal{W}}} = \sqrt{\sum_{n=1}^{\infty} \alpha_{n,1}^2}$. One can equivalently define $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and $\|\cdot\|_{\mathcal{V}}$ to be the inner-product and the induced norm in the Hilbert space $\mathcal{V}$. When the context is clear, we drop the subscript and simply write $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$.

A linear operator $f : \mathcal{V} \to \mathcal{W}$ is a mapping that preserves the linear structure of the input. That is, $f(c_1 v_1 + c_2 v_2) = c_1 f(v_1) + c_2 f(v_2)$ for any $c_1, c_2 \in \mathbb{R}$ and $v_1, v_2 \in \mathcal{V}$. Let $\mathcal{L}(\mathcal{V}, \mathcal{W})$ denote the set of all linear operators from $\mathcal{V}$ to $\mathcal{W}$. A linear operator $f : \mathcal{V} \to \mathcal{W}$ is bounded if there exists a constant $c > 0$ such that $\|f(v)\| \leq c \|v\|$ for all $v \in \mathcal{V}$. The quantity $\|f\|_{\text{op}} := \inf\{c \geq 0 : \|f(v)\| \leq c \|v\|, \forall v \in \mathcal{V}\}$ is called the operator norm of f. The operator norm induces the set of bounded linear operators, $\mathcal{B}(\mathcal{V}, \mathcal{W}) = \{f \in \mathcal{L}(\mathcal{V}, \mathcal{W}) \mid \|f\|_{\text{op}} < \infty\}$, which is a Banach space with $\|\cdot\|_{\text{op}}$ as the norm.

For an operator $f \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, let $f^\star : \mathcal{W} \to \mathcal{V}$ denote the adjoint of f. We can use f and $f^\star$ to define a self-adjoint, non-negative operator $f^\star f : \mathcal{V} \to \mathcal{V}$. Moreover, the absolute value operator is defined as $|f| := (f^\star f)^{\frac{1}{2}}$, which is the unique non-negative operator such that $|f| \circ |f| = f^\star f$. Given any operator $g : \mathcal{V} \to \mathcal{V}$, the trace of g is defined as $\text{tr}(g) = \sum_{n=1}^{\infty} \langle g(e_n), e_n \rangle$, where $\{e_n\}_{n=1}^{\infty}$ is any orthonormal basis of $\mathcal{V}$. The notion of trace and absolute value allows us to define the p-Schatten norm of f,

$$\|f\|_p = \Big(\text{tr}(|f|^p) \Big)^{\frac{1}{p}},$$

for all $p \in [1, \infty)$. Accordingly, we can define the p-Schatten class as

$$S_p(\mathcal{V}, \mathcal{W}) = \{f \in \mathcal{L}(\mathcal{V}, \mathcal{W}) \mid f \text{ is compact and } \|f\|_p < \infty\}.$$

A linear operator $f : \mathcal{V} \to \mathcal{W}$ is compact if the closure of the set $\{f(v) \mid v \in \mathcal{V}, \|v\| \leq 1\}$ is compact. For a compact linear operator $f : \mathcal{V} \to \mathcal{W}$, there exists a sequence of orthonormal basis $\{\phi_n\}_{n=1}^{\infty} \subset \mathcal{V}$ and $\{\varphi_n\}_{n=1}^{\infty} \subset \mathcal{W}$ such that $f = \sum_{n=1}^{\infty} s_n(f)\, \varphi_n \otimes \phi_n$, where $s_n(f) \downarrow 0$ and $\varphi_n \otimes \phi_n$ denote the tensor product between φ_n and ϕ_n. This is the singular value decomposition of f and the sequence $\{s_n(f)\}_{n=1}^{\infty}$ are the singular values of f. For $p \in [1, \infty)$, the p-Schatten norm of a compact operator is equal to the ℓ_p norm of the sequence $\{s_n(f)\}_{n \geq 1}$,

$$\|f\|_p = \left(\sum_{n=1}^{\infty} s_n(f)^p \right)^{\frac{1}{p}}.$$

On the other hand, for a compact operator f, the ℓ_∞ norm of its singular values is equal to its operator norm, $\|f\|_{\text{op}} = \|f\|_\infty = \sup_{n \geq 1} |s_n(f)|$. Accordingly, for compact operators, the

operator norm is referred to as ∞-Schatten norm, which induces the class

$$S_\infty(\mathcal{V}, \mathcal{W}) = \{f \in \mathcal{L}(\mathcal{V}, \mathcal{W}) \mid f \text{ is compact and } \|f\|_\infty < \infty\}.$$

Therefore, $S_\infty(\mathcal{V}, \mathcal{W}) \subset \mathcal{B}(\mathcal{V}, \mathcal{W})$. For a comprehensive treatment of the theory of Hilbert spaces and linear operators, we refer the reader to Conway [1990] and Weidmann [2012].

5.2.2 Online Learning

Let $\mathcal{X} \subseteq \mathcal{V}$ denote the instance space, $\mathcal{Y} \subseteq \mathcal{W}$ denote the target space, and $\mathcal{F} \subseteq \mathcal{L}(\mathcal{V}, \mathcal{W})$ denote the hypothesis class. In online linear operator learning, a potentially adversarial nature plays a sequential game with the learner over T rounds. In each round $t \in [T]$, the nature selects a labeled instance $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ and reveals x_t to the learner. The learner then uses all past examples $\{(x_i, y_i)\}_{i=1}^{t-1}$ and the newly revealed instance x_t to make a prediction $\hat{y}_t \in \mathcal{Y}$. Finally, the adversary reveals the target y_t, and the learner suffers the loss $\|\hat{y}_t - y_t\|_{\mathcal{W}}^2$. Given $\mathcal{F}$, the goal of the learner is to make predictions such that its regret, defined as a difference between the cumulative loss of the learner and the best possible cumulative loss over operators in $\mathcal{F}$, is small.

Definition 25 (Online Linear Operator Learnability)**.** *A linear operator class $\mathcal{F} \subseteq \mathcal{L}(\mathcal{V}, \mathcal{W})$ is online learnable if there exists an algorithm $\mathcal{A}$ such that its expected regret is*

$$\mathrm{R}_{\mathcal{A}}(T, \mathcal{F}) := \sup_{(x_1,y_1),\ldots,(x_T,y_T)} \mathbb{E}\left[\sum_{t=1}^{T} \|\mathcal{A}(x_t) - y_t\|^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \|f(x_t) - y_t\|^2\right]$$

is a non-decreasing, sublinear function of T.

Unlike when $\mathcal{V}$ is finite-dimensional, the class $\mathcal{F} = \mathcal{L}(\mathcal{V}, \mathcal{W})$ is not online learnable when $\mathcal{V}$ is infinite-dimensional (see Section 5.4). Accordingly, we are interested in understanding for which subsets $\mathcal{F} \subset \mathcal{L}(\mathcal{V}, \mathcal{W})$ is online learning possible. Beyond online learnability, we are also interested in understanding when a probabilistic property called the sequential uniform convergence holds for the loss class $\{(x, y) \mapsto \|f(x) - y\|^2 : f \in \mathcal{F}\}$.

Definition 26 (Sequential Uniform Convergence)**.** *Let $\{(X_t, Y_t)\}_{t=1}^{T}$ be an arbitrary sequence of random variables defined over an appropriate probability space on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{C} = \{\mathcal{C}_t\}_{t=0}^{T-1}$ be an arbitrary filtration such that (X_t, Y_t) is $\mathcal{C}_t$-measurable. Given a linear operator class $\mathcal{F} \subseteq \mathcal{L}(\mathcal{V}, \mathcal{W})$, we say that sequential uniform convergence holds for a loss class $\{(x, y) \mapsto$*

$\|f(x) - y\|^2 : f \in \mathcal{F}\}$ *if*

$$\limsup_{T\to\infty} \sup_{\mathbf{P}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{T}\sum_{t=1}^{T} \left(\|f(X_t) - Y_t\|^2 - \mathbb{E}[\|f(X_t) - Y_t\|^2 \mid \mathcal{C}_{t-1}] \right) \right| \right] = 0.$$

Here, the supremum is taken over all joint distributions $\mathbf{P}$ *of* $\{(X_t, Y_t)\}_{t=1}^T$.

A general complexity measure called the sequential Rademacher complexity characterizes sequential uniform convergence [Rakhlin et al., 2015a,b].

Definition 27 (Sequential Rademacher Complexity)**.** *Let* $\sigma = \{\sigma_i\}_{i=1}^T$ *be a sequence of independent Rademacher random variables and* $(x, y) = \{(x_t, y_t)\}_{t=1}^T$ *be a sequence of functions* $(x_t, y_t) : \{-1, 1\}^{t-1} \to \mathcal{X} \times \mathcal{Y}$*. Then, the sequential Rademacher complexity of the loss class* $\{(v, w) \mapsto \|f(v) - w\|^2 : f \in \mathcal{F}\}$ *is defined as*

$$\mathrm{Rad}_T(\mathcal{F}) = \sup_{x,y} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \sigma_t \|f(x_t(\sigma_{<t})) - y_t(\sigma_{<t})\|^2\right],$$

where $\sigma_{<t} = (\sigma_1, \ldots, \sigma_{t-1})$.

If there exists a $B > 0$ such that $\sup_{f,v,w} \|f(v) - w\|^2 \le B$, then Theorem 1 of Rakhlin et al. [2015b] implies that the sequential uniform convergence holds for the loss class $\{(v, w) \mapsto \|f(v) - w\|^2 : f \in \mathcal{F}\}$ if and only if $\mathrm{Rad}_T(\mathcal{F}) = o(T)$. Given this equivalence, in this work, we only rely on the sequential Rademacher complexity of $\mathcal{F}$ to study its sequential uniform convergence property.

5.3 Schatten Operators are Online Learnable

In this section, we show that every uniformly bounded subset of $S_p(\mathcal{V}, \mathcal{W})$ is online learnable. Despite not making any distributional assumptions, the rates in Theorem 21 match the lowerbounds in the batch settig established in Section 5.4.1. This complements the results by Rakhlin and Sridharan [2014], who show that the rates for scalar-valued regression with squared loss are similar for online and PAC learning.

Theorem 21 (Uniformly Bounded Subsets of $S_p(\mathcal{V}, \mathcal{W})$ are Online Learnable)**.** *Fix* $c > 0$*. Let* $\mathcal{X} = \{v \in \mathcal{V} \mid \|v\| \le 1\}$ *denote the instance space,* $\mathcal{Y} = \{w \in \mathcal{W} \mid \|w\| \le c\}$ *denote the target space, and* $\mathcal{F}_p = \{f \in S_p(\mathcal{V}, \mathcal{W}) \mid \|f\|_p \le c\}$ *be the hypothesis class for* $p \in [1, \infty]$*. Then,*

$$\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}_p) \le 2\,\mathrm{Rad}_T(\mathcal{F}_p) \le 6c^2\, T^{\max\left\{\frac{1}{2}, 1-\frac{1}{p}\right\}}.$$

Theorem 21 implies the regret $O(\sqrt{T})$ for $p \in [1,2]$ and the regret $O(T^{1-\frac{1}{p}})$ for $p > 2$. When $p = \infty$, the regret bound implied by Theorem 21 is vacuous. Indeed, in Section 5.4, we prove that any uniformly bounded subset of $S_\infty(\mathcal{V}, \mathcal{W})$ is not online learnable.

Our proof of Theorem 21 relies on Lemma 10 which shows that the q-Schatten norm of Rademacher sums of rank-1 operators concentrates for every $q \geq 1$. The proof of Lemma 10 is in Appendix D.1.

Lemma 10 (Rademacher Sums of Rank-1 Operators)**.** *Let $\sigma = \{\sigma_i\}_{i=1}^T$ be a sequence of independent Rademacher random variables and $\{(v_t, w_t)\}_{t=1}^T$ be any sequence of functions $(v_t, w_t) : \{-1,1\}^{t-1} \to \{v \in \mathcal{V} : \|v\| \leq c_1\} \times \{w \in \mathcal{W} : \|w\| \leq c_2\}$. Then, for any $q \geq 1$, we have*

$$\mathbb{E}\left[\left\|\sum_{t=1}^T \sigma_t\, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right\|_q\right] \leq c_1\, c_2\, T^{\max\left\{\frac{1}{2}, \frac{1}{q}\right\}}$$

Lemma 10 extends Lemma 1 in [Tabaghi et al., 2019] to the non-i.i.d. setting. In particular, the rank-1 operator indexed by t can depend on the Rademacher subsequence $\sigma_{<t}$, whereas they only consider the case when the rank-1 operators are independent of the Rademacher sequence. In addition, Tabaghi et al. [2019] use a non-trivial result from convex analysis, namely the fact that $A \mapsto \mathrm{tr}(h(F))$ is a convex functional on the set $\{F \in \mathcal{T} \mid \mathrm{spectra}(F) \subseteq [\alpha, \beta]\}$ for any convex function h and the class of finite-rank self-adjoint operators $\mathcal{T}$. Our proof of Lemma 10, on the other hand, only uses standard inequalities.

Equipped with Lemma 10, our proof of Theorem 21 follows by upper bounding the sequential Rademacher complexity of the loss class. Although this proof of online learnability is non-constructive, we can use Proposition 1 from [Rakhlin et al., 2012b] to design an explicit online learner that achieves the matching regret given access to an oracle that computes the sequential Rademacher complexity of the class. Moreover, online mirror descent (OMD) with the $\|f\|_p^p$ regularizer also achieves the rates established in Theorem 21. In particular, OMD with the strongly convex regularizer $\|f\|_2^2$ guarantees regret $O(\sqrt{T})$ for $p = 2$. The $O(\sqrt{T})$ regret bound for $\mathcal{F}_2$ immediately implies an $O(\sqrt{T})$ regret bound for all $\mathcal{F}_p \subseteq \mathcal{F}_2$ in $p \in [1,2]$ by monotonicity. For $p > 2$, the Clarkson-McCarthy inequality [Bhatia and Holbrook, 1988] implies that $\|f\|_p^p$ is p-uniformly convex and thus OMD with this regularizer obtains the regret of $O(T^{1-\frac{1}{p}})$ [Sridharan and Tewari, 2010, Srebro et al., 2011]. That said, Theorem 21 establishes a stronger guarantee– not only are these classes online learnable but they also enjoy sequential uniform convergence.

5.3.1 Examples of p-Schatten class

In this section, we provide examples of operator classes with uniformly bounded p-Schatten norm.

Uniformly bounded operators w.r.t. $\|\cdot\|_{\mathbf{op}}$ when either $\mathcal{V}$ or $\mathcal{W}$ is finite-dimensional. If either the input space $\mathcal{V}$ or the output space $\mathcal{W}$ is finite-dimensional, then the class of bounded linear operators $\mathcal{B}(\mathcal{V}, \mathcal{W})$ is p-Schatten class for every $p \in [1, \infty]$. This is immediate because for every $f \in \mathcal{B}(\mathcal{V}, \mathcal{W})$, either the operator $f^\star f : \mathcal{V} \to \mathcal{V}$ or $f f^\star : \mathcal{W} \to \mathcal{W}$ is a bounded operator that maps between two finite-dimensional spaces. Let $\|f\|_{\text{op}} \leq c$ and $\min\{\dim(\mathcal{V}), \dim(\mathcal{W})\} = d < \infty$. Since $f^\star f$ and $f f^\star$ have the same singular values and one of them has rank at most d, both of them must have rank at most d. Let $s_1 \geq s_2 \ldots \geq s_d \geq 0$ denote all singular values of $f^\star f$. Then, $\|f\|_p = \left(\sum_{i=1}^d s_i^p\right)^{\frac{1}{p}} \leq c\, d^{\frac{1}{p}} < \infty$, where we use the fact that $s_i \leq c$ for all i. Since $\|f\|_2 \leq c\sqrt{d}$, Theorem 21 implies that $\mathcal{F} = \{f \in \mathcal{B}(\mathcal{V}, \mathcal{W}) \mid \|f\|_{\text{op}} \leq c\}$ is online learnable with regret at most $6c^2 d\sqrt{T}$.

Kernel Integral Operators. Let $\mathcal{V}$ denote a Hilbert space of functions defined on some domain Ω. Then, a kernel $K : \Omega \times \Omega \to \mathbb{R}$ defines an integral operator $f_K : \mathcal{V} \to \mathcal{W}$ such that $f_K(v(r)) = \int_\Omega K(r, s)\, v(s)\, d\mu(s)$, for some measure space (Ω, μ). Now define a class of integral operators,

$$\mathcal{F} = \left\{ f_K \,:\, \int_\Omega \int_\Omega |K(r, s)|^2 \, d\mu(r)\, d\mu(s) \leq c^2 \right\},$$

induced by all the kernels whose L^2 norm is bounded by c. It is well known that $\|f\|_2 \leq c$ for every $f \in \mathcal{F}$ (see [Conway, 1990, Page 267] and [Weidmann, 2012, Theorem 6.11]) . Thus, Theorem 21 implies that $\mathcal{F}$ is online learnable with regret $6c^2\sqrt{T}$.

5.4 Lower Bounds and Hardness Results

In this section, we establish lower bounds for learning uniformly bounded subsets of $S_p(\mathcal{V}, \mathcal{W})$ for $p \in [1, \infty]$.

Theorem 22 (Lower Bounds for Uniformly Bounded Subsets of $S_p(\mathcal{V}, \mathcal{W})$)**.** *Fix $c > 0$. Let $\mathcal{X} = \{v \in \mathcal{V} \mid \|v\| \leq 1\}$ denote the instance space, $\mathcal{Y} = \{w \in \mathcal{W} \mid \|w\| \leq c\}$ denote the target space, and $\mathcal{F}_p = \{f \in S_p(\mathcal{V}, \mathcal{W}) \mid \|f\|_p \leq c\}$ be the hypothesis class for $p \in [1, \infty]$. Then, we have*

$$\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}_p) \geq c^2\, T^{1 - \frac{1}{p}}.$$

Theorem 22 shows a linear lowerbound of $c^2\,T$ for $p = \infty$, thus implying that the class $\mathcal{F}_\infty$ *is not online learnable*. For $p \in [2, \infty)$, the lowerbound in Theorem 22 matches the upperbound in Theorem 21 up to a factor of 6. However, in the range $p \in [1, 2)$, our upperbound saturates at the rate $\sqrt{T}$, while the lower bound gets progressively worse as p decreases. It remains an open problem to find the optimal regret of learning $\mathcal{F}_p$ for $p \in [1, 2)$.

Proof. (of Theorem 22) Fix an algorithm $\mathcal{A}$, and consider a labeled stream $\{(e_t, c\,\sigma_t\psi_t)\}_{t=1}^T$ where $\sigma_t \sim \text{Unif}(\{-1, 1\})$. Then, the expected loss of $\mathcal{A}$ is

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \|\mathcal{A}(e_t) - c\,\sigma_t\psi_t\|^2\right] &\geq \sum_{t=1}^{T} \left(\mathbb{E}\left[\|\mathcal{A}(e_t) - c\,\sigma_t\psi_t\|\right]\right)^2 \\
&= \sum_{t=1}^{T} \left(\mathop{\mathbb{E}}_{\mathcal{A}}\left[\frac{1}{2}\|\mathcal{A}(x_t) - c\,\psi_t\| + \frac{1}{2}\|\mathcal{A}(x_t) + c\,\psi_t\|\right]\right)^2 \\
&\geq \sum_{t=1}^{T} \left(\frac{1}{2}\|c\,\psi_t - (-c\,\psi_t)\|\right)^2 = \sum_{t=1}^{T} c^2\ \|\psi_t\|^2 = c^2\,T.
\end{aligned}$$

The first inequality above is due to Jensen's, whereas the second inequality is the triangle inequality.

To establish the upper bound on the optimal cumulative loss amongst operators in $\mathcal{F}_p$, consider the operator $f_{\sigma,p} := \sum_{t=1}^{T} \frac{c\,\sigma_t}{T^{1/p}}\ \psi_t \otimes e_t$. As the singular values of $f_{\sigma,p}$ are $\{c\,\sigma_t T^{-1/p}\}_{t=1}^T$, we have

$$\|f_{\sigma,p}\|_p = \left(\sum_{t=1}^{T} \left|\frac{c\,\sigma_t}{T^{1/p}}\right|^p\right)^{1/p} = \left(\sum_{t=1}^{T} \frac{c^p}{T}\right)^{1/p} = c \quad \text{for } p \in [1, \infty).$$

Similarly, $\|f_{\sigma,\infty}\|_\infty = \left\|\sum_{t=1}^{T} c\sigma_t\psi_t \otimes e_t\right\|_\infty = \max_{t\geq 1} |c\,\sigma_t| = c$. That is, $f_{\sigma,p} \in \mathcal{F}_p$ for all $p \geq 1$. Thus, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\inf_{f\in\mathcal{F}_p} \sum_{t=1}^{T} \|f(e_t) - c\sigma_t\psi_t\|^2\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \|f_{\sigma,p}(e_t) - c\sigma_t\psi_t\|^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\|\frac{c\,\sigma_t}{T^{1/p}}\psi_t - c\sigma_t\psi_t\right\|^2\right] \\
&= \sum_{t=1}^{T} c^2 \left(1 - \frac{1}{T^{1/p}}\right)^2 \\
&\leq \sum_{t=1}^{T} c^2 \left(1 - \frac{1}{T^{1/p}}\right) = c^2\,T - c^2\,T^{1-\frac{1}{p}}.
\end{aligned}$$

Therefore, we have shown that the regret of $\mathcal{A}$ is

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\mathcal{A}(e_t)-c\,\sigma_t\psi_t\|^2-\inf_{f\in\mathcal{F}_p}\sum_{t=1}^{T}\|f(e_t)-c\,\sigma_t\psi_t\|^2\right]\geq c^2\,T^{1-\frac{1}{p}}.$$

Our proof uses a random adversary, and the expectation above is taken with respect to both the randomness of the algorithm and the stream. However, one can use the probabilistic method to argue that for every algorithm, there exists a fixed stream forcing the claimed lowerbound. This completes our proof. ■

5.4.1 Lower Bounds in the Batch Setting

In the batch setting, the learner is provided with $n\in\mathbb{N}$ i.i.d. samples $S=\{(x_i,y_i)\}_{i=1}^n$ from a joint distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$ that is unknown to the learner. Using the sample S, the learner then finds a predictor $\hat{f}_n\in\mathcal{Y}^{\mathcal{X}}$ using some learning rule. We will abuse notation and use $\hat{f}_n$ to denote both the learning rule and the predictor returned by it. Given a linear operator class $\mathcal{F}\subseteq\mathcal{L}(\mathcal{V},\mathcal{W})$, the goal of the learner is to find an estimator $\hat{f}_n$ with a small worst-case expected excess risk

$$\mathcal{E}_n(\mathcal{F},\hat{f}_n):=\sup_{\mathcal{D}}\mathop{\mathbb{E}}_{S_n\sim\mathcal{D}^n}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left\|\hat{f}_n(x)-y\right\|^2\right]-\inf_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\|f(x)-y\|^2\right]\right].$$

The minimax excess risk for learning the function class $\mathcal{F}$ is then defined as $\mathcal{E}_n(\mathcal{F})=\inf_{\hat{f}_n}\mathcal{E}(\mathcal{F},\hat{f})$, where the infimum is over all possible learning rules. We adopt the minimax perspective to define agnostic batch learnability.

Definition 28 (Batch Learnability)**.** *A linear operator class $\mathcal{F}\subseteq\mathcal{L}(\mathcal{V},\mathcal{W})$ is batch learnable if and only if* $\limsup_{n\to\infty}\mathcal{E}_n(\mathcal{F})=0$.

Our results in Section 5.3 immediately provide an upperbound on $\mathcal{E}_n(\mathcal{F})$ because $\mathcal{E}_n(\mathcal{F})$ is upper bounded by the batch Rademacher complexity of $\mathcal{F}$, which is further upper bounded by its sequential analog. Similar upperbounds on batch Rademacher complexity of $\mathcal{F}$ were also provided by Tabaghi et al. [2019]. In this section, we complement these results by providing lower bounds on $\mathcal{E}_n(\mathcal{F})$.

Theorem 23 (Batch Lower Bounds for Uniformly Bounded Subsets of $S_p(\mathcal{V},\mathcal{W})$)**.** *Fix $c>0$. Let $\mathcal{X}=\{v\in\mathcal{V}\mid\|v\|\leq 1\}$ denote the instance space, $\mathcal{Y}=\{w\in\mathcal{W}\mid\|w\|\leq c\}$ denote the target space, and $\mathcal{F}_p=\{f\in S_p(\mathcal{V},\mathcal{W})\mid\|f\|_p\leq c\}$ be the hypothesis class for $p\in[1,\infty]$. Then, we have*

$$\mathcal{E}_n(\mathcal{F})\geq\frac{c^2}{12}\max\left\{n^{-\frac{1}{p-1}},n^{-\frac{2}{p}}\right\}.$$

Theorem 23 shows a non-vanishing lowerbound of $\frac{c^2}{12}$ for $p = \infty$, immediately implying that the class $\mathcal{F}_\infty$ *is not batch learnable*. For $p \in [2, \infty)$, Tabaghi et al. [2019] provides an upperbound of $O(n^{-\frac{1}{p}})$, whereas our lowerbound is $\Omega(n^{-\frac{1}{p-1}})$. Additionally, for $p \in [1, 2)$, there is also a gap between our lowerbound of $\Omega(n^{-\frac{2}{p}})$ and Tabaghi et al. [2019]'s upperbound of $O(n^{-\frac{1}{2}})$. Thus, it remains to find the optimal rates for learning $\mathcal{F}_p$ for every $p \in [1, \infty)$.

5.5 Online Learnability without Sequential Uniform Convergence

In learning theory, the uniform law of large numbers is intimately related to the learnability of a hypothesis class. For instance, a binary hypothesis class is PAC learnable if and only if the hypothesis class satisfies the i.i.d. uniform law of large numbers [Shalev-Shwartz and Ben-David, 2014]. An online equivalent of this result states that a binary hypothesis class is *online* learnable if and only if the hypothesis class satisfies the sequential uniform law of large numbers [Rakhlin et al., 2015b]. However, in a recent work, Hanneke et al. [2023] show that uniform convergence and learnability are not equivalent for online multiclass classification. A key factor in Hanneke et al. [2023]'s proof is the unboundedness of the size of the label space. This unboundedness is critical as the equivalence between uniform convergence and learnability continues to hold for multiclass classification with a finite number of labels [Daniely et al., 2011]. Nevertheless, the number of labels alone cannot imply a separation. This is true because a real-valued function class (say $\mathcal{G} \subseteq [-1, 1]^{\mathcal{X}}$ where the size of label space is uncountably infinite) is online learnable with respect to absolute/squared-loss if and only if the uniform convergence holds [Rakhlin et al., 2015a]. In this section, we show an analogous separation between uniform convergence and learnability for online linear operator learning. As the unbounded label space was to Hanneke et al. [2023], the infinite-dimensional nature of the target space is critical to our construction exhibiting this separation. Mathematically, a unifying property of Hanneke et al. [2023]'s and our construction is the fact that the target space $\mathcal{Y}$ is not *totally bounded* with respect to the pseudometric defined by the loss function.

The following result establishes a separation between uniform convergence and online learnability for bounded linear operators. In particular, we show that there exists a class of bounded linear operators $\mathcal{F}$ such that the sequential uniform law of large numbers does not hold, but $\mathcal{F}$ is online learnable.

Theorem 24 (Sequential Uniform Convergence $\not\equiv$ Online Learnability)**.** *Let* $\mathcal{X} = \{v \in \mathcal{V} \mid \sum_{n=1}^{\infty} |c_n| \leq 1$ *where* $v = \sum_{n=1}^{\infty} c_n e_n\}$ *be the instance space and* $\mathcal{Y} = \{v \in \mathcal{V} \mid \|v\| \leq 1\}$ *be the target space. Then, there exists a function class* $\mathcal{F} \subset S_1(\mathcal{V}, \mathcal{V})$ *such that the following holds:*

(i) $\mathrm{Rad}_T(\mathcal{F}) \geq \frac{T}{2}$

(ii) $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}) \leq 2 + 8\sqrt{T \log(2T)}$.

Proof. For a natural number $k \in \mathbb{N}$, define an operator $f_k : \mathcal{V} \to \mathcal{V}$ as

$$f_k := \sum_{n=1}^{\infty} b_k[n] \; e_k \otimes e_n = e_k \otimes \sum_{n=1}^{\infty} b_k[n]\, e_n \tag{5.1}$$

where b_k is the binary representation of the natural number k and $b_k[n]$ is its n^{th} bit. Define $\mathcal{F} = \{f_k \mid k \in \mathbb{N}\} \cup \{f_0\}$ where $f_0 = 0$.

We begin by showing that $\mathcal{F} \subset S_1(\mathcal{V}, \mathcal{V})$. For any $\alpha, \beta \in \mathbb{R}$ and $v_1, v_2 \in \mathcal{V}$, we have

$$f_k(\alpha v_1 + \beta v_2) = \sum_{n=1}^{\infty} b_k[n] \;\; \langle e_n, \alpha v_1 + \beta v_2 \rangle \, e_k = \alpha f_k(v_1) + \beta f_k(v_2).$$

Thus, f_k is a linear operator. Note that f_k is defined in terms of singular value decomposition, and has only one non-zero singular value along the direction of e_k. Therefore,

$$\|f_k\|_1 = \sum_{n=1}^{\infty} b_k[n] \leq \log_2(k) + 1,$$

where we use the fact that there can be at most $\log_2(k) + 1$ non-zero bits in the binary representation of k. This further implies that $\|f_k\|_p \leq \|f_k\|_1 \leq \log_2(k) + 1 < \infty$ for all $p \in [1, \infty]$. Note that each f_k maps a unit ball in $\mathcal{V}$ to a subset of $\{\alpha\, e_k : |\alpha| \leq \log_2(k) + 1\}$, which is a compact set for every $k \in \mathbb{N}$. Thus, for every $k \in \mathbb{N}$, f_k is a compact operator and $f_k \in S_1(\mathcal{V}, \mathcal{V})$. We trivially have $f_0 \in S_1(\mathcal{V}, \mathcal{V})$.

Proof of (i). Let $\sigma = \{\sigma_t\}_{t=1}^T$ be a sequence of i.i.d. Rademacher random variables. Consider a sequence of functions $(x, y) = \{x_t, y_t\}_{t=1}^T$ such that $x_t(\sigma_{<t}) = e_t$ and $y_t(\sigma_{<t}) = 0$ for all $t \in [T]$. Note that our sequence $\{e_t\}_{t=1}^T \subseteq \mathcal{X}$. Then, the sequential Rademacher complexity of the loss class is

$$\begin{aligned}
\mathrm{Rad}_T(\mathcal{F}) = \sup_{x,y} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t})) - y_t(\sigma_{<t})\right\|^2\right] &\geq \mathbb{E}\left[\sup_{k \in \mathbb{N}} \sum_{t=1}^{T} \sigma_t \left\|f_k(e_t)\right\|^2\right] \\
&= \mathbb{E}\left[\sup_{k \in \mathbb{N}} \sum_{t=1}^{T} \sigma_t \, b_k[t]\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\sigma_t = 1\}\right] = \frac{T}{2}.
\end{aligned}$$

Here, we use the fact that $f_k(e_t) = b_k[t]\, e_k$ and $\mathbb{P}[\sigma_t = 1] = \frac{1}{2}$. As for the inequality $\sup_{k \in \mathbb{N}} \sum_{t=1}^{T} \sigma_t \, b_k[t] \geq \sum_{t=1}^{T} \mathbb{1}\{\sigma_t = 1\}$, note that for any sequence $\{\sigma_t\}_{t=1}^T$, there exists a

$k \in \mathbb{N}$ (possibly of the order $\sim 2^T$) such that $b_k[t] = 1$ whenever $\sigma_t = 1$ and $b_k[t] = 0$ whenever $\sigma_t = -1$.

Proof of (ii). We now construct an online learner for $\mathcal{F}$. Let $(x_1, y_1) \ldots, (x_T, y_T) \in \mathcal{X} \times \mathcal{Y}$ denote the data stream. Since y_t is an element of unit ball of $\mathcal{V}$, we can write $y_t = \sum_{n=1}^{\infty} c_n(t) e_n$ such that $\sum_{n=1}^{\infty} c_n^2(t) \leq 1$. For each $t \in [T]$, define a set of indices $S_t = \{n \in \mathbb{N} : |c_n(t)| \geq \frac{1}{2\sqrt{T}}\}$. Since

$$1 \geq \|y_t\|^2 = \sum_{n=1}^{\infty} c_n^2(t) \geq \sum_{n \in S_t} c_n^2(t) \geq \sum_{n \in S_t} \frac{1}{4T} = \frac{|S_t|}{4T},$$

we have $|S_t| \leq 4T$. Let $\text{sort}(S_i)$ denote the ordered list of size $4T$ that contains elements of S_i in descending order. If S_i does not contain $4T$ indices, append 0's to the end of $\text{sort}(S_i)$. We let $\text{sort}(S_i)[j]$ denote the j^{th} element of the ordered list $\text{sort}(S_i)$.

For each $i \in [T]$ and $j \in [4T]$, define an expert E_i^j such that

$$E_i^j(x_t) = \begin{cases} 0, & t \leq i \\ f_k(x_t), & t > i \end{cases}, \qquad \text{where } k = \text{sort}(S_i)[j].$$

An online learner $\mathcal{A}$ for $\mathcal{F}$ runs multiplicative weights algorithm using the set of experts $\mathcal{E} = \{E_i^j \mid i \in [T], j \in [4T]\}$. It is easy to see that $\|f_k(x)\| \leq 1$ for all $x \in \mathcal{X}$. Thus, for any $\hat{y}_t, y_t \in \mathcal{Y}$, we have $\|\hat{y}_t - y_t\|^2 \leq 4$. Thus, for an appropriately chosen learning rate, the multiplicative weights algorithm guarantees (see Theorem 21.11 in Shalev-Shwartz and Ben-David [2014]) that the regret of $\mathcal{A}$ satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\mathcal{A}(x_t) - y_t\|^2\right] \leq \inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \|E(x_t) - y_t\|^2 + 4\sqrt{2T \ln(|\mathcal{E}|)}.$$

Note that $|\mathcal{E}| \leq 4T^2$, which implies $4\sqrt{2T \ln(|\mathcal{E}|)} \leq 8\sqrt{T \ln(2T)}$. We now show that

$$\inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \|E(x_t) - y_t\|^2 \leq \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \|f(x_t) - y_t\|^2 + 2.$$

Together, these two inequalities imply that the expected regret of $\mathcal{A}$ is $\leq 2 + 8\sqrt{T \ln(2T)}$. The rest of the proof is dedicated to proving the latter inequality.

Let $f_{k^\star} \in \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \|f(x_t) - y_t\|^2$. Let $t^\star \in [T]$ be the first time point such that $k^\star \in S_{t^\star}$ and suppose it exists. Let $r^\star \in [4T]$ be such that $k^\star = \text{sort}(S_{t^\star})[r^\star]$. By definition of the experts, we have

$$E_{t^\star}^{r^\star}(x_t) = f_{k^\star}(x_t) \quad \text{for } t > t^\star,$$

thus implying that $\sum_{t>t^\star} \left\| E^{r^\star}_{t^\star}(x_t) - y_t \right\|^2 = \sum_{t>t^\star} \|f_{k^\star}(x_t) - y_t\|^2$. Therefore, it suffices to show that

$$\sum_{t\le t^\star} \left\| E^{r^\star}_{t^\star}(x_t) - y_t \right\|^2 \le \sum_{t\le t^\star} \|f_{k^\star}(x_t) - y_t\|^2 + 2.$$

As $E^{r^\star}_{t^\star}(x_t) = 0$ for all $t \le t^\star$, proving the inequality above is equivalent to showing

$$\sum_{t\le t^\star} \|y_t\|^2 \le \sum_{t\le t^\star} \|f_{k^\star}(x_t) - y_t\|^2 + 2.$$

Since $\|y_{t^\star}\|^2 \le 1$, we trivially have $\|y_{t^\star}\|^2 \le \|f_{k^\star}(x_{t^\star}) - y_{t^\star}\|^2 + 1$. Thus, by expanding the squared norm, the problem reduces to showing

$$\sum_{t<t^\star} \left(2\left\langle f_{k^\star}(x_t), y_t \right\rangle - \|f_{k^\star}(x_t)\|^2 \right) \le 1.$$

We prove the inequality above by establishing

$$2\left\langle f_{k^\star}(x_t), y_t \right\rangle - \|f_{k^\star}(x_t)\|^2 \le \frac{1}{T} \quad \text{for all } t < t^\star.$$

Let $x_t = \sum_{n=1}^\infty \alpha_n(t) e_n$. We have $f_{k^\star}(x_t) = \sum_{n=1}^\infty b_{k^\star}[n] \left\langle x_t, e_n \right\rangle e_{k^\star} = \left(\sum_{n=1}^\infty b_{k^\star}[n]\alpha_n(t)\right) e_{k^\star}$. Defining $a_{k^\star}(t) = \left(\sum_{n=1}^\infty b_{k^\star}[n]\alpha_n(t)\right)$, we can write

$$f_{k^\star}(x_t) = a_{k^\star}(t) e_{k^\star} \quad \text{and} \quad \|f_{k^\star}(x_t)\| = |a_{k^\star}(t)|.$$

So, it suffices to show that $2\, a_{k^\star}(t)\, c_{k^\star}(t) - |a_{k^\star}(t)|^2 \le \frac{1}{T}$ for all $t < t^\star$. To prove this inequality, we consider the following two cases:

(I) Suppose $|a_{k^\star}(t)| > 2|c_{k^\star}(t)|$. Then, $2\, a_{k^\star}(t)\, c_{k^\star}(t) - |a_{k^\star}(t)|^2 < |a_{k^\star}(t)|^2 - |a_{k^\star}(t)|^2 = 0$.

(II) Suppose $|a_{k^\star}(t)| \le 2|c_{k^\star}(t)|$. Then, $2\, a_{k^\star}(t)\, c_{k^\star}(t) - |a_{k^\star}(t)|^2 \le 4\, |c_{k^\star}(t)|^2 < 4 \left(\frac{1}{2\sqrt{T}}\right)^2 = \frac{1}{T}$ because $k^\star \notin S_t$ for all $t < t^\star$.

In either case, $2\, a_{k^\star}(t)\, c_{k^\star}(t) - |a_{k^\star}(t)|^2 \le \frac{1}{T}$ for all $t < t^\star$.

Finally, suppose that such a $t^\star$ does not exist. Then, our analysis for the case $t \le t^\star$ above shows that the expert E^1_T that predicts $E^1_T(x_t) = 0$ for all $t \le T$ satisfies $\sum_{t=1}^T \|E^1_T(x_t) - y_t\|^2 \le \sum_{t=1}^T \|f_{k^\star}(x_t) - y_t\|^2 + 2$. ■

5.5.1 Batch Learnability without Uniform Convergence

Although we state Theorem 24 in the online setting, an analogous result also holds in the batch setting. To establish the batch analog of Theorem 24, consider f_k defined in (5.1) and

define a class $\mathcal{F} = \{f_k \mid k \in \mathbb{N}\} \cup \{f_0\}$ where $f_0 = 0$. This is the same class considered in the proof of Theorem 24. Recall that in our proof of Theorem 24 (i), we choose a sequence of labeled examples $\{e_t, 0\}_{t=1}^T$ that is independent of the sequence of Rademacher random variables $\{\sigma_t\}_{t=1}^T$. Thus, our proof shows that the i.i.d. version of the Rademacher complexity of $\mathcal{F}$, where the labeled samples are independent of Rademacher variables, is also lower bounded by $\frac{T}{2}$. This implies that the class $\mathcal{F}$ does not satisfy the uniform law of large numbers in the i.i.d. setting. However, using the standard online-to-batch conversion techniques, we can convert our online learner for $\mathcal{F}$ to a batch learner for $\mathcal{F}$ [Cesa-Bianchi et al., 2004]. This shows a separation between uniform convergence and batch learnability of bounded linear operators.

5.6 Discussion

In this work, we study the online learnability of bounded linear operators between two infinite-dimensional Hilbert spaces. In Theorems 21 and 22, we showed that

$$c^2\, T^{1-\frac{1}{p}} \leq \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}_p) \leq 6c^2\, T^{\max\left\{\frac{1}{2}, 1-\frac{1}{p}\right\}},$$

for every $p \in [1, \infty]$, where $\mathcal{F}_p := \{f \in S_p(\mathcal{V}, \mathcal{W}) : \|f\|_p \leq c\}$. Note that the upperbound and lowerbound match $p \geq 2$. However, for $p \in [1, 2)$, the upperbound saturates at $\sqrt{T}$, while the lower bound gets progressively worse as p decreases. Given this gap, we leave it open to resolve the following question.

$$\text{What is } \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}_p) \text{ for } p \in [1, 2)?$$

We conjecture that lowerbound is loose for $p \in [1, 2)$, and one can obtain faster rates using some adaptation of the seminal Vovk-Azoury-Warmuth forecaster [Vovk, 2001, Azoury and Warmuth, 2001].

Section 5.5 shows a separation between sequential uniform convergence and online learnability for bounded linear operators. The separation is exhibited by a class that lies in $S_1(\mathcal{V}, \mathcal{W})$, but is *not* uniformly bounded. In this work, we established that there is no separation between online learnability and sequential uniform convergence for any subset of $S_p(\mathcal{V}, \mathcal{W})$ with uniformly bounded p-Schatten norm for $p \in [1, \infty)$. However, it is unknown whether this is also true for $S_\infty(\mathcal{V}, \mathcal{W})$. This raises the following natural question.

$$\text{Is } \mathrm{Rad}_T(\mathcal{F}) = o(T) \text{ if and only if } \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}) = o(T) \text{ for every } \mathcal{F} \subseteq \{f \in S_\infty(\mathcal{V}, \mathcal{W}) \mid \|f\|_\infty \leq c\}?$$

Finally, in this work, we showed that a uniform bound on the p-Schatten norm for any $p \in [1, \infty)$ is sufficient for online learnability. However, the example in Theorem 24 shows that a uniform upper bound on the norm is not necessary for online learnability. Thus, it is an interesting future direction to fully characterize the landscape of learnability for bounded linear operators. In addition, it is also of interest to extend these results to nonlinear operators.

Part II
Operator Learning: A Learning Theoretic Perspective

CHAPTER 6

Controlling Statistical, Discretization, and Truncation Errors in Learning Fourier Linear Operators

In this chapter[1], we study learning-theoretic foundations of operator learning, using the linear layer of the Fourier Neural Operator architecture as an illustrative example. To briefly recap our discussion from Chapter 1, the central goal of operator learning is to use statistical methods to estimate an unknown operator between function spaces. A primary application is the development of fast data-driven methods to approximate the solution operator of partial differential equations (PDEs) [Li et al., 2021, Kovachki et al., 2023]. For example, consider our running example of the heat equation

$$\frac{\partial u}{\partial t} = \tau \, \nabla^2 u,$$

where $u : [0,1]^d \to \mathbb{R}$ vanishes on the boundary. The solution operator for this equation is a linear operator $\exp(\tau t \nabla^2) := \sum_{k=0}^{\infty} (\tau t \, \nabla^2)^k / k!$. Fixing some time point (say $t = 1$), our objective is to learn the solution operator $\mathcal{L} := \exp(\tau \nabla^2)$.

Given the training data $(v_1, w_1), \ldots, (v_n, w_n)$ where $w_i = \mathcal{L} v_i$, operator learning entails using statistical methods to estimate the solution operator $\widehat{\mathcal{L}}_n$. Then, given a new input v, one can get the approximate solution $\widehat{w} = \widehat{\mathcal{L}}_n v$. The goal is to develop the estimation rule such that $\widehat{w}$ is close to the actual solution $w = \mathcal{L} v$ under some appropriate metric.

Traditionally, given an input function v, one would use numerical methods such as finite differences to get a numerical solution. The solver starts from scratch for every new function v of interest and can be computationally slow and expensive. This can be limiting in some

[1]This chapter is based on: Unique Subedi and Ambuj Tewari (2025). *Controlling Statistical, Discretization, and Truncation Errors in Learning Fourier Linear Operators.* Transactions on Machine Learning Research.

applications such as engineering design where the solution needs to be evaluated for many different instances of the input functions. To solve this problem, operator learning aims to learn surrogate models that significantly increase speed for solution evaluation compared to traditional solvers while sacrificing a small degree of accuracy.

In this work, rather than focusing on specific PDEs, we adopt a broader perspective and study the learning-theoretic foundations of operator learning. For this task, we use the linear layer of the influential Fourier Neural Operator (FNO) architecture proposed by Li et al. [2021] as our model problem. While our results offer some practical and theoretical insights into the FNO, it is important to emphasize that our primary objective is neither to advance the practical implementation of operator learning nor to develop deeper insights into the FNO architecture itself. Instead, our objective is to rigorously understand the statistical learning aspects of the operator learning paradigm. Our primary goal is to understand how operator learning differs from traditional machine learning settings and to identify the new techniques required to build a rigorous learning-theoretic foundation for this emerging area.

To this end, we start by identifying the distinct types of errors that are unique to operator learning. In addition to the standard statistical error arising from a finite sample size, operator learning introduces a discretization error due to the functional data being available only on a finite grid of domain points. Furthermore, ignoring high-frequency Fourier modes lead to a truncation error. Lastly, we introduce a Discrete Fourier Transform (DFT)-based estimator for our model problem and demonstrate how these errors can be systematically quantified for this estimator.

6.1 Neural Operators

To formally define our problem setting, we need to introduce neural operators from [Kovachki et al., 2023]. Let $\mathcal{V}$ be a vector space of functions from a bounded subset $\mathcal{X} \subseteq \mathbb{R}^d$ to $\mathbb{R}^p$, and $\mathcal{W}$ to be a vector space of functions from $\mathcal{Y} \subseteq \mathbb{R}^d$ to $\mathbb{R}^q$. Given a function $v \in \mathcal{V}$, a single layer of neural operator $\mathtt{N}_t : \mathcal{V} \to \mathcal{W}$ is a mapping such that

$$(\mathtt{N}_t v)(y) = \sigma\Big(\left(\mathcal{K}_{\theta_t} v\right)(y) + b_t(y) \Big) \qquad \forall y \in \mathcal{Y},$$

where $\left(\mathcal{K}_{\theta_t} v\right)(y) = \int_{\mathcal{X}} k_{\theta_t}(y, x)\, v(x)\, dx$. The function $b_t : \mathcal{Y} \to \mathbb{R}^q$ is a bias function in $\mathcal{W}$, the function $\sigma : \mathbb{R}^q \to \mathbb{R}^q$ is a point-wise non-linear activation, and the transformation $v \mapsto \mathcal{K}_{\theta_t} v$ is an integral kernel transform of v using some kernel $k_{\theta_t} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^{q \times p}$. These layers are then composed to get a neural operator architecture.

Parametrizing $\mathcal{K}_{\theta_t}$ in terms of k_{θ_t} can be impractical due to the computational cost of calculating the integral in for each layer. Thus, a significant area of research in neural

operators focuses on developing innovative parametrizations of $\mathcal{K}_{\theta_t}$ that facilitate more efficient computation. One such parametrization gives rise to a well-known architecture called the Fourier Neural Operator.

6.1.1 Fourier Neural Operator (FNO)

In this section, we present a brief, non-rigorous overview of the FNO. A more formal treatment, along with new insights into its parametrization, is provided in Appendix E.1.

We consider the setup from the work of Li et al. [2021]. Let $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d \simeq [0,1]^d$ be a d-dimensional periodic torus. Assume the kernel k_θ is translation invariant–that is, $k_\theta(y,x) = k_\theta(y-x)$. This implies that $\mathcal{K}_\theta$ is a convolution operator. Then, the Convolution Theorem implies that

$$\mathcal{K}_\theta v = \mathcal{F}^{-1}\Big(\mathcal{F}(k_\theta)\,\mathcal{F}(v)\Big),$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are Fourier and Inverse Fourier transform respectively. The key insight in FNO is that instead of parametrizing the kernel k_θ, we parametrize its Fourier transform $\mathcal{F}(k_\theta)$ directly. That is, we parametrize the kernel transform operator as

$$\mathcal{K}_\beta v = \mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big).$$

This is a linear operator and will be referred to as *Fourier linear operator*. When $|\Lambda_\beta(\cdot)|_{\ell^1} < \infty$, we can write this

$$(\mathcal{K}_\beta v)(y) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,y\rangle}\,\Lambda_\beta(m)\ (\mathcal{F}v)(m) \qquad \forall y \in \mathcal{Y}.$$

There are two practical challenges in implementing the operator $\mathcal{K}_\beta$. First, the implementation involves an infinite sum over $\mathbb{Z}^d$. Second, the Fourier transform $\mathcal{F}v$ cannot be computed exactly since the function v is only available on a finite grid of domain points. To address the first challenge, a large $K \in \mathbb{N}$ is fixed and we sum only over $m \in \mathbb{Z}^d$ such that $|m|_{\ell^\infty} \le K$. The second challenge is addressed by approximating $\mathcal{F}v$ using the Discrete Fourier Transform (DFT) of v over the finite grid of domain points, which can be efficiently computed using Fast Fourier Transform (FFT) algorithms. The solution to the second challenge motivates our DFT-based least-squares estimator.

6.1.2 Our Contribution

In this work, we study the error bounds of learning the operator class $\{v \mapsto \mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big) : \beta \in \mathcal{B}\}$, where $\mathcal{B}$ is some parameter space that will be specified later. We study this simple setup to conceptually separate the paradigm of operator learning from its commonly used instantiation using neural network architectures. By eliminating the complexities associated with neural networks, studying this linear class can provide insights that are broadly applicable to both algorithm design and theoretical analysis.

We assume that $\mathcal{V} = \mathcal{W} = \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, a $(s, 2)$-Sobolev space of real-valued functions defined on the d-dimensional periodic torus. See Section 6.4.3 for an explanation on why $\mathcal{V}$ and $\mathcal{W}$ need to be function space with higher-order smoothness to achieve a vanishing error. We work in the agnostic (misspecified) setting and analyze the DFT-based least-squares estimator (see Section 6.4.2 for more details). Specifically, for some universal constant $c_1 > 0$, we show that the excess risk of the DFT-based least-squares estimator is at most

$$c_1 \left(\frac{1}{\sqrt{n}} + \frac{1}{N^s} + \frac{1}{K^{2s}} \right).$$

The term $1/\sqrt{n}$ is the usual *statistical/estimation* error due to a finite sample size. The term $1/K^{2s}$ is the *truncation* error incurred because the learner only works with the low Fourier modes m such that $|m|_{\ell^\infty} \leq K$. Finally, the term $1/N^s$ is the *discretization* error due to functions being accessible to the learner only on the uniform grid of size N^d of $[0, 1]^d$. This error quantifies the generalization error of an estimator trained on a grid of size N^d but evaluated at full resolution ($N \to \infty$). It formalizes the concept of multiresolution generalization (operators trained at lower resolution have good generalization even when evaluated in higher resolution)–a phenomenon frequently observed in practice [Li et al., 2021, Section 5].

Additionally, we establish the lower bound on excess risk, showing that it is at least

$$c_2 \left(\frac{1}{n} + \frac{1}{N^{2s}} + \frac{1}{K^{2s}} \right)$$

for some $c_2 > 0$. Our analysis is non-asymptotic and the precise form of the constants c_1 and c_2 are provided in Theorems 25 and 26 respectively.

6.2 Related Works

After Li et al. [2021] proposed Fourier Neural Operators (FNOs), there has been a surge of interest in this architecture. The number of applied works is too vast and not entirely relevant to list here, so we focus on related theoretical works. One of the earliest theoretical analyses of FNOs was the universal approximation result by Kovachki et al. [2021].

More closely related to our work is a recent study on the sample complexity of various operator classes, including FNOs, by Kovachki et al. [2024a]. Their scope is broader than ours as they address a general class of nonlinear operators. However, their results do not imply ours. They treat the truncation parameter K as a part of the model rather than a variable that the learning algorithm can choose. Their error bounds are based on metric entropy analysis, which leads to a suboptimal dependence on K and the input dimension d. Specifically, their bounds break down as $K \to \infty$ and suffer from the curse of dimensionality in d. In contrast, our work establishes statistical error bounds using sharp Rademacher analysis, avoiding both dependence on K and the curse of dimensionality in d. An interesting future direction is to extend our Rademacher-based analysis to capture function classes at the level of generality considered in Kovachki et al. [2024a]. We also note that Rademacher-based analysis has also been used by Raman et al. [2024b], Tabaghi et al. [2019] to study Schatten operators between Hilbert spaces. Kim and Kang [2024] also bound the Rademacher complexity of FNOs, but the bound is rather loose and even non-vanishing in some cases. Finally, the analysis by Liu et al. [2024] and Liu et al. [2025] also share our motivation of quantifying the statistical error in operator learning.

A recent work by Lanthaler et al. [2024] aligns with our goal of quantifying the discretization error of FNOs. In fact, the key ideas used in the proof of lemmas 26 and 27 in discretization error analysis is drawn from their work. However, the nature of their results differs from ours. To discuss the difference precisely, let Ψ be a trained Fourier Neural Operator and v be an input function available to the learner only over a discrete grid of domain points of size N. Denote v^N as the set of discrete values of v available to the learner. Lanthaler et al. [2024] bound the term $\|\Psi v - \Psi v^N\|$, quantifying the error incurred in the forward pass due to the function being available only over a discrete grid. Essentially, this only captures errors incurred during the test time but does not quantify the discretization error incurred during training. In contrast, our focus is on quantifying the generalization error of an operator trained on a grid of size N^d but evaluated at full resolution ($N \to \infty$), a type of multiresolution generalization [Li et al., 2021, Section 5].

Finally, we also note that or setup is closely related to the function-to-function regression often studied in the functional data analysis (FDA) literature. For example, the linear layer

of a neural operator $v \mapsto \mathcal{K}v + b$ is a well-studied model in FDA [Wang et al., 2016, Equation 15]. Even a single layer of a neural operator $v \mapsto \sigma(\mathcal{K}v + b)$ has been examined in FDA literature as multi-index functional models [Wang et al., 2016, Equation 13], [Chen et al., 2011]. That said, the overall goal of the FDA differs slightly from that of operator learning. In FDA, the focus is on statistical inference, typically using RKHS-based frameworks under some assumptions about the data-generating process. As a result, FDA methods often do not always scale to large datasets. In contrast, operator learning primarily aims at prediction, seeking to develop surrogate models that approximate numerical PDE solvers [Li et al., 2021, Kovachki et al., 2024b]. The emphasis is on creating computationally efficient methods that can be used to train large models and handle large datasets. However, we believe that the intersection of these two fields can benefit both. The theoretical tools developed in FDA literature can be applied to the analysis of operator learning methods, while the computational advances in operator learning can help scale FDA methods.

6.3 Preliminaries

6.3.1 Notation

Let $\mathbb{N}$ be natural numbers and $\mathbb{Z}$ be integers. Define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. $\mathbb{R}$ and $\mathbb{C}$ denote real and complex numbers respectively. For any $\eta \in \mathbb{R}^d$, we let $|\eta|_\infty := \max_{1 \le i \le d} |\eta_i|$ denote the ℓ^∞ norm. For any complex number $z \in \mathbb{C}$ such that $z = a + b\,\mathrm{i}$, we use $|z| = \sqrt{a^2 + b^2}$ and $\bar{z} = a - \mathrm{i}\,b$ denotes complex conjugate. For any $x, y \in \mathbb{R}^d$, the term $\langle x, y \rangle$ denotes the Euclidean inner product. Occasionally, the inner products on other Hilbert spaces such as L^2 will be distinguished from the Euclidean one with the subscript such as $\langle \cdot, \cdot \rangle_{L^2}$. However, when the context is clear, we will use $\langle \cdot, \cdot \rangle$ to denote canonical inner products on the respective Hilbert spaces.

Given $K \in \mathbb{N}$, we define $\mathbb{Z}^d_{\le K} = \{m \in \mathbb{Z}^d : |m|_\infty \le K\}$ and $\mathbb{Z}^d_{>K} := \mathbb{Z}^d \backslash \mathbb{Z}^d_{\le K}$. For a sequence $s := \{s_k\}_{k \in \mathbb{Z}^d}$, we will also use $|s|_{\ell^p}$ to denote the ℓ^p norm of s. Moreover, we let $\mathbb{T}^d \simeq [0, 1]^d$ denote a d-dimensional periodic torus. See [Grafakos, 2008, Chapter 3] for more details on the torus. Throughout the paper, for any $m \in \mathbb{Z}^d$, we use $\varphi_m : \mathbb{T}^d \to \mathbb{C}$ to denote the function $\varphi_m(x) = e^{2\pi\,\mathrm{i}\langle m, x \rangle}$. The sequence $\{\varphi_m\}_{m \in \mathbb{Z}^d}$ will be referred to as Fourier basis.

6.3.2 L^2-Spaces and Fourier Analysis

Define

$$L^2(\mathbb{T}^d, \mathbb{R}) := \left\{ u : \mathbb{T}^d \to \mathbb{R} \mid \int_{\mathbb{T}^d} |u(x)|^2 \, dx < \infty \right\}.$$

Recall that $L^2(\mathbb{T}^d, \mathbb{R})$ is a Hilbert space with inner-product $\langle u, v \rangle_{L^2} = \int_{\mathbb{T}^d} u(x)\, v(x)\, dx$. The norm induced by this inner product will be denoted as $\|\cdot\|_{L^2}$. The sequence $\{\varphi_m\}_{m \in \mathbb{Z}^d}$ forms an orthonormal basis for $L^2(\mathbb{T}^d, \mathbb{R})$. That is, for any $u \in L^2(\mathbb{T}^d, \mathbb{R})$, we can write $u = \sum_{m \in \mathbb{Z}^d} \langle u, \varphi_m \rangle_{L^2}\, \varphi_m$, where the convergence is in L^2-norm. The celebrated Parseval's identity then implies that $\|u\|_{L^2}^2 = \sum_{m \in \mathbb{Z}^d} |\, \langle u, \varphi_m \rangle_{L^2}\, |^2$.

Since $\mathbb{T}^d$ is identified with a bounded set $[0,1]^d$, the condition $u \in L^2(\mathbb{T}^d, \mathbb{R})$ implies that u is integrable. That is, $\int_{\mathbb{T}^d} |u(x)|\, dx < \infty$. For integrable functions, $\mathcal{F}$ denotes the Fourier transform operator such that $\mathcal{F}u : \mathbb{Z}^d \to \mathbb{C}$ is a complex-valued function on $\mathbb{Z}^d$ defined as

$$(\mathcal{F}u)(m) = \int_{\mathbb{T}^d} u(x)\, e^{-2\pi\, \mathrm{i} \langle m, x \rangle}\, dx.$$

Note that we have $(\mathcal{F}u)(m) = \langle u, \varphi_m \rangle$. We let $\mathcal{F}^{-1}$ denote the operator that satisfies $(\mathcal{F}^{-1}\mathcal{F})\,(u) = u$. $\mathcal{F}^{-1}$ will be referred to as inverse Fourier transform. Note that even when u is a real-valued function, $\langle u, \varphi_m \rangle$ may lie in $\mathbb{C}$.

6.3.3 Sobolev Spaces

Fix $s \in \mathbb{N}$ and define

$$\mathcal{H}^s(\mathbb{T}^d, \mathbb{R}) = \Big\{ u \in L^2 \Big|\, \partial^k u \in L^2(\mathbb{T}^d, \mathbb{R}) \text{ for all } k \in \mathbb{N}_0^d \;\&\; |k|_\infty \leq s \Big\}.$$

Here, $\partial^k u$ is the k^{th} partial derivatives. The space $\mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, also referred to as $(s,2)$-Sobolev space, is a Hilbert space with an inner product

$$\langle u, v \rangle_{\mathcal{H}^s} := \sum_{k \in \mathbb{N}_0^d\, :\, |k|_\infty \leq s} \Big\langle \partial^k u, \partial^k v \Big\rangle_{L^2},$$

which naturally induces the norm $\|u\|_{\mathcal{H}^s} := \sqrt{\langle u, u \rangle_{\mathcal{H}^s}}$. In this paper, we often assume that $s > d/2$. This ensures that (see Lemma 25) $\sum_{m \in \mathbb{Z}^d} |\, \langle u, \varphi_m \rangle\, | < \infty$, which implies uniform convergence of the Fourier series over $\mathbb{T}^d$.

Note that it is more common to define Sobolev spaces with multi-indices k such that $|k|_1 \leq s$. We chose the restriction $|k|_\infty \leq s$ simply for the convenience of computation. However, as d is finite and all ℓ_p norms on a d-dimensional space are equivalent up to a factor of d.

6.4 Learning Fourier Linear Operators

In this section, we establish excess risk bounds of learning the operator class $\{v \mapsto \mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big)\ :\ \beta \in \mathcal{B}\}$, where $\mathcal{B}$ is some parameter space. Here, we only consider the case where $\mathcal{V}, \mathcal{W} \subseteq L^2(\mathbb{T}^d, \mathbb{R})$. This is different from the usual setting in the literature, where $\mathcal{V}$ and $\mathcal{W}$ are Banach spaces of vector-valued functions. First, a significant number of PDEs of practical interest describe how scalar-valued functions evolve. Since not much is known from a theoretical standpoint even for scalar-valued functions, we believe that this is a good start. Second, assuming $\mathcal{V}, \mathcal{W}$ to be a subset of L^2 (a Hilbert space) does not result in any meaningful loss of generality from a practical standpoint. In practice, one must discretize the domain and work with function values over a discrete grid, which effectively requires a bounded domain. This essentially means working with bounded functions on a bounded domain, all of which are L^2 integrable.

For scalar-valued functions, Λ_β is a scalar-valued function defined on modes $\mathbb{Z}^d$. Since the function is only defined on a countable domain, we can also represent it by a scalar-valued sequence $\{\Lambda_\beta(m)\}_{m\in\mathbb{Z}^d}$. Henceforth, we will drop the β and just write $\{\lambda_m\}_{m\in\mathbb{Z}^d}$, denoting λ_m's to be the parameters themselves. For the convenience of notation, we will also use λ to denote the sequence $\{\lambda_m\}_{m\in\mathbb{Z}^d}$ and write $\mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(\cdot)\Big)$. Fixing some $C > 0$, the class of interest can be written as

$$\Big\{v \mapsto \mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(v)\Big)\ :\ |\lambda|_{\ell^1} \le C\Big\}.$$

A starting point of our work is the following result on the decomposition of Fourier linear operators.

Proposition 1. *If $\lambda \in \ell^1(\mathbb{Z}^d)$, then*

$$\mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(u)\Big) = \sum_{m\in\mathbb{Z}^d} \lambda_m\ \varphi_m\ \langle \varphi_{-m}, u\rangle_{L^2}\,, \tag{6.1}$$

where the equality holds for every $u \in L^2(\mathbb{T}^d, \mathbb{R})$.

Here, $\varphi_m \otimes \varphi_{-m}$ is a rank-1 operator such that $(\varphi_m \otimes \varphi_{-m})(u) = \langle \varphi_{-m}, u\rangle_{L^2}\ \varphi_m$. The equality in (6.1) means $\mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(u)\Big) = \sum_{m\in\mathbb{Z}^d} \lambda_m\ \varphi_m\ \langle \varphi_{-m}, u\rangle_{L^2}$ for all $u \in L^2(\mathbb{T}^d, \mathbb{R})$, where the sum converges uniformly over $x \in \mathbb{T}^d$. We provide the proof of Proposition 1 in Appendix E.2.

Given Proposition 1, we can write our class as $\{\sum_{m\in\mathbb{Z}^d} \lambda_m\ \varphi_m \otimes \varphi_{-m}\ :\ |\lambda|_{\ell^1} \le C\}$. This representation is preferable for the following reasons. First, it highlights the fact that the Fourier basis is just one of the design choices for singular vectors that may be replaced with

any other orthonormal sequences. Second, this representation also allows us to drop the constraint that $\lambda \in \ell^1$, which is a rather artificial constraint required only to ensure that the operator $\mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(\cdot)\Big)$ is a well-defined object. However, $\sum_{m\in\mathbb{Z}^d} \lambda_m\ \varphi_m \otimes \varphi_{-m}$ is still well-defined even when $\lambda \in \ell^\infty$ (in fact, it is a bounded operator). Therefore, for some fixed $C > 0$, we will instead study the class of operators

$$\mathcal{T} := \left\{ \sum_{m\in\mathbb{Z}^d} \lambda_m\ \varphi_m \otimes \varphi_{-m} \;\Big|\; |\lambda|_{\ell^\infty} \le C \right\}.$$

Since the class $\Big\{v \mapsto \mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(\cdot)\Big) :\ |\lambda|_{\ell^1} \le C\Big\}$ is contained in the class $\mathcal{T}$, any guarantee (in terms of upper bound) for $\mathcal{T}$ also holds for the ℓ^1 constrained class.

Remark. The class $\mathcal{T}$ should remind readers of de Hoop et al. [2023], who also consider the problem of singular value inference of an operator under fixed singular vectors. However, their setting differs from ours in two significant ways. First, they only consider the well-specified setting with an additive noise model, whereas we adopt a fully agnostic viewpoint. Second, they do not account for possible discretization errors, assuming that their input and output functions are fully available to the learner.

6.4.1 Problem Setting and Error Types

We adopt the framework of statistical learning and study the rates of error in learning the class $\mathcal{T}$. In statistical learning, the learner is provided with $n \in \mathbb{N}$ i.i.d samples $S_n = \{(v_i, w_i)\}_{i=1}^n$ from some unknown distribution μ on $\mathcal{V} \times \mathcal{W}$. We adopt a fully agnostic viewpoint and do not make any assumptions about the data-generating process. Next, using the sample S_n and some prespecified learning rule, the learner then finds an estimator $\widehat{T} \in \mathcal{T}$. We will abuse notation and denote $\widehat{T}$ to be both the learning rule and the estimator output by the learner. For an estimator $\widehat{T}$, we can define its expected excess risk as

$$\mathcal{E}_n(\widehat{T}, \mathcal{T}, \mu) = \mathop{\mathbb{E}}_{S_n\sim\mu^n} \left[\mathop{\mathbb{E}}_{(v,w)\sim\mu} \left[\|\widehat{T}v - w\|_{L^2}^2 \right] - \inf_{T\in\mathcal{T}} \mathop{\mathbb{E}}_{(v,w)\sim\mu} \left[\|Tv - w\|_{L^2}^2 \right] \right].$$

Formally, the goal of the learner is to output the estimator such that $\mathcal{E}_n(\widehat{T}, \mathcal{T}, \mu) \to 0$ as $n \to \infty$. In traditional settings, the excess risk $\mathcal{E}_n(\widehat{T}, \mathcal{T}, \mu)$ is usually referred to as the statistical error of the learner. This error arises because the learner is trying to find the optimal operator in $\mathcal{T}$ for distribution μ while only having access to finitely many samples from the distribution. However, unlike traditional statistical learning settings, in operator

learning, there are two additional errors beyond the statistical error: discretization error and truncation Error.

Discretization Error: The discretization error arises because the learner only has access to $(v_i, w_i) \sim \mu$ over some discrete grid of domain points. In this work, we assume that each v_i and w_i are available on a uniform grid

$$\mathrm{G} := \left\{ m/N : m \in \{0, \ldots, N-1\}^d \right\}$$

of $[0,1]^d$ for some prespecified $N \in \mathbb{N}$. That is, the learner only has access to $\{v_i(x) : x \in \mathrm{G}\}$ and $\{w_i(x) : x \in \mathrm{G}\}$. Although other grids are also used in practice, the use of FNO requires uniform griding. This is because the main benefit of FNO is its computationally efficient approximation of Fourier transform through fast Fourier transform (FFT) algorithms, which requires uniform grids.

Truncation Error: To see where the truncation error comes from, note that the representation of any estimator $T \in \mathcal{T}$ requires specifying an infinite sequence $\{\lambda_m\}_{m \in \mathbb{Z}^d}$. However, the infinite sequence cannot be implemented in a computer. Thus, for a practical implementation [Li et al., 2021], one picks a large $K \in \mathbb{N}$ and specifies the finite rank operator

$$T_K = \sum_{m \in \mathbb{Z}^d_{\leq K}} \lambda_m \, \varphi_m \otimes \varphi_{-m}.$$

While the truncation error is specific to our class of interest $\mathcal{T}$, a similar "truncation" error occurs in any model class. Such error arises because operator learning is inherently an infinite-dimensional problem, yet any computation we perform is limited to some finite-dimensional subspace.

6.4.1.1 Further Connection to FDA.

The operator T_K is related to functional PCA-based estimators common in the FDA literature. Given n i.i.d. function pairs $\{(v_i, w_i)\}_{i \leq n}$, the least-squares estimator solves $\sum_{i=1}^n w_i \otimes v_i = L \circ (\sum_{i=1}^n v_i \otimes v_i)$, which is under-specified in infinite-dimensional spaces. To address this, one computes a pseudo-inverse $(\sum_{i=1}^n v_i \otimes v_i)^\dagger$ by fixing an orthonormal basis $\{\psi_t\}_{t \in \mathbb{N}}$. With eigendecomposition $\sum_{i=1}^n v_i \otimes v_i = \sum_{t \geq 1} \eta_t \, \psi_t \otimes \psi_t$, the pseudo-inverse becomes $\sum_{t \geq 1} \mathbb{1}[\eta_t > 0] \eta_t^{-1} \psi_t \otimes \psi_t$, yielding the estimator $\widehat{L} = (\sum_{i=1}^n w_i \otimes v_i) \left(\sum_{t \geq 1} \mathbb{1}[\eta_t > 0] \, \eta_t^{-1} \, \psi_t \otimes \psi_t \right)$. In practice, the sum is truncated at some $t \leq \tau$.

Estimators of this type have been studied in works such as Hörmann and Kidziński [2015], Reimherr [2015], Yao et al. [2005] under well-specified models. These approaches generally re-

quire learning the basis functions ψ_t's and the truncation parameter from the data to achieve the guarantees established in these studies, which often introduces significant computational challenges. In contrast, we work in the potentially misspecified (agnostic) setting, and K depends only on the sample size n to achieve $\sqrt{n}$-risk consistency. Additionally, FDA-based approaches typically assume exact access to the functions, which is unrealistic in practice. Instead, we explicitly account for the discretization error that arises when functions are only available on a finite grid.

6.4.2 A Constrained Least-Squares Estimator

In this section, we specify our primary estimator of interest. Let $T = \sum_{m\in\mathbb{Z}^d} \lambda_m \, \varphi_m \otimes \varphi_{-m}$. For any $v \in \mathcal{V}$, we have $Tv = \sum_{m\in\mathbb{Z}^d} \lambda_m \, \langle \varphi_{-m}, v\rangle \, \varphi_m$. As we only require ℓ^∞ norm of λ to be bounded by C, we only get the convergence of the sum $\sum_{m\in\mathbb{Z}^d} \lambda_m \, \langle \varphi_{-m}, v\rangle \, \varphi_m$ in L^2 norm rather than uniform. Since $\{\varphi_m\}_{m\in\mathbb{Z}^d}$ is an orthonormal basis of $L^2(\mathbb{T}^d, \mathbb{R})$, Parseval's identity implies

$$\begin{aligned}\|Tv - w\|_{L^2}^2 &= \sum_{m\in\mathbb{Z}^d} |\, \langle Tv - w, \varphi_m\rangle_{L^2} \,|^2 \\ &= \sum_{m\in\mathbb{Z}^d} |\lambda_m \, \langle \varphi_{-m}, v\rangle_{L^2} - \langle \varphi_{-m}, w\rangle_{L^2} \,|^2.\end{aligned}$$

To see why the last equality is true, note that $\langle Tv, \varphi_m\rangle = \lambda_m \, \langle \varphi_{-m}, v\rangle$ and $\langle w, \varphi_m\rangle_{L^2} = \overline{\langle \varphi_m, w\rangle}_{L^2} = \langle \varphi_{-m}, w\rangle_{L^2}$ as w is real-valued. Thus, given $\{(v_i, w_i)\}_{i=1}^n$, the least-squares estimator over the class $\mathcal{T}$ is an operator T specified by the sequence $\{\lambda_m\}_{m\in\mathbb{Z}^d}$, which is obtained by solving the optimization problem

$$\min_{\{\lambda_m \,:\, m\in\mathbb{Z}^d\}} \; \frac{1}{n}\sum_{i=1}^n \sum_{m\in\mathbb{Z}^d} \Big|\lambda_m \, \langle \varphi_{-m}, v_i\rangle_{L^2} - \langle \varphi_{-m}, w_i\rangle_{L^2}\Big|^2 \quad \text{subject to } \sup_{m\in\mathbb{Z}^d} |\lambda_m| \le C.$$

However, this estimator cannot be implemented for two reasons. First, there is an infinite sum over $\mathbb{Z}^d$. Second the learner only has access to (v_i, w_i) through $v_i^N := \{v_i(x) \,:\, x \in \mathrm{G}\}$ and $w_i^N := \{w_i(x) \,:\, x \in \mathrm{G}\}$, and thus the L^2 inner products cannot be computed exactly. Both of these issues can be resolved by considering the operator specified by the finite length sequence $\widehat{\lambda}(N) = \{\hat{\lambda}_m : m \in \mathbb{Z}^d_{\le K}\}$ obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^n \sum_{m\in\mathbb{Z}^d_{\le K}} \Big|\lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m)\Big|^2$$

subject to $\sup_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m| \leq C$. DFT, which stands for Discrete Fourier Transform, is the numerical approximation of $\langle \varphi_{-m}, u\rangle_{L^2}$ and is defined formally as

$$\mathrm{DFT}(u)(-m) := \frac{1}{N^d} \sum_{x\in \mathrm{G}} u(x)\, e^{-2\pi\,\mathrm{i}\langle x,m\rangle}.$$

To indicate the dependence of both truncation value K and grid-size N^d, let us denote the estimator obtained by solving this problem to be $\widehat{T}_K^N$ where

$$\widehat{T}_K^N := \sum_{m\in\mathbb{Z}^d_{\leq K}} \widehat{\lambda}_m(N)\ \varphi_m \otimes \varphi_{-m}. \tag{6.2}$$

The estimator $\widehat{T}_K^N$ is the closest implementable version of the least-squares estimator for our setting.

6.4.3 Error Bounds

In this section, we study how $\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu)$ decay as a function of n, K and N. Note that we have only specified that $\mathcal{V}$ and $\mathcal{W}$ are subsets of $L^2(\mathbb{T}^d, \mathbb{R})$, but have not specified their precise form. A natural choice would be $\mathcal{V} = \mathcal{W} = \{u \in L^2(\mathbb{T}^d, \mathbb{R}) : \|u\|_{L^2} \leq 1\}$, the unit ball of $L^2(\mathbb{T}^d, \mathbb{R})$. However, it turns out that $\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu)$ does not vanish under such $\mathcal{V}$ and $\mathcal{W}$.

To see this, let $K \in \mathbb{N}$ be a truncation parameter chosen by the learner. Define $\mu = \mathrm{Uniform}(\{(\psi_m, \psi_m) : 2^K < |m|_\infty < 2^{K+1}\})$ that is only supported on large modes. Here, $\psi_m = 2^{-1/2}(\varphi_m + \varphi_{-m})$ is the symmetrized, real-valued version of m-th Fourier mode. Note that we can choose a distribution as a function of K because the truncation parameter K can depend on the sample size n, but not on the exact realization of the samples.

For any sample size n and the estimator $\widehat{T}_K^N$ produced by the learner, $\widehat{T}_K^N v = 0$ almost surely for $(v, w) \sim \mu$. Thus, we have $\mathbb{E}_{(v,w)\sim\mu}\left[\|\widehat{T}_K^N v - w\|_{L^2}^2\right] = \mathbb{E}_{(v,w)\sim\mu}\left[\|w\|_{L^2}^2\right] = 1$, as $w = \psi_m$ for some $2^K < |m|_\infty < 2^{K+1}$ almost surely and $\|\psi_m\|_{L^2} = 1$ for any $m \in \mathbb{Z}^d_{>0}$.

Next, let $C = 1$ and define $T^\star = \sum_{m\in\mathbb{Z}^d} \varphi_m \otimes \varphi_{-m}$. It is easy to see that $T^\star \psi_k = 2^{-\frac{1}{2}}(T^\star\,\varphi_k + T^\star\,\varphi_{-k}) = 2^{-\frac{1}{2}}\,(\varphi_{-k} + \varphi_k) = \psi_k \quad \forall k \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$. As $T^\star \in \mathcal{T}$, we obtain $\inf_{T\in\mathcal{T}} \mathbb{E}_{(v,w)\sim\mu}\left[\|Tv - w\|_{L^2}^2\right] \leq \mathbb{E}_{(v,w)\sim\mu}\left[\|T^\star v - w\|_{L^2}^2\right] = 0$. Thus, we have established

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) \geq 1.$$

This shows that merely bounding the L^2 norm of v, w is not sufficient to achieve a vanishing error. So, we need a stronger assumption on the input and output functions.

The inductive bias in FNOs is that the functions are sufficiently smooth so that the higher Fourier modes can be safely ignored. We will also adopt this viewpoint and assume that $\mathcal{V}$ and $\mathcal{W}$ are smooth subsets of $L^2(\mathbb{T}^d, \mathbb{R})$. In particular, we will assume that $\mathcal{V} = \mathcal{W} = \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, a $(s, 2)$-Sobolev space (see Section 6.3.3). For any $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, we are guaranteed that $\langle \varphi_{-m}, u \rangle_{L^2} \to 0$ sufficiently fast as $|m|_\infty \to \infty$. This allows us to ignore higher Fourier modes while only incurring small error. The following Theorem, whose proof is deferred to Apendix E.4, makes these arguments precise and provides an upper bound on the excess risk of $\widehat{T}_K^N$ in terms of n, N, and K.

Theorem 25 (Upper Bound)**.** *Let $\mathcal{V} = \mathcal{W} = \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$ for $s > d/2$ and μ be any distribution on $\mathcal{V} \times \mathcal{W}$ for which $\exists B > 0$ such that $\|v\|_{\mathcal{H}^s} \leq B$ and $\|w\|_{\mathcal{H}^s} \leq B$ almost surely. Then, for n iid samples $\{(v_i, w_i)\}_{i=1}^n \sim \mu^n$ accessible to the learner over the N-uniform grid of $[0, 1]^d$, the estimator $\widehat{T}_K^N$ defined in (6.2) for $N > \max\{5, 2K\}$ satisfies*

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) \leq 8B^2(C+1)^2 \left(\frac{1}{\sqrt{n}} + \frac{2^s \sqrt{\pi^d}}{N^s} + \frac{1}{K^{2s}} \right).$$

The terms $O(1/\sqrt{n})$, $O(1/N^s)$, and $O(1/K^{2s})$ are the estimator's statistical, discretization, and truncation errors respectively. For most practical applications of interest, we have $d = 3$ (functions defined on spatial coordinates). Since $\sqrt{\pi^d} \leq 6$ in these cases, the exponential dependence of the discretization error on d is not an issue. Finally, choosing $N \geq n^{\frac{1}{2s}}$ and $K \geq n^{\frac{1}{4s}}$, Theorem 25 guarantees the $\sqrt{n}$– risk consistency of the estimator $\widehat{T}_K^N$.

Proof Technique for Upper Bound: Here, we highlight here the key technical novelties of our proof techniques and the implications of our results. To establish the upper bound, we first decompose the excess risk into three components: (1) the risk gap between the optimal operator in $\mathcal{T}$ for the distribution μ and its truncated counterpart, (2) the uniform deviation between the true empirical risk on the sample and its numerical approximation on the discrete grid, and (3) the uniform deviation between the empirical risk and the actual risk. This decomposition, introduced at the beginning of Appendix E.4, is not limited to the linear setting and can also be applied to analyze general non-linear operator classes. Given such decomposition, bounding the truncation error is straightforward using standard Fourier series properties for Sobolev spaces. The discretization error, however, requires nontrivial analysis to show that controlling the error of DFT suffices. Importantly, while the lower bound on the DFT error likely bounds the discretization error below, an upper bound on the DFT error does not always translate to an upper bound for the trained operator. For example, this is not true if one adds non-smooth activation such as RELU to our model. For statistical error, standard techniques yield a bound of $\sqrt{\frac{K^d}{n}}$, which does not allow taking

$K \to \infty$. Our key contribution is a refined analysis that achieves a $\frac{1}{\sqrt{n}}$ bound independent of K^d. The K-independent bound is especially notable because K in FNOs is analogous to the width in standard neural networks, where generalization bounds are known to be width-independent [Golowich et al., 2018]. Our results provide initial evidence that similar K-free generalization bounds may be achievable for FNOs.

Our next result, proved in Appendix E.5, provides a lower bound on the rates at which $\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu)$ decay.

Theorem 26 (Lower Bound)**.** *Let $\mathcal{V} = \mathcal{W} = \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$ for $s > d/2$ and $C = 1$. Given $n, N, K \in \mathbb{N}$, there exists a distribution on μ on $\mathcal{V} \times \mathcal{W}$ for which $\exists B > 0$ such that $\|v\|_{\mathcal{H}^s} \leq B$ and $\|w\|_{\mathcal{H}^s} \leq B$ almost surely and for n iid samples $\{(v_i, w_i)\}_{i=1}^n \sim \mu^n$ accessible over the N-uniform grid of $[0,1]^d$, the estimator $\widehat{T}_K^N$ defined in (6.2) for $N^s \geq \sqrt{2}B$ satisfies*

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) \geq \frac{B^2}{3(s+1)} \left(\frac{1}{8n} + \frac{1}{N^{2s}} + \frac{2}{(K+2)^{2s}} \right).$$

Although the lower bound on truncation error matches with the upper bound, there is a gap in the statistical and discretization error. We leave closing this gap for future work.

6.4.4 On Possible Extensions and Refinements of our Error Bounds

The smoothness assumptions in our work are primarily needed to control truncation and discretization errors. The lower bound in Section 6.4.3 shows that some regularity, specifically $s > 0$, is necessary for achieving a vanishing truncation error. This condition is also sufficient for our upper bound on the truncation error. The stronger requirement $s > d/2$ is required to ensure that the DFT-based estimator approximates the true Fourier coefficients. Moreover, even when $s = 0$, a statistical rate of $1/\sqrt{n}$ independent of K can still be obtained under alternative assumptions. For example, if the operator's spectrum lies in $\ell^2(\mathbb{Z}^d)$, making it Hilbert-Schmidt, results from Tabaghi et al. [2019], Raman et al. [2024b] imply that such a rate is possible without any smoothness assumptions.

Additionally, in our analysis of the discretization error (Appendix E.4.2), the key quantity we control is the difference between the DFT approximation and the true Fourier coefficient, namely

$$|\operatorname{DFT}(u^N)(-m) - \langle \varphi_{-m}, u \rangle|.$$

The assumption of a uniform grid is used only to bound the numerical integration error introduced by the DFT. In principle, any numerical integration method can be applied to a

non-uniform grid, and as long as its error vanishes with increasing grid resolution. For non-uniform grids with structure, such as those based on the roots of orthogonal polynomials, Gaussian quadrature rules may be used with standard accuracy guarantees. On unstructured grids, Monte Carlo methods with estimated importance weights can be used, although their convergence can be slow or the error may not vanish if the estimated weights have high variance.

6.5 Experiments

In this section, we present experiments demonstrating that our estimator achieves vanishing errors. We pick $d = 2$, and the input functions v are sampled i.i.d. from $\mathcal{N}(0, 10^2(-\nabla^2 + \mathbf{I})^{-\gamma})$, a widely used distribution for generating training data in the operator learning literature (see Li et al. [2021], Kovachki et al. [2023]). Since γ governs the decay rate of the eigenvalues of the covariance operator for this distribution, it directly controls the average smoothness of the samples v. For our experiments, we set $\gamma = 2$ as this is the smallest integer value that ensures $\gamma > d/2$ for $d = 2$.

To generate training data, we define a random operator

$$T^\star := \sum_{m \in \mathbb{Z}^d} \lambda_m \, \varphi_m \otimes \varphi_{-m},$$

where φ_m's are Fourier modes and $\lambda_m \sim \text{Unif}(-2, 2)$. For a given input v, the corresponding output is generated as $w = T^\star v + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, (-\nabla^2 + \mathbf{I})^{-3})$. Noise is sampled from a higher-order smooth space to ensure that its addition does not alter the smoothness of w. In actual implementation, the transformation $T^\star v$ is implemented on some $N \times N$ grid using Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT). The sum over $\mathbb{Z}^d$ is truncated at a Nyquist limit of $N/2$.

Recall that, our estimator in Section 6.4.2 is obtained by solving a convex optimization problem for λ_m's for $m \in \mathbb{Z}^d_{\leq K}$. So, we implement the optimization routine for our estimator using stochastic gradient descent with a projection step to ensure $|\widehat{\lambda}_m| \leq 2$.

Figures 6.1, 6.2, and 6.3 show the statistical, truncation, and discretization errors, respectively. The y-axis in all these figures represents the relative mean-squared testing error:

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{\|w_i^{\text{true}} - w_i^{\text{predicted}}\|_{L^2}^2}{\|w_i^{\text{true}}\|_{L^2}},$$

evaluated using $n_{\text{test}} = 100$ i.i.d. samples. Additional experimental results are presented in Appendix E.6. The corresponding code is available at

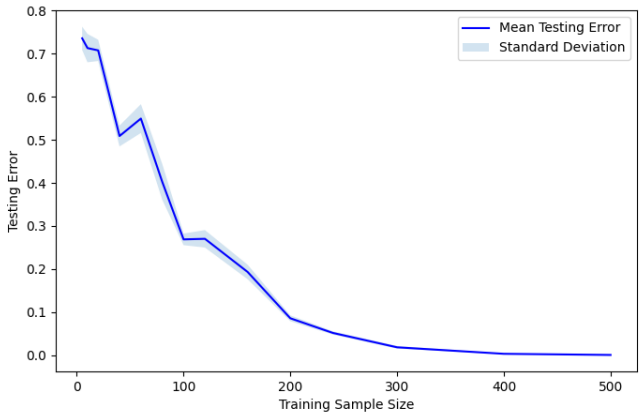


Figure 6.1: Statistical error of the estimator.

https://github.com/unique-subedi/fourier-linear-operators.

6.5.1 Statistical Error

Both training and testing are carried out on a 128×128 grid, with the estimator implemented using $K = 64$ modes. Error bands are included to account for fluctuations in the estimated parameters at small sample sizes, showing results from 5 independent runs. The model is trained and tested at the same resolution at the Nyquist limit of $K = 32$ modes to ensure that the reported error isolates statistical error with the minimum possible truncation and discretization errors. The smallest error is $\sim 6 \times 10^{-4}$ for the sample size of 500.

6.5.2 Truncation Error

Training and testing data are generated on a 128×128 grid, with the estimator trained using $n = 500$ samples. Error bands are omitted as the estimates are almost identical due to a large sample size. Both training and testing are conducted at the same resolution to remove discretization error, with the sample size selected to minimize statistical error, ensuring that the reported error isolates the truncation error effectively. The testing error converges to around 7.9×10^{-4} at the Nyquist limit of $K = 64$.

6.5.3 Discretization Error

Testing data is generated on a 512×512 grid. The estimator is trained using $n = 500$ samples on grids of varying sizes $N \times N$, where $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. For each training grid of size $N \times N$, truncation is performed at the Nyquist limit ($K = N/2$).

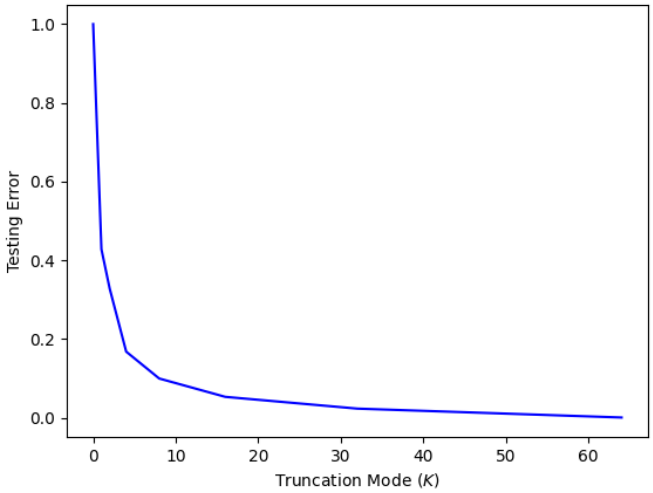


Figure 6.2: Truncation error of the estimator.

The trained estimators are subsequently evaluated at the higher testing resolution of 512×512 to quantify discretization error. The testing error converges to around 6×10^{-4} when the estimator is trained at a full grid size of 512×512 with 500 training samples.

6.5.4 Summary of Experimental Findings

Our experiments in this section confirm that all three sources of error—statistical, truncation, and discretization—can be independently reduced by increasing the respective parameters n, K, and N.

However, the observed convergence rates, as reported in Appendix E.6, reveal some gaps compared to our theoretical predictions. In particular, the statistical error appears to decay faster than expected, and may depend on the smoothness parameter γ. This suggests that a refined, distribution-dependent analysis could yield sharper bounds beyond the worst-case setting. The largest discrepancy arises in the discretization error, where we observed almost uniform rate for all smoothness parameters. While our theory assumes the test resolution $N_2 \to \infty$, the experiments use a fixed resolution $N_2 = 512$ for computational reasons. This mismatch may account for the unexpectedly uniform decay rate across different smoothness levels. A more detailed analysis under finite resolutions could help explain this gap.

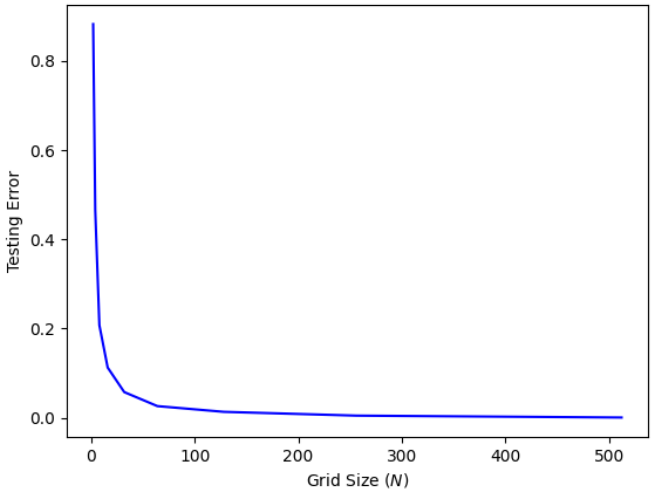


Figure 6.3: Discretization error of the estimator.

6.6 Discussion

In this chapter, we established the excess risk error bounds of learning the core linear layer $v \mapsto \mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big)$ of Fourier neural operators. A natural future direction is to extend these results to single layer Fourier neural operator, $v \mapsto \sigma\left(\mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big) + b\right)$ and then to multiple layers. Although simple metric entropy-based analysis gives a bound on statistical error even for single layer neural operator, such a bound is vacuous when $K \to \infty$. It would be interesting to see if we can get a meaningful statistical rate even at the limit of $K \to \infty$. One can view K as an analog of the width of traditional neural networks. Thus, analysis of $v \mapsto \sigma\left(\mathcal{F}^{-1}\Big(\Lambda_\beta\ \mathcal{F}(v)\Big) + b\right)$ as $K \to \infty$ can lead to a neural tangent kernel theory [Jacot et al., 2018] for operator learning. These insights will help us better understand width vs depth tradeoffs in operator learning. One approach to extending our statistical rates to nonlinear operator could be to carry out a Rademacher analysis similar to that of Golowich et al. [2018] for finite-dimensional neural networks. The main technical challenge arises from the nature of the pointwise nonlinearity σ. In finite dimensions, for $v \in \mathbb{R}^p$, Golowich et al. [2018] exploit the identity $\sigma(v) = \sum_{j=1}^{p} \sigma(\langle v, e_j \rangle) e_j$, where $\{e_j\}$ is the standard basis. However, this identity no longer holds in the infinite-dimensional setting considered in Kovachki et al. [2023], where σ is applied pointwise in the spatial domain rather than in the spectral domain.

For discretization error, we consider the setup where the training data is available on a grid of size N^d but the trained operator is evaluated at full resolution ($N \to \infty$). It would be interesting to study the discretization error when the training data is available at resolution

N_1, but the trained operator is evaluated at resolution N_2. Such a theory would provide a more fine-grained quantification of multi-resolution generalization error observed in practices [Li et al., 2021]. Additionally, a key practical limitation of our analysis is its reliance on uniform grids. As discussed in Section 6.4.4, the uniform grid assumption is used solely to bound the numerical integration error from the DFT. In principle, any integration scheme on a non-uniform grid could replace the DFT, as long as the approximation error vanishes with grid refinement. For structured grids (e.g., nodes from orthogonal polynomials), Gaussian quadrature can be used directly. For unstructured grids, one could apply Monte Carlo methods with importance weights, though these may exhibit slow convergence or non-vanishing bias when the weight variance is high. Extending our analysis to such general grids and formalize how discretization error interacts with other sources of error remains an interesting future direction.

CHAPTER 7

On the Benefits of Active Data Collection for Operator Learning

In this chapter[1], we study the data complexity of operator learning. Specifically, given an operator F, how many input-output pairs $\{(f_j, \mathrm{F}(f_j))\}_{j=1}^n$ are necessary to estimate an operator $\widehat{\mathrm{F}}_n$ such that $\widehat{\mathrm{F}}_n(f) \approx \mathrm{F}(f)$ for all relevant f? This question has been studied in several specific contexts, such as for linear operators with fixed singular value decomposition by de Hoop et al. [2023] and Subedi and Tewari [2025b], for Lipschitz operators by Liu et al. [2024], and for the random feature model by Nelsen and Stuart [2021]. These are just a few representative works, and we refer the reader to [Kovachki et al., 2024b, Section 5] for a more comprehensive review of such sample complexity results.

A common theme of these data complexity analyses is that they are conducted within the framework of traditional statistical learning [Kovachki et al., 2023, Section 2.2]. In the statistical setting, the learner has access to training samples $\{(f_j, \mathrm{F}(f_j))\}_{j=1}^n$, where $f_j \sim_{\text{iid}} \mu$ from some probability measure μ, and the objective is to produce an estimator $\widehat{\mathrm{F}}_n$ such that $\widehat{\mathrm{F}}_n(f) \approx \mathrm{F}(f)$ on average over test samples $f \sim \mu$. This scenario is also referred to as the passive learning setting. Under reasonable non-trivial assumptions, the best achievable rate of error convergence in this setting is $\sim 1/n$, when $\widehat{\mathrm{F}}_n$ is evaluated under an appropriate metric, say the p-th power of the L^p_μ-Bochner norm.

However, as already alluded to in Chapter 1, the iid-based statistical model is likely not right framework to study operator learning for PDEs. This is because the learner can generate any training data by querying the numerical solver, and thus has no reason to be limited to iid samples from some source distribution. In fact, as generating training data requires computationally expensive numerical solvers, the learner *should* ideally generate data adaptively to ensure that the computational cost of training is justified by saving during evaluation. The model where the learner can adaptively select the data is referred to as active

[1]This chapter is based on: Unique Subedi and Ambuj Tewari (2025). *On the Benefits of Active Data Collection in Operator Learning*. International Conference on Machine Learning (ICML).

learning model in statistical learning theory. We will propose an active learning model and argue why this is the right model to study operator learning for surrogate modeling of PDE.

Thus, in this chapter, we study the data complexity of operator learning where the learner is not restricted to iid samples from a source distribution and can use active data collection strategies. We focus on the case where the operator of interest F is a bounded *linear* operator. Although a distribution is not required to specify the training data in this setup, we still specify a distribution to evaluate the proposed estimator. We will assume that the estimator is evaluated with respect to input samples drawn from μ, a distribution with zero mean and the covariance structure defined by a continuous kernel K. Such distributions include Gaussian processes with common covariance kernels. For a given covariance kernel K, our main result provides an active data collection strategy, an estimation rule, and establishes an error bound for the proposed estimator in terms of the eigenvalue decay of the integral operator of K. Formally, if $\lambda_1 \geq \lambda_2 \geq \dots$ are the eigenvalues of the integral operator of K, there exists an active data collection strategy and estimation rule such that the estimator obtained using n actively collected input-output pairs achieves the following error bound:

$$\varepsilon^2 \sum_{i=1}^{n} \lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{i=n+1}^{\infty} \lambda_i.$$

The ε^2 captures the error of approximation oracle $\mathcal{O}$ for F that the learner has access to. For example, $\mathcal{O}$ could be the PDE solver used to generate training data for operator learning. Generally, $\varepsilon > 0$ is the irreducible error of the bound. The second term $O(\sum_{i>n} \lambda_i)$ is a reducible error, which goes to 0 as $n \to \infty$ for continuous kernels K under the bounded domain. For example, for the covariance operator $\alpha(-\nabla^2 + \beta \mathbf{I})^{-\gamma}$ used by Li et al. [2021] and Kovachki et al. [2023], we show that the reducible error vanish at the rate $\lesssim n^{-\left(\frac{2\gamma}{d}-1\right)}$. Taking $2\gamma \gg d$, one can achieve *any* polynomial rate of decay. In fact, given any rate $R_n \to 0$ as $n \to \infty$, one can always construct a continuous kernel K such that the reducible error decays faster than R_n. Thus, arbitrarily fast rates can be obtained using active data collection strategies. Our main result is formalized in Theorem 27, and the proof is based on the celebrated Karhunen–Loève decomposition for functions drawn from μ with covariance kernel K.

Furthermore, in Theorem 28, we show that, irrespective of the decay rate of the eigenvalues of the covariance kernel K, there always exists a bounded linear operator F and a distribution μ with covariance kernel K such that the minimax estimation error fails to converge to 0 under *any* passive (i.i.d.) data collection strategy. In particular, for every n, even when

$\varepsilon = 0$, we establish the minimax lower bound of

$$\|\mathrm{F}\|_{\mathrm{op}}^2 \ \lambda_1$$

under any passive data collection strategy. That is, the lower bound does not vanish even as $n \to \infty$. Collectively, Theorems 27 and 28 establish a clear advantage of active data collection strategy for operator learning.

7.1 Related Works

Recent work by Musekamp et al. [2024] considers active methods for operator learning. However, in contrast to our approach of using linearity of the solution operator and the distributional family of interest, their methods rely on estimating uncertainty and identifying coreset. Additionally, their study is purely empirical and lacks theoretical guarantees. In a similar spirit, Li et al. [2024a] study using active learning to select input functions from multiresolution datasets to lower the data cost.

On the theoretical side, a closely related work is by Kovachki et al. [2024a], who allow for active data collection strategies. However, the upper bound in [Kovachki et al., 2024a, Theorem 3.3] is derived assuming that input functions $v_1, \ldots, v_n$ are drawn i.i.d. from μ. Their proof, based on standard empirical risk minimization (ERM) analysis, achieves a convergence rate that, at best, matches the Monte Carlo rate of $n^{-1/2}$. Moreover, their lower bounds apply to both active and passive data collection strategies, suggesting that, for the nonparametric operator classes considered by Kovachki et al. [2024a], active learning provides no clear advantage over passive approaches. Exploring whether an adaptive data collection strategy, informed by the covariance of μ and targeting smaller subclasses within these broad nonparametric classes, could yield faster convergence rates remains an interesting direction for future research.

Additionally, Boullé et al. [2023] shares our objective of achieving faster convergence rates for PDEs with linear solution operators, but there are notable differences between their results and ours. First, their approach requires stronger control over the Hilbert-Schmidt norm of the operator F, whereas we only require control over the operator norm. Notably, the Hilbert-Schmidt norm can be arbitrarily larger than the operator norm. Second, their estimator uses the specific structure of F, particularly the Green's function, while we rely on black-box access to F via an ε-approximate oracle. Lastly, although both approaches introduces a term measuring the quality of training data (ε in our bound and Γ_ε in theirs), their definition of Γ_ε is more technical and less intuitive. However, their guarantee is stronger, as their upper bound applies uniformly to any L^2-integrable input function, whereas our

guarantees hold in expectation for inputs drawn from the distribution μ.

Our work considers the setting where μ is defined by a stochastic process with a specific covariance structure. Such a μ was taken to be a Gaussian process with mean zero and covariance given by $\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$ in [Bhattacharya et al., 2021, Li et al., 2021, Kovachki et al., 2023]. The use of Karhunen–Loève decomposition for generating input functions is also discussed by Boullé and Townsend [2024, Section 4.1]. Our upper bound also share conceptual similarities with results in Lanthaler et al. [2022], Lanthaler [2023], who established approximation error bounds, rather than estimation, in terms of s eigenvalues of covariance operator.

Finally, we highlight the ICML 2024 tutorial by Azizzadenesheli [2024], who mentions active data collection as an important future direction for operator learning. We also acknowledge the extensive literature on the learning-theoretic foundations of active learning [Settles, 2009]. The active learning framework we adopt is known as the membership query model, which has a rich history in learning theory Angluin [1988]. A more detailed discussion of various active learning models within the learning theory literature is deferred to Section 7.3.4.

7.2 Preliminaries

7.2.1 Notation

Let $\mathbb{R}, \mathbb{C}$ denote the set of real and complex numbers respectively. The set $\mathbb{N}$ and $\mathbb{Z}$ denote the natural numbers and integers. Define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For any $x \in \mathbb{R}^d$, we use $|x|_p$ to denote the ℓ^p norm of x. Given a set $\mathcal{X} \subseteq \mathbb{R}^d$, we use $L^2(\mathcal{X})$ to denote the space of squared integrable *real-valued* functions on $\mathcal{X}$ under some base measure ν. For any $u \in L^2(\mathcal{X})$, we define $\|u\|_{L^2}^2 := \int_{\mathcal{X}} |u(x)|^2 \, d\nu(x)$. The notation ν is reserved for the base measure on $\mathcal{X}$, whereas μ will be used to denote the probability distribution over $L^2(\mathcal{X})$. For a linear operator $\mathrm{F} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$, we define $\|\mathrm{F}\|_{\mathrm{op}} := \sup\{\|\mathrm{F}\, v\|_{L^2} : \|v\|_{L^2} = 1\}$. We use GP to denote Gaussian Process.

7.2.2 Distribution Over Function Space

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be any compact set, $\mathcal{B}(\mathcal{X})$ denote the Borel sigma-algebra, and ν denote some finite measure on $\mathcal{X}$ (that is, $\nu(\mathcal{X}) < \infty$). Generally, we will take ν to be Lebesgue measure on $\mathcal{X}$ but sometimes it may be useful to take a weighted measure such as $\propto e^{-\alpha^2|x|^2} \, dx$. Denote $L^2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \nu)$ to be the set of all squared integrable functions on $\mathcal{X}$. From here on, we will drop the dependence on $\mathcal{B}(\mathcal{X})$ and ν, and just write $L^2(\mathcal{X})$. Let $(\Omega, \Sigma, \mathbf{P})$ denote a

probability space. We will consider a sequence of real-valued random variables $\{h_x : x \in \mathcal{X}\}$ defined over the probability space $(\Omega, \Sigma, \mathbf{P})$ that is centered, squared integrable, and has *continous* covariance kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Recall that covariance kernels are symmetric and positive definite. More precisely, for any $x, y \in \mathcal{X}$, the random variables h_x satisfies

$$\begin{aligned}\mathbb{E}[h_x] &= \int_\Omega h_x(\omega)\, d\mathbf{P}(\omega) = 0 \\ \mathbb{E}[h_x^2] &= \int_\Omega |h_x(\omega)|^2\, d\mathbf{P}(\omega) < \infty \\ \mathbb{E}[h_x\, h_y] &= \int_\Omega h_x(\omega)\, h_y(\omega)\, d\mathbf{P}(\omega) = K(y, x).\end{aligned}$$

Next, we use this process to define a probability distribution over $L^2(\mathcal{X})$. To that end, it will be more convenient to write the process as a function $h : \mathcal{X} \times \Omega \to \mathbb{R}$. By definition, $h(x, \cdot)$ is Σ-measurable for every $x \in \mathcal{X}$. However, this is not enough to argue that h is a random element of $L^2(\mathcal{X})$. Thus, to ensure measurability, we will only consider stochastic processes h that satisfy the following: (i) The process h is measurable with respect to product sigma algebra $\mathcal{B}(\mathcal{X}) \times \Sigma$ and (ii) For every $\omega \in \Omega$, the sample path $h(\cdot, \omega) : \mathcal{X} \to \mathbb{R}$ is an element of $L^2(\mathcal{X})$. Conditions (i) and (ii) ensure that $\omega \mapsto h(\cdot, \omega)$ is a measurable function from Ω to $L^2(\mathcal{X})$ [Hsing and Eubank, 2015, Theorem 7.4.1]. In other words, h is a $L^2(\mathcal{X})$ valued random variable. We can now meaningfully talk about probability distribution over $L^2(\mathcal{X})$ induced by the stochastic process h.

Accordingly, given a continuous covariance kernel K, let $\mathcal{P}(K)$ denote the set of all centered and squared-integrable stochastic processes with covariance kernel K indexed by $\mathcal{X}$ that satisfies conditions (i) and (ii) above. With a slight abuse of notation, we will also use $\mathcal{P}(K)$ to denote the set of all distributions over $L^2(\mathcal{X})$ induced by these stochastic processes. Each element $\mu \in \mathcal{P}(K)$ is now a probability distribution over $L^2(\mathcal{X})$.

7.2.3 Problem Setting and Goal

Let $\mathrm{F} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ denote the operator of interest. One should think of F as the solution operator of the PDE. The goal is to estimate a surrogate $\widehat{\mathrm{F}}_n$ using n input/output functions such that

$$\sup_{\mu \in \mathcal{P}(K)} \mathop{\mathbb{E}}_{v \sim \mu} \left[\left\| \widehat{\mathrm{F}}_n(v) - \mathrm{F}(v) \right\|_{L^2}^2 \right] \tag{7.1}$$

is small. In the absence of additional knowledge about the image space of the solution operator F, minimizing this objective is the most natural choice. Accordingly, for a fixed μ, the L^p_μ-Bochner norm has been a standard error metric in the operator learning literature (see [Kovachki et al., 2023, Section 2.2], [Liu et al., 2024]). The $p = 2$ case, in particular, is

of practical significance, as its empirical counterpart is the widely used mean squared loss. Regarding the family of probability distributions, our proposed family $\mathcal{P}(K)$ aims to unify and generalize marginal distributions on input functions commonly used in practice Li et al. [2021], Lu et al. [2021]. This family also aligns with the recommendation of Boullé and Townsend [2024]. Other families of probability distributions, such as the set of all compactly supported measures on a Hilbert space, have been used in theoretical analyses (e.g., Liu et al. [2024]). Extending our result to include other distribution families of theoretical or applied interest is left for future work.

Throughout this work, we will assume that the learner knows the covariance kernel K.

Assumption 3. *The learner knows the kernel K.*

Although not always explicitly stated, this has been a standard assumption in the operator learning literature. For example, Li et al. [2021] and Kovachki et al. [2023] generate their input functions, both during training and testing, from a Gaussian process with the covariance kernel K such that its associated integral operator is $\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$ for some constants $\alpha, \beta, \gamma > 0$. Thus, all the empirical performances observed in these works are in a setup similar to those described above. Additionally, Boullé and Townsend [2024, Section 4.1.1] also suggests generating source terms (input functions) from Gaussian processes with standard covariance kernels such as RBF, Mattern, etc.

Additionally, from a learning-theoretic perspective, assuming knowledge of the kernel K is arguably without loss of generality. In active learning, it is common to assume access to an unlimited pool of unlabeled samples $v_1, \ldots, v_m \sim_{\text{iid}} \mu$, where $\mu \in \mathcal{P}(K)$, and focus on minimizing label complexity—the number of labeled samples requested [Hanneke, 2013]. This aligns with our setting, where labeling (e.g., solving a PDE) is the primary cost. Given such unlabeled samples, one can estimate the covariance operator as

$$\Sigma_m = \frac{1}{m-1} \sum_{i=1}^{m} (v_i - \bar{v}_m) \otimes (v_i - \bar{v}_m)$$

where $\bar{v}_m = \frac{1}{m} \sum_{i=1}^{m} v_i$. Since $\mathbb{E}[\|v_i\|^2] < \infty$, Theorem 8.1.2 of [Hsing and Eubank, 2015] guarantees that $\Sigma_m \to \Sigma$ almost surely in Hilbert-Schmidt norm, where Σ is the integral operator associated with K. While our work assumes Σ has a finite trace norm, this is not required to recover its eigenfunctions: convergence in Hilbert-Schmidt norm suffices for accurate spectral approximation. Thus, assuming access to the eigenfunctions of K is reasonable in theory, even if it may be computationally demanding in practice.

Once the input functions are generated, the learner has to use numerical solvers to PDE numerically in order to generate the solution function. In this work, we will make the

following assumption about learner's access to the PDE solver.

Assumption 4. *The learner only has black-box access to* F *through an ε-approximate oracle $\mathcal{O}$ that satisfies*

$$\sup_{v \in L^2(\mathcal{X})} \|\mathcal{O}(v) - \mathrm{F}(v)\|_{L^2}^2 \leq \varepsilon^2.$$

From an implementation standpoint, it might seem unnatural to consider $\mathcal{O}(v)$ for a function $v \in L^2(\mathcal{X})$, especially since most PDE solvers usually only take function values over a discrete grid as an input. Nevertheless, the oracle is an abstract object, and the grid can be integrated into its definition. For example, given any function v, the oracle first extracts the values of v on a grid $\{x_1, \ldots, x_m\}$ and produces output values on the same or a different grid. On the output side, the oracle may then construct an actual function, either through trigonometric interpolation or simply by setting the function values to zero outside the grid points. Thus, we do not specify these implementation details of the oracle and instead characterize it solely by accuracy parameter ε.

In general, ε primarily reflects the discretization error for finite-difference type methods and truncation error for spectral methods, but it may also include measurement errors or errors resulting from the early stopping of some iterative routine. Therefore, ε can be broadly viewed as quantifying the quality of the training data. From this perspective, ε represents the irreducible error in (7.1). Specifically, there exists a function $g : [0, \infty) \to [0, \infty)$ such that (7.1) is bounded below by $g(\varepsilon)$, even as $n \to \infty$. There is extensive literature that attempts to quantify ε for various oracles (PDE solvers), and we can use these results readily to establish bounds on the irreducible error in our context. For example, for spectral solvers truncated to the first N basis functions where the input and output functions are s-times continuously differentiable, we typically have $\varepsilon \sim N^{-s/d}$. Here, d is the dimension of the domain Ω.

7.3 Upper Bounds Under Active Data Collection

In Section 7.2.3, we discussed the problem setting and the goal. Next, we specify how the learner can collect the training data $(v_1, w_1), \ldots, (v_n, w_n)$. In a departure from the standard statistical learning setting, where the training data is obtained as iid samples from the distribution under which the learner is evaluated, we investigate active data collection strategies. In active data collection strategies, the learner can pick *any* source terms $v_1, \ldots, v_n$ and use the oracle to obtain $w_i = \mathcal{O}(v_i)$. Since the goal is to provide guarantees under samples from the distribution $\mu \in \mathcal{P}(K)$, the learner *can* use the knowledge of K to pick source terms. For a given oracle with accuracy ε, covariance kernel K, and the desired accuracy $\eta > 0$,

the goal of the learner is to develop an active data collection strategy for the source terms and an estimation rule to produce $\widehat{\mathrm{F}}$ such that the accuracy of η can be obtained with the fewest number of oracle calls. Or equivalently, achieve an optimal decay in the upperbound of (7.1) for $n \in \mathbb{N}$ number of oracle calls. Under this model, we provide an upperbound on (7.1) when F is a bounded linear operator.

Theorem 27 (Upper Bound)**.** *Suppose* F *is a bounded linear operator. There exists a deterministic data collection strategy and a deterministic estimation rule such that the estimate* $\widehat{\mathrm{F}}_n$ *produced after* n *calls to oracle* $\mathcal{O}$ *satisfies*

$$\sup_{\mu \in \mathcal{P}(K)} \mathop{\mathbb{E}}_{v \sim \mu} \left[\left\| \widehat{\mathrm{F}}_n(v) - \mathrm{F}(v) \right\|_{L^2}^2 \right] \leq \varepsilon^2 \sum_{i=1}^{n} \lambda_i + \|\mathrm{F}\|_{op}^2 \sum_{i>n} \lambda_i.$$

Here, $\lambda_1 \geq \lambda_2 \geq \dots$ *are the eigenvalues of the integral operator of* K *defined as* $(\mathcal{I}_K v)(\cdot) = \int_{\mathcal{X}} K(\cdot, x)\, v(x)\, d\nu(x)$.

The first term above is the irreducible error, which depends on the quality of the training data. For the second term, Hsing and Eubank [2015, Theorem 4.6.7] implies that

$$\sum_{i=1}^{\infty} \lambda_i = \int_{\mathcal{X}} K(x,x)\, d\nu(x) \leq \sup_x |K(x,x)|\, \nu(\mathcal{X}) < \infty.$$

This is finite because ν is a finite measure on $\mathcal{X}$, and $K(x,x)$ is a continuous function on a compact domain, making it bounded. As a result, the second term in the upper bound of Theorem 27 vanishes as $n \to \infty$. In Section 7.3.3, we apply Theorem 27 to derive precise rates for several common covariance kernels.

7.3.1 Data Collection Strategy and The Estimator

Here, we specify the data collection strategy and the estimator that achieves the claimed guarantee in Theorem 27. Let $\{\lambda_j, \varphi_j\}_{j=1}^{\infty}$ be the sequence of eigenpairs of K defined by solving the Feldholm integral equation

$$\int_{\mathcal{X}} K(y,x)\, \varphi_j(x)\, d\nu(x) = \lambda_j\, \varphi_j(y), \qquad y, x \in \mathcal{X}.$$

Given the Oracle call budget of n, the input functions that the learner selects are $\varphi_1, \varphi_2, \dots, \varphi_n$ as source terms. For each $i \in [n]$, the learner makes an oracle call and obtains $w_i = \mathcal{O}(\varphi_i)$. Then, we consider the estimator

$$\widehat{\mathrm{F}}_n := \sum_{i=1}^{n} w_i \otimes \varphi_i.$$

More precisely, this estimation rule yields an operator $\widehat{\mathrm{F}}_n$ such that $\widehat{\mathrm{F}}_n v = \sum_{i=1}^{n} w_i \langle \varphi_i, v \rangle_{L^2}$ for any $v \in L^2(\mathcal{X})$. Appendix F.1.1 provides an overview of the process for deriving this estimator starting from a least-squares estimation rule. Furthermore, Appendix F.3 discusses methods for approximating the eigenfunctions φ_i when the Fredholm integral equation cannot be solved exactly.

7.3.2 Sketch of a Proof of Theorem 27

We now provide a high-level, non-rigorous sketch of a proof of Theorem 27, and defer a full proof to Appendix F.1.

To bound the risk of the estimator specified above, we first rewrite the risk using Karhunen–Loève Theorem. Pick any $v \sim \mu$. Since v is defined using a centered and squared-integrable stochastic process with continuous covariance kernel K, the celebrated Karhunen–Loève Theorem [Hsing and Eubank, 2015, Theorem 7.3.5] states that

$$v(\cdot) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(\cdot),$$

where ξ_j's are random variables defined as $\xi_j := \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{X}} v(x)\, \varphi_j(x)\, d\nu(x)$. It turns out that ξ_j's are uncorrelated random variables with mean 0 and variance 1. That is, $\mathbb{E}[\xi_j] = 0$ and $\mathbb{E}[\xi_i\, \xi_j] = \mathbb{1}[i = j]$.

This decomposition allows us to rewrite expectation over μ in terms of expectation over the randomness of the sequence $(\xi_j)_{j \geq 1}$, which is more tractable. For simplicity, assume that $\varepsilon = 0$. Then, using Karhunen–Loève expansion, we can show that

$$\mathbb{E}_{v \sim \mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}(v)\right\|_{L^2}^2\right] \leq \mathbb{E}_{\xi}\left[\left\|\mathrm{F}\left(\textstyle\sum_{j>n} \sqrt{\lambda_j}\, \xi_j\, \varphi_j\right)\right\|_{L^2}^2\right],$$

which can then be further upper bounded by $\|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j>n} \lambda_j$ using properties of ξ_i's.

There are two primary challenges in completing this argument in a fully rigorous manner. First, we must address the fact that the oracle $\mathcal{O}$ is only ε-approximate for any $\varepsilon > 0$. Second, the convergence statement for $\sum_{j=1}^{\infty} \sqrt{\lambda_j} \xi_j \varphi_j(\cdot)$ is quite specific, requiring careful attention when applying this result.

7.3.3 Examples of Covariance Kernels

To make the upperbound in Theorem 27 more concrete, let us consider a few specific covariance kernels K. While not all claims are rigorously proven in this subsection, a detailed and formal treatment of the material can be found in Appendix F.2.

7.3.3.1 Fractional Inverse of Shifted Laplacian

Li et al. [2021], Kovachki et al. [2023] generated input functions from $\mathrm{GP}(0, \alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma})$ for some constants $\alpha, \beta, \gamma > 0$. Here, ∇^2 is the Laplacian operator defined as

$$\nabla^2 v = \sum_{j=1}^{d} \frac{\partial^2 v}{\partial x_j^2}.$$

In this section, we will consider $\mathcal{X}$ to be a d-dimensional periodic torus $\mathbb{T}^d$ and the base measure ν is Lebesgue. We identify $\mathbb{T}^d$ by $[0,1]^d$ with periodic boundary conditions. Let us define a function $\varphi_m : \mathbb{T}^d \to \mathbb{C}$ as $\varphi_m(x) = e^{2\pi\,\mathrm{i}\, m\cdot x}$ for every $m \in \mathbb{Z}^d$. Recall that φ_m is the eigenfunction of ∇^2 with eigenvalue $-4\pi^2|m|_2^2$. In particular,

$$\nabla^2 e^{2\pi\,\mathrm{i}\, m\cdot x} = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}\, e^{2\pi\,\mathrm{i}\, m\cdot x} = -4\pi^2|m|_2^2\, e^{2\pi\,\mathrm{i}\, m\cdot x}.$$

Since $\{\varphi_m : m \in \mathbb{Z}^d\}$ forms a complete orthonormal system in $L^2(\mathbb{T}^d)$, there are no other eigenfunctions of ∇^2. A simple algebra shows that φ_m's are also the eigenfunctions of $(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$ with eigenvalues being $(\beta + 4\pi^2|m|_2^2)^{-\gamma}$.

Using this fact in Theorem 27 yields the upper bound

$$\leq \varepsilon^2 \left(\alpha\beta^{-\gamma} + \alpha + \frac{\alpha}{2\gamma - d}\right) + \frac{\alpha\, \|\mathrm{F}\|_{\mathrm{op}}^2}{2\gamma - d}\, \frac{1}{n^{\frac{2\gamma}{d}-1}}.$$

When $2\gamma/d - 1 > 0$, the reducible error above goes to 0 when $n \to \infty$. Again, as an example, Li et al. [2021] uses $\alpha = 7^{3/2}$, $\beta = 49$ and $\gamma = 2.5$ in their experiment for $2d$-Navier Stokes. In this case, $2\gamma/d = 2.5$, yielding the convergence rate of $n^{-1.5}$ for the reducible error. Note that this rate is faster than the usual passive statistical rate of $1/n$. However, for any value τ, one can take $\gamma = d(\tau+1)/2$ to get the rate of $n^{-\tau}$. Thus, every polynomial rate is possible depending on the choice of γ.

7.3.3.2 RBF Kernel

Let $\mathcal{X} = \mathbb{R}$ and $K(x,y) = \exp\left(-\frac{1}{2\ell^2}|x-y|^2\right)$ for all $x, y \in \mathbb{R}$ and $\ell > 0$. For now, let ν is a Gaussian measure with mean 0 and variance σ^2 on $\mathbb{R}$. Using the known results on eigenfunctions of RBF kernel in terms of Hermite polynomials [Williams and Rasmussen, 2006, Section 4.3.1], we show that there exists $\gamma \in (0,1)$ such that the upper bound in Theorem 27 is

$$\leq \frac{1}{(1-\gamma)} \left(\varepsilon^2 + \|\mathrm{F}\|_{\mathrm{op}}^2\, \gamma^n\right).$$

That is, the reducible error vanishes exponentially fast as $n \to \infty$. In Appendix F.2, we also show that a rate faster than any polynomial rate can be achieved for RBF kernel on $\mathbb{R}^d$.

7.3.3.3 Brownian Motion

Let us consider the case where $\mathcal{X} = [0, 1]$, the base measure ν is Lebegsue, and the stochastic process in Section 7.2.2 is Brownian motion. Recall that the Brownian motion is a Gaussian process with covariance kernel $K(s, t) = \min(s, t)$ for all $s, t \in [0, 1]$. It is well-known [Hsing and Eubank, 2015, Example 4.6.3] that the eigenfunctions of K can be written in terms of sine waves. A simple analysis can then be used to establish an upper bound of

$$\leq \frac{\varepsilon^2}{2} + \|\mathrm{F}\|_{\mathrm{op}}^2 \, \frac{1}{\pi^2} \, \frac{2}{2n-1}.$$

Therefore, the reducible error vanishes at rate $\sim n^{-1}$.

7.3.4 Comparison to Traditional Active Learning

The active learning framework we adopt in this work is referred to as the membership query model, which has a longstanding history in the learning theory literature Angluin [1988, 2001]. However, in traditional learning settings, the membership query model—where the learner can request labels for any unlabeled instance—is generally unrealistic. For example, in the context of human data, it may not be feasible to generate a label for an individual with an arbitrary feature vector, as such a person may not exist in reality. As a result, other active learning frameworks, such as the stream-based sampling model Atlas et al. [1989] and the pool-based model Lewis and Gale [1994], Hanneke [2013], have gained prominence in the recent literature. These models restrict the learner to requesting labels for instances sampled from a specific distribution, making them more practical for many real-world applications. For a comprehensive review of active learning models, their history, and key results, we refer readers to Settles [2009]. That said, we believe that the membership query model is the right model for developing surrogates for solution operators of PDEs. This is because a PDE solver can provide a solution to any query of an input function within an appropriate function space.

7.4 Lower Bounds on Passive Learning

In this section, we establish a lower bound on (7.1) for any passive data collection strategy. As usual, the kernel K is known to the learner. Nature selects a distribution $\mu_\star \in \mathcal{P}(K)$,

and the learner receives n i.i.d. samples $v_1, v_2, \ldots, v_n \sim \mu_\star$. For each $i \in [n]$, the learner queries the oracle $\mathcal{O}$ to produce $w_i = \mathcal{O}(v_i)$. It is important to emphasize that the learner can only make oracle calls for the i.i.d. samples $v_1, v_2, \ldots, v_n$. If the learner were allowed to make oracle calls for other input functions, the learner could simply disregard these i.i.d. samples and implement the "active strategy" from Section 7.3.1. Such restriction on oracle calls still includes most passive learning rules of interest, such as arbitrary empirical risk minimization (ERM), regularized least-squares estimators, and parametric operators trained with stochastic gradient descent.

Using these n training points $\{(v_i, w_i)\}_{i \leq n}$, the learner then constructs an operator $\widehat{\mathrm{F}}_n$. Since the learner only has access to samples from $\mu_\star$, it is unrealistic to expect a uniform guarantee over the entire family $\mathcal{P}(K)$ as established in Theorem 27. Therefore, in this section, the learner will be evaluated solely under the distribution $\mu_\star$. The objective is to minimize the expected loss under $\mu_\star$, defined as

$$\mathop{\mathbb{E}}_{v_{1:n} \sim \mu_\star^n} \left[\mathop{\mathbb{E}}_{v \sim \mu_\star} \left[\left\| \widehat{\mathrm{F}}_n(v) - \mathrm{F}(v) \right\|_{L^2}^2 \right] \right].$$

Moreover, establishing any meaningful lower bound on this risk requires imposing some restriction on the oracle $\mathcal{O}$. To understand why, consider the case where F is a finite-rank operator that only maps to the span of $\{\psi_1, \psi_2, \ldots, \psi_N\}$ for some orthonormal sequence $\psi_1, \ldots, \psi_N$ in $L^2(\mathcal{X})$. Now, consider an oracle $\mathcal{O}$ such that for any $v \in L^2(\mathcal{X})$, it outputs

$$\mathcal{O}(v) = \mathrm{F}(v) + \chi \, \psi_{N+1},$$

where $\chi \in \mathbb{R}$ and ψ_{N+1} is a unit norm function in $L^2(\mathcal{X})$ that is orthogonal to all ψ_j for $1 \leq j \leq N$. If $|\chi| \leq \varepsilon$, it is easy to see that

$$\sup_{v \in L^2(\mathcal{X})} \|\mathcal{O}(v) - \mathrm{F}(v)\|_{L^2}^2 = \|\chi \, \psi_{N+1}\|_{L^2}^2 = |\chi|^2 \leq \varepsilon^2.$$

Thus, $\mathcal{O}$ is a valid oracle according to Assumption 4. However, in principle, it is possible to encode the entire identity of F in a real number χ. Thus, the learner could determine the identity of F with just a single call to $\mathcal{O}$, making any attempt at establishing a lower bound futile.

This problem may still persist even when $\varepsilon = 0$. Consider the case where ν is the Lebesgue measure, and the oracle is of the form

$$\mathcal{O}(v) = \mathrm{F}(v) + \chi \mathbb{1}[x = x_0]$$

for some $x_0 \in \mathcal{X}$. Then, for any $v \in L^2(\mathcal{X})$, we have $\|\mathcal{O}(v) - \mathrm{F}(v)\|_{L^2}^2 = \|\chi \mathbb{1}\{x = x_0\}\|_{L^2}^2 = 0$ as $\nu(\{x_0\}) = 0$. This shows that the oracle can still reveal the identity of F in regions of the domain that have zero measure under ν. Therefore, to avoid these pathological edge cases, we will assume that the oracle is perfect.

Definition 29 (Perfect Oracle)**.** *$\mathcal{O}$ is a perfect oracle for* F *if, for every $v \in L^2(\mathcal{X})$, we have*

$$\Big(\mathcal{O}(v)\Big)(x) = \Big(\mathrm{F}(v)\Big)(x) \quad \forall x \in \mathcal{X}.$$

In other words, the perfect oracle $\mathcal{O}$ produces exactly the same function that F does—nothing more, nothing less. With this assumption, we are in the usual realizable setting often considered in statistical learning theory. That is, the learner has access to n samples $\{(v_i, \mathrm{F}(v_i))\}_{i=1}^n$, where $v_1, \ldots, v_n$ are drawn iid from some distribution μ.

Theorem 28 provides a lower bound on the risk of any estimator under such passive data collection strategy.

Theorem 28 (Lowerbound)**.** *Fix any continuous covariance kernel K with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$. Then, there exists a solution operator* F*, accessible to the learner through a perfect oracle $\mathcal{O}$, such that the following holds: for every $n \in \mathbb{N}$, there exists a distribution $\mu \in \mathcal{P}(K)$ such that, under any estimation rule within a passive data collection strategy, the risk of the resulting estimator $\widehat{\mathrm{F}}_n$ is*

$$\mathop{\mathbb{E}}_{v_{1:n} \sim \mu^n} \left[\mathop{\mathbb{E}}_{v \sim \mu} \left[\left\| \widehat{\mathrm{F}}_n(v) - \mathrm{F}(v) \right\|_{L^2}^2 \right] \right] \geq \frac{\|\mathrm{F}\|_{op}^2}{2} \sum_{j=1}^{m} \lambda_j$$

for every fixed $m \in \mathbb{N}$.

Specifically, for $m = 1$, we obtain a lower bound of $\frac{1}{2} \|\mathrm{F}\|_{\mathrm{op}}^2 \lambda_1$. This provides a non-vanishing lower bound for any non-trivial operator F and covariance kernel K.

Our lower bound is constructive: we explicitly define a difficult distribution for the learner. We construct a distribution μ over input functions such that, along each eigenfunction direction φ_j, it places mass 0 with probability $1-p$, and $\pm 1/\sqrt{p}$ with equal probability $p/2$. This yields a sparse distribution with rare but large spikes. A careful argument shows that this construction defines a valid distribution in $\mathcal{P}(K)$ for any $p > 0$. When p is small, the learner observes mostly zero inputs during training with probability at least $1/2$, yet the expected squared error along each direction is $(1/\sqrt{p})^2 \cdot p = 1$, leading to a non-vanishing error. The full proof is provided in Appendix F.4.

7.5 Experiments

In this section, we conduct numerical studies comparing our active data collection strategy with passive data collection (random sampling) for learning solution operators for the Poisson and Heat Equations. For the actively collected data, we implement the *linear estimator* defined in Section 7.3.1. On the other hand, for passively collected data, we use a least-squares estimator, where the pseudoinverse is computed numerically. Recall that, given input-output functions $\{v_i, w_i\}_{i=1}^n$, the least-squares estimator has a form $L = (\sum_{i=1}^n w_i \otimes v_i)(\sum_{i=1}^n v_i \otimes v_i)^\dagger$. For the actively collected data in Section 7.3.1, the v_i's are orthogonal, which yielded a simple and natural pseudoinverse (see Appendix F.1.1). However, for the passively collected data, the v_i's may not be orthogonal anymore and the pseudoinverse does not have a nice closed form. Thus, we use standard numerical techniques to compute the pseudo-inverse $(\sum_{i=1}^n v_i \otimes v_i)^\dagger$. However, in practice, one rarely uses linear estimators for passively collected data. Thus, we also compare our method against the Fourier Neural Operator [Li et al., 2021], the most popular architecture for operator learning. Our code is available at https://github.com/unique-subedi/active-operator-learning.

7.5.1 Poisson Equation

Let $\mathcal{X} = [0,1]^2$. Consider Poisson equation with Dirichlet boundary conditions:

$$-\nabla^2 u = f, \quad u(x) = 0 \quad \forall x \in \text{boundary}(\mathcal{X}),$$

where ∇^2 is the Laplace operator. The objective is to learn the solution operator that maps the source function f to the solution u. This solution operator is the inverse of the Laplacian, which is a compact linear operator since $\mathcal{X}$ is bounded. For the passive data collection strategy, the input functions f are independently sampled as $f \sim \text{GP}(0, 50^2(-\nabla^2 + \mathbf{I})^{-2})$, where GP denotes Gaussian Process.

The solution u is computed using the finite-difference method. Both linear estimators and Fourier Neural Operators (FNO) are trained on n such independently sampled pairs (f, u). For testing, 100 additional source functions $f \sim \text{GP}(0, 50^2(-\nabla^2 + \mathbf{I})^{-2})$ are generated, and their corresponding solutions u are also obtained via the finite-difference method. Both active and passive estimators are evaluated on this test set, with the performance measured using the mean-squared relative error:

$$\text{Error} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{\left\| u_i^{\text{true}} - u_i^{\text{predicted}} \right\|_{L^2}^2}{\| u_i^{\text{true}} \|_{L^2}^2}.$$

We report the relative error instead of the absolute error to normalize for potential arbitrary scaling due to the norms of the true solution function. The FNO model has four Fourier layers and $N/2$ Fourier modes, where N denotes the number of grid points along each spatial dimension. In our experiments, all computations are carried out on a 64×64 grid, so $N = 64$. Figures (7.1) and (7.2) show the testing error as a function of the training sample size. The performance of FNO on active data is not included in this figure due to its poor results. However, Figure (F.1) in the Appendix includes the error curve for FNO trained with active data, alongside the results for passive data.

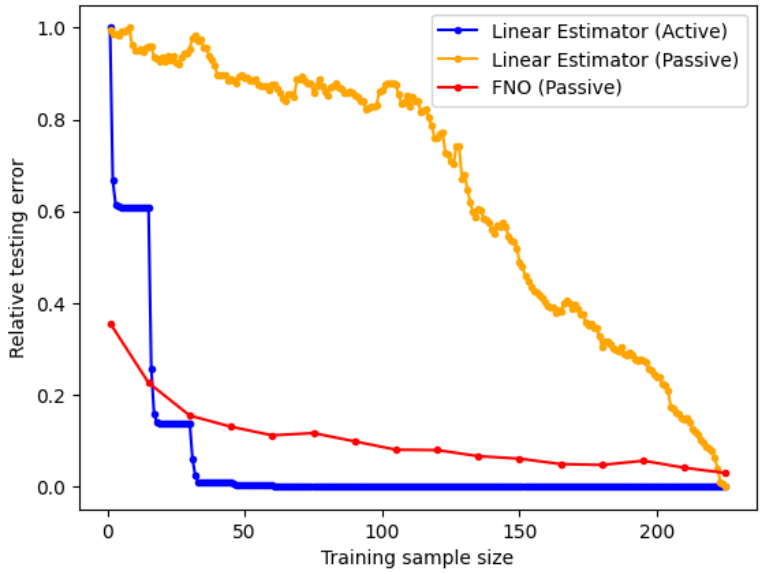


Figure 7.1: Error Plots for various estimators for Poisson Equation. The blue curve shows the performance of our linear estimator on actively collected data. The orange and red curves include the linear estimator's and FNO's performance on passively collected data. Figure 7.2 shows the same plot in log-scale.

While the convergence guarantee of our estimator is formally established only for the covariance operator $50^2(-\nabla^2 + \mathbf{I})^{-\gamma}$ with $\gamma > 1$ as $d = 2$, we observe that the estimator demonstrates robust convergence even when $\gamma \leq 1$ in the context of Poisson equation. Figure (7.3) presents the convergence rate of our estimator in log scale, using actively collected data across various values of γ.

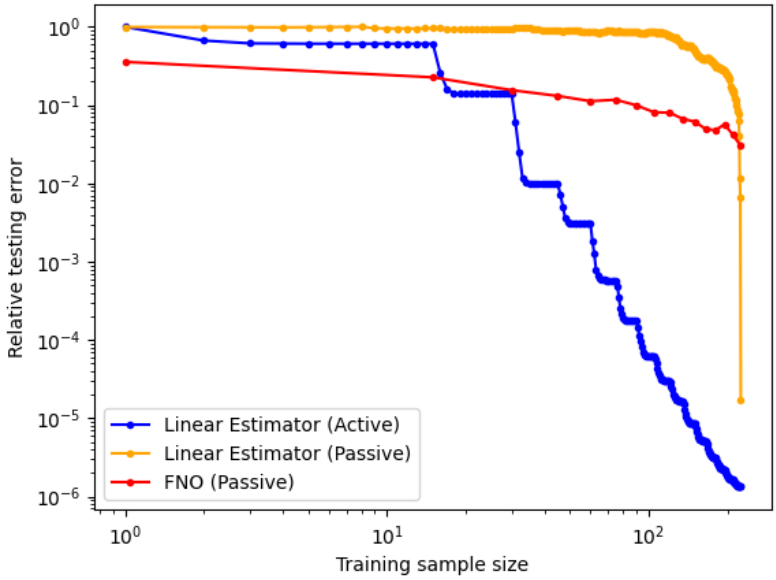


Figure 7.2: Error Plots for Poisson Equation in log-log scale.

7.5.2 Heat Equation

Consider the heat equation

$$\frac{\partial u}{\partial t} = \tau \nabla^2 u,$$

where $u : [0,1]^2 \to \mathbb{R}$ vanishes on the boundary. The solution operator for this equation is given by $\exp(\tau t \nabla^2)$, and the solution at time $t \geq 0$ can be expressed as $u_t = \exp(\tau t \nabla^2) u_0$. Fixing $t = 1$, our objective is to learn the solution operator $\exp(\tau \nabla^2)$. This operator is defined as

$$\exp(\tau \nabla^2) = \sum_{k=0}^{\infty} \frac{(\tau \nabla^2)^k}{k!},$$

which is a bounded linear operator. As in the previous case, we sample n initial conditions $u_0 \sim \mathrm{GP}(0, (-\nabla^2 + \mathbf{I})^{-1.5})$. For each initial condition, we use the finite difference method with forward-time discretization to compute the solution u_1 at $t = 1$. This is done using 1000 time discretization steps on a 64×64 grid. For our experiments, we set $\tau = 10^{-2}$. As τ is the step size in the forward Euler method, choosing a larger τ would result in instability in the numerical PDE solver.

All estimators are evaluated on a test set of size 100, drawn from the same distribution as the training data. Figure (7.4) presents the relative testing errors. Furthermore, the error plot for the Fourier Neural Operator (FNO) trained on actively collected data is shown in

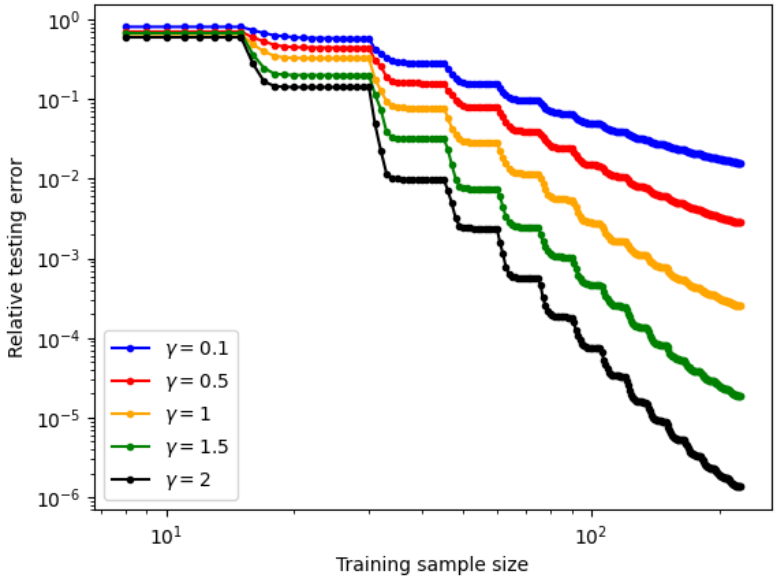


Figure 7.3: Convergence rate of the active linear estimator for Poisson equation with actively collected data for different values of γ.

Figure (F.2) in the Appendix. Finally, Figure (F.3) in the Appendix shows the convergence rates of the active linear estimator for different values of γ.

Our experimental results verify the theoretical advantage of active data collection strategies over passive sampling, as established in Theorem 27. These findings highlight the practical utility of active learning frameworks in improving data efficiency for operator learning.

7.6 Discussion

In this work, we show that arbitrarily fast rates can be achieved with an active data collection strategy when the operator of interest is a bounded linear operator and the input functions are drawn from centered distributions with continuous covariance kernels. A natural extension of these results would involve non-linear operators. Specifically, one might ask whether there exists a natural class of non-linear operators that permits such fast rates when input functions are drawn from centered distributions with continuous covariance kernels. A natural starting point might be to consider the RKHS of operators. Additionally, given that functional PCA is the estimation of truncated Karhunen–Loève decomposition, it would be interesting to explore whether a variant of a PCANet-based architecture could achieve fast rates with active data collection.

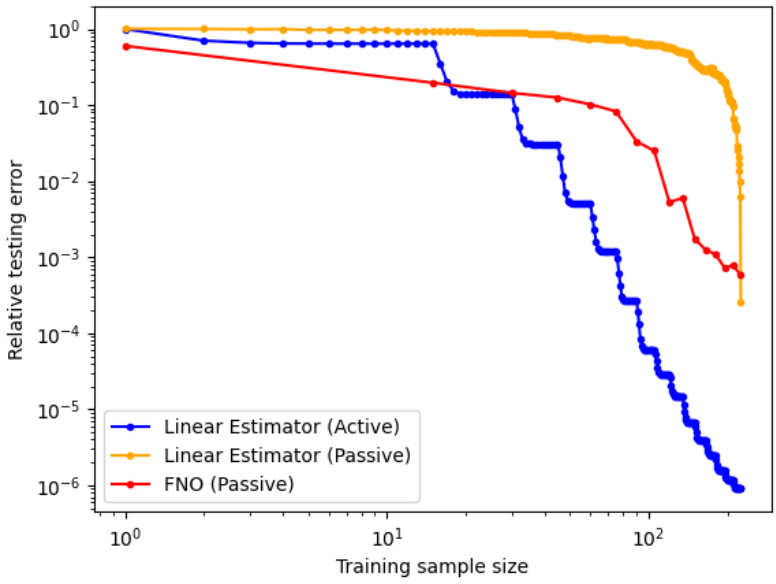


Figure 7.4: Error Plots for the Heat equation in log-scale.

CHAPTER 8

Is Zero-Shot Super-Resolution Possible in Operator Learning?

Neural operators are a class of neural-network-based operator models designed to learn complex non-linear operators [Kovachki et al., 2023]. They have demonstrated remarkable empirical success across a wide range of PDEs arising in practice, from quantum mechanics [Mizera, 2023] to fluid dynamics [Wang et al., 2024]. In addition, a growing body of work has reported an intriguing capability of neural operators, commonly referred to as zero-shot super-resolution [Li et al., 2021].

Although operator learning concerns mappings between functions defined on continuous domains, the learner often has access only to function values on a predefined discrete grid for computational feasibility. In this setting, zero-shot super-resolution refers to the phenomenon in which models trained using supervision on a coarse output grid exhibit good performance when evaluated on a substantially finer grid at test time, *without any additional retraining or fine-tuning.* Despite being frequently reported in empirical studies [Jiang et al., 2023, Yang et al., 2024, Sinha et al., 2025], zero-shot super-resolution remains poorly understood from a theoretical perspective. Thus, in this chapter[1], we formally define the problem of zero-shot super-resolution and initiate a systematic theoretical investigation of zero-shot super-resolution in operator learning.

Before presenting our results, we emphasize two important points. First, the learner is always evaluated through its predicted outputs; therefore, zero-shot super-resolution is fundamentally a property of generalization across *output resolutions*. Second, the "zero-shot" property requires that the learner does not observe output functions at finer resolutions during training. However, the learner may still have access to higher-resolution input functions, or even continuum inputs, during training. While the setting in which input functions are available at higher resolution than output functions may appear uncommon, it is not

[1]This chapter is based on: Unique Subedi and Ambuj Tewari (2026+). *Is Zero-Shot Super-Resolution Possible in Operator Learning?* In preparation.

unrealistic. In many applications, input functions are specified by the practitioner and are therefore relatively inexpensive to obtain at high resolution. In contrast, generating output functions often requires running expensive PDE solvers, making high-resolution outputs significantly more costly. The ability of many operator-learning models to handle inputs at varying resolutions is commonly referred to as discretization invariance. Although related, discretization invariance and zero-shot super-resolution are distinct concepts (see Section 8.2.1).

A first natural question in this direction is whether zero-shot super-resolution is possible at all. Our first main result is an impossibility theorem establishing that zero-shot super-resolution inference can fail even in extremely benign settings. We construct a class of rank-one operators for which learning is trivial when the training and testing grids coincide, yet for which no estimator, including empirically successful ones such as Fourier Neural Operators and DeepONets, achieves vanishing error when evaluated on a finer grid than the one used for training. We further support our theoretical results with numerical experiments that demonstrate the failure modes predicted by our lower bound (see Figure 8.1). This lower bound complements recent empirical findings of Gao et al. [2025], Sakarvadia et al. [2025], which provide evidence that zero-shot super-resolution can fail in certain regimes. Taken together, these results establish that zero-shot super-resolution is impossible without additional structural assumptions.

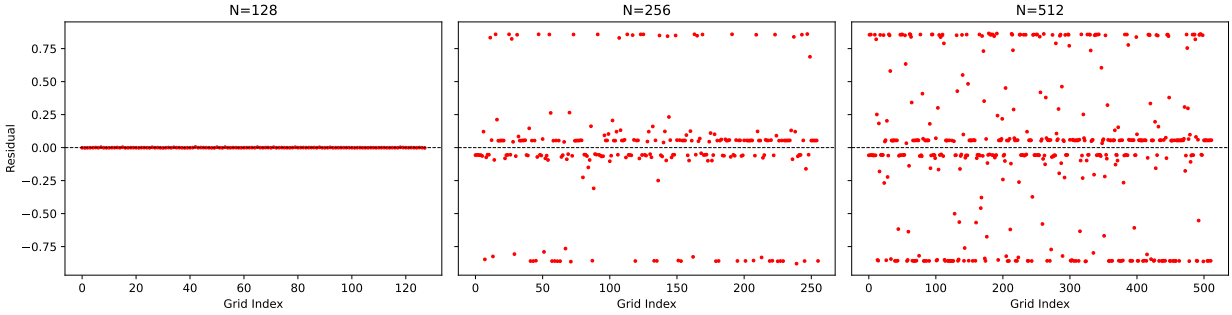


Figure 8.1: Residuals of the predicted output $\hat{w}$ and the ground truth w at test resolutions $N = 128, 256, 512$, with training performed at $N = 128$. While residuals are negligible at the training resolution, errors grow substantially as the grid is refined, illustrating the failure of zero-shot super-resolution.

Given our lower bound, the next natural question is when zero-shot super-resolution is possible. We show that zero-shot super-resolution can indeed be achieved when the output functions produced by both the ground-truth operator and the learned operator are Hölder continuous. Under this assumption, we derive a zero-shot generalization bound that decomposes the test-time error on a finer grid into the error on the coarse grid and a term

capturing extrapolation across resolutions. Finally, we connect our assumptions to classical PDE theory and modern neural-operator architectures. We show that Hölder regularity arises naturally for solutions of many elliptic and parabolic PDEs via standard regularity results, and we verify that common neural-operator architectures produces continuous outputs under mild conditions on their kernels and biases.

8.1 Related Works

Zero-shot super-resolution in operator learning was first reported by Li et al. [2021] in the context of Fourier Neural Operators. Since then, the ability of neural operators to exhibit zero-shot super-resolution has been reported in a number of empirical studies across diverse applied settings [Jiang et al., 2023, Luo et al., 2024, Yang et al., 2024, Yasuda and Onishi, 2025, Sinha et al., 2025].

Such cross-resolution generalization is feasible in the first place because these models are discretization invariant, in the sense that they can be evaluated on grids with resolutions different from those used during training. These discretization-invariant operator-learning frameworks were by developed in a series of works, with representative examples including [Bhattacharya et al., 2021, Li et al., 2020b, 2021, Lu et al., 2021, Nelsen and Stuart, 2021, Bartolucci et al., 2023]. We refer the reader to the survey articles by Boullé and Townsend [2024], Kovachki et al. [2024b], Azizzadenesheli et al. [2024], and Subedi and Tewari [2026] for a more comprehensive overview.

Although zero-shot super-resolution has been widely reported in empirical studies, its failures have also been documented. For example, Li et al. [2024b] observed that Fourier Neural Operators fail to match the behavior of Kolmogorov flow in frequency regimes not present in the training data, and proposed physics-informed constraints as a corrective measure. Similarly, Raonic et al. [2023] noted the inability of common operator-learning models such as Fourier Neural Operators, DeepONets, and Galerkin Transformers to achieve zero-shot super-resolution for transport equations, and hypothesized that this limitation arises from a lack of translation equivariance. They empirically showed that convolutional neural operators, which are translation equivariance, can generalize to unseen resolutions for these equations. Moreover, Gao et al. [2025] reported analogous failures of standard operator-learning methods in cross-resolution generalization and proposed techniques to mitigate them. Gao et al. [2025] also established an upper bound on the difference in prediction error between operators evaluated at two different resolutions and inferred, based on the upper bound, that the error can grow as the resolution gap increases. While such upper bounds are useful for building intuition, they do not conclusively rule out zero-shot super-resolution, as doing so

requires a corresponding lower bound on the error.

A recent empirical study by Sakarvadia et al. [2025] systematically evaluated the methods proposed in Li et al. [2024b], Raonic et al. [2023], Gao et al. [2025] and demonstrated that these approaches still fail to achieve reliable cross-resolution inference. Based on these findings, Sakarvadia et al. [2025] concluded that zero-shot super-resolution is fundamentally an out-of-distribution generalization problem and proposed multi-resolution training as a remedy. While this strategy can improve performance at higher resolutions, it breaks the zero-shot paradigm by allowing the model access to higher-resolution data during training.

8.2 Problem Formulation

Let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ be bounded domains. Define $L^2(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} |f(x)|^2 \, dx < \infty\}$ as the space of square-integrable functions over $\mathcal{X}$. Let $\mathcal{V} \subseteq L^2(\mathcal{X})$ and $\mathcal{W} \subseteq L^2(\mathcal{Y})$ be Banach spaces of functions, and suppose the ground truth operator of interest is $\mathrm{G} : \mathcal{V} \to \mathcal{W}$. For example, G might represent a PDE solution operator.

In the statistical learning setting, the learner observes n i.i.d. samples $\{(v_i, w_i)\}_{i=1}^n$, where $v_i \sim \mu$ and $w_i = \mathrm{G}(v_i)$ (or a noisy version thereof). Crucially, we assume that outputs w_i are only available on a discrete output grid $\mathcal{Y}_{\text{train}} \subset \mathcal{Y}$, with $|\mathcal{Y}_{\text{train}}| < \infty$. For now, we assume that the inputs v_i are accessible over the entire continuous domain $\mathcal{X}$. The case where inputs are observed only on a discrete grid $\mathcal{X}_{\text{train}}$ is discussed in Section 8.6, with additional related remarks provided in Section 8.2.1. Using the observed data, the learner constructs an estimator $\widehat{\mathrm{F}}_n$. For simplicity, we use the same notation to refer both to the learned operator and to the learning rule. Although $\widehat{\mathrm{F}}_n$ can in principle be any map $\mathcal{V} \to \mathcal{W}$, in practice it is usually chosen from a restricted function class $\mathcal{F} \subseteq \mathcal{W}^{\mathcal{V}}$, such as neural operators [Kovachki et al., 2023]. Note that the ground truth operator G may not belong to $\mathcal{F}$.

At the very least, the learner's goal is to minimize the expected error on the training grid defined as

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}) := \mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{train}}|}\sum_{y\in\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}(v)(y)\right)^2\right].$$

However, often at test time, the output of the estimator is evaluated on a finer grid $\mathcal{Y}_{\text{test}} \supseteq \mathcal{Y}_{\text{train}}$, where $|\mathcal{Y}_{\text{test}}| < \infty$. The test-time error is defined as

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{test}}) := \mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}(v)(y)\right)^2\right].$$

We emphasize that the estimator is trained using supervision only on the coarse grid $\mathcal{Y}_{\text{train}}$, yet is expected to generalize to the finer grid $\mathcal{Y}_{\text{test}}$. This is precisely the central question of zero-shot super-resolution. Accordingly, we seek conditions under which a small expected error $\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}})$ on the training grid guarantees a correspondingly small expected error on the finer testing grid $\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{test}})$, even though the learning algorithm has never observed outputs on $\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}$. Formally, we ask whether there exists a condition such that, for every $\varepsilon > 0$, if there exists n sufficiently large for which

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}) \leq \varepsilon,$$

then

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{test}}) \leq c\,\varepsilon$$

for some universal constant $c > 0$. Ideally, we want the aforementioned condition to be distribution-free. If such condition exists, then we say that zero-shot super-resolution is possible for the tuple $(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}, \mathcal{Y}_{\text{test}})$.

8.2.1 Zero-Shot Super-Resolution vs Discretization Invariance

In the operator learning community, the concept of zero-shot super-resolution is sometimes conflated with a related notion of discretization invariance, as formalized in [Kovachki et al., 2023, Definition 4]. An operator F is said to be discretization invariant with respect to a sequence of nested input grids $\{\mathcal{X}_k\}_{k\in\mathbb{N}}$, where $\mathcal{X}_k \subseteq \mathcal{X}_{k+1}$ and $\lim_{k\to\infty} \mathcal{X}_k = \mathcal{X}$, if

$$\lim_{k\to\infty} \sup_{v\in\mathcal{A}} \left\| \mathrm{F}(v_{|\mathcal{X}_k}) - \mathrm{F}(v) \right\|_{L^2(\mathcal{Y})} = 0$$

for every compact subset $\mathcal{A} \subseteq \mathcal{V}$, where $v_{|\mathcal{X}_k}$ denotes the restriction of v to the grid $\mathcal{X}_k$. This definition presupposes a meaningful way to evaluate $\mathrm{F}(v_{|\mathcal{X}_k})$.

Importantly, discretization invariance focuses solely on input resolution. It ensures that as the input grid becomes increasingly refined, the operator's output converges to its continuum evaluation. However, zero-shot super-resolution concerns a different question: how well an operator trained on a coarse *output* grid generalizes to a finer one. Since discretization invariance is defined using the continuum norm in the output space, it provides no information about generalization across different *output grids*.

To clearly separate these two notions, we begin by assuming that input functions v are available over the entire domain $\mathcal{X}$. This removes the issue of input discretization, allowing us to focus purely on output resolution. In fact, discretization invariance is trivially satisfied in this setting, as $\mathcal{X}_k = \mathcal{X}$ for all k. Despite this, our impossibility result in Section 8.3

shows that zero-shot super-resolution remains a fundamentally nontrivial problem, even when discretization invariance holds.

8.3 Impossibility of Zero-Shot Super-Resolution: A Lower Bound

We now show that generalization to finer output grids cannot always be guaranteed, even for simple operator learning problems. Specifically, we construct a setting where learning is trivial when the training and testing grids coincide ($\mathcal{Y}_{\text{train}} = \mathcal{Y}_{\text{test}}$), but impossible when the model is trained on a coarse grid and evaluated on a finer one.

Let $\mathcal{X} = \mathcal{Y} = [0, 1)$, where the exclusion of the endpoint is simply so that our construction allows periodic boundary conditions. Let $N_1, N_2 \in \mathbb{N}$ with $N_2 > N_1$. The learner observes outputs $w_i = \mathrm{G}(v_i)$ only on the coarse, uniformly spaced grid of size N_1 but evaluated on the finer grid of size N_2. That is,

$$\mathcal{Y}_{\text{train}} := \left\{0, \tfrac{1}{N_1}, \ldots, \tfrac{N_1-1}{N_1}\right\} \quad \text{and} \quad \mathcal{Y}_{\text{test}} := \left\{0, \tfrac{1}{N_2}, \ldots, \tfrac{N_2-1}{N_2}\right\}.$$

To ensure that $\mathcal{Y}_{\text{train}} \subseteq \mathcal{Y}_{\text{test}}$, we assume $N_2 = mN_1$ for some $m \in \mathbb{N}$, as done in prior works. For example, Li et al. [2021] uses $m = 2^k$ for $k \geq 1$.

We consider a learning problem where the ground truth operator G belongs to a known class $\mathcal{G}$, though the learner does not know which specific operator in the class is realized. This reflects practical scenarios where structural knowledge about the PDE allows the learner to constrain the solution operator to a known class, even if the specific instance depends on problem-specific parameters. The result below shows that learning G is straightforward when the training and testing grids match, but zero-shot super-resolution is impossible when the evaluation grid is finer than the training grid.

Theorem 29 (Impossibility of Zero-Shot Super-Resolution)**.** *Let $\mathcal{X} = \mathcal{Y} = [0, 1)$, and let $\mathcal{G}$ be a known class of rank-one linear operators from $L^2(\mathcal{X})$ to $L^2(\mathcal{Y})$ such that $\| \mathrm{G} \|_{\text{op}} \leq 1$ for all $\mathrm{G} \in \mathcal{G}$. Let μ be a probability measure supported on $L^2(\mathcal{X})$ such that every $v \sim \mu$ satisfies*

$$a \leq |v(x)| \leq 1 \quad \textit{for all } x \in \mathcal{X},$$

for some constant $a > 0$. Then there exists a ground-truth operator $\mathrm{G} \in \mathcal{G}$ such that, for any $N_1, N_2 \in \mathbb{N}$ with $N_2 = mN_1$, the following statements hold.

(i) (Perfect learning on the training grid) There exists an estimator $\widehat{\mathrm{F}}_1' : L^2(\mathcal{X}) \to L^2(\mathcal{Y})$, depending only on a single sample (v_1, w_1) with $v_1 \sim \mu$ and the output $w_1 = \mathrm{G}(v_1)$

observed only on $\mathcal{Y}_{\text{train}}$*, such that*

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{train}}|}\sum_{y\in\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_1'(v)(y)-\mathrm{G}(v)(y)\right)^2\right]=0.$$

(ii) (Failure of zero-shot super-resolution) For every estimator $\widehat{\mathrm{F}}_n : L^2(\mathcal{X}) \to L^2(\mathcal{Y})$ *trained on samples* $\{(v_i, w_i)\}_{i=1}^n$*, where* $v_i \sim \mu$ *and each output* $w_i = \mathrm{G}(v_i)$ *is observed only on the training grid* $\mathcal{Y}_{\text{train}}$*, the expected test error on the finer grid satisfies*

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right)^2\right] \;\geq\; \frac{a^2}{4}\left(1-\frac{1}{m}\right).$$

Part (i) shows that the problem is trivially learnable when the operator is evaluated on the same grid it was trained on. However, the lower bound in part (ii) remains non-zero for any $m > 1$ even with infinitely many samples ($n \to \infty$), implying that zero-shot super-resolution is impossible in this setting. Importantly, our lower bound does not rely on any structural assumptions about the estimator $\widehat{\mathrm{F}}_n$. In particular, the estimator need not be linear, unlike the ground truth operator G. As a result, the bound applies even to highly expressive nonlinear models, including neural operators such as Fourier Neural Operators (FNOs). The proof of Theorem 29 is provided in Appendix G.1.

8.4 A Generalization Bound for Zero-Shot Super-Resolution

Given the impossibility result established in Theorem 29, a natural next step is to identify sufficient conditions under which zero-shot super-resolution becomes possible. A close inspection of the proof of Theorem 29 suggests that the lack of regularity of the output functions is a primary source of difficulty. Intuitively, without continuity, the output functions can vary arbitrarily between observed grid points, making any attempt to infer their values at unseen locations fundamentally ill-posed. This observation motivates the question of whether continuity of the output functions is sufficient to guarantee zero-shot super-resolution. While continuity is indeed sufficient, it is a qualitative property and does not yield quantitative error bounds. To obtain meaningful rates, we therefore impose a stronger regularity assumption. Specifically, we assume that the output functions are uniformly Hölder continuous and derive quantitative generalization bounds for zero-shot super-resolution under this assumption.

Recall that a function $w : \mathcal{Y} \to \mathbb{R}$ is said to be uniformly Hölder continuous with exponent

$\alpha \in (0, 1]$ on $\mathcal{Y}$ if

$$[w]_\alpha := \sup_{\substack{y_1, y_2 \in \mathcal{Y} \\ y_1 \neq y_2}} \frac{|w(y_1) - w(y_2)|}{|y_1 - y_2|_2^\alpha} < \infty.$$

Let $C^{0,\alpha}(\mathcal{Y})$ denote the space of real-valued functions on $\mathcal{Y}$ with finite Hölder exponent. In this section, we consider the case where the output space $\mathcal{W}$ is constrained to functions that are uniformly Hölder continuous for some constant $c > 0$m

$$\mathcal{W} \subseteq \left\{ w \in C^{0,\alpha}(\mathcal{Y}) \,:\, [w]_\alpha \leq c \right\}.$$

This regularity assumption allows us the learner to extrapolate between grid points in the output domain $\mathcal{Y}$.

In addition to regularity assumptions on the output functions, it is also necessary to ensure that the training and testing grids are sufficiently close for meaningful extrapolation to be possible. For example, consider a setting in which the training grid is the uniform discretization of $[0, 1)$ with N points, $\mathcal{Y}_{\text{train}} = \left\{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\right\}$, while the testing grid is given by $\mathcal{Y}_{\text{test}} = \mathcal{Y}_{\text{train}} \cup (\mathcal{Y}_{\text{train}} + 5)$, that is, the union of the same uniform grid on $[0, 1)$ and a shifted copy on $[5, 6)$. Although $\mathcal{Y}_{\text{train}} \subseteq \mathcal{Y}_{\text{test}}$, it is unreasonable to expect any meaningful extrapolation to the additional test points in $[5, 6)$ based solely on observations from $[0, 1)$. Such pathological cases typically do not arise in existing empirical studies, which usually assume that both $\mathcal{Y}_{\text{train}}$ and $\mathcal{Y}_{\text{test}}$ are uniform discretizations of the same domain, with the testing grid being a refinement of the training grid. However, since our goal is to establish results at a higher level of generality, potentially accommodating non-uniform or even arbitrary grids, we introduce quantitative notions of similarity between the training and testing grids that are sufficient to allow meaningful extrapolation.

Definition 30. *Let $\mathcal{Y}_{\text{train}}, \mathcal{Y}_{\text{test}} \subset \mathcal{Y} \subseteq \mathbb{R}^d$ be finite sets.*

(i) The coverage between the training and testing grids is defined as

$$\beta := \max_{y \in \mathcal{Y}_{\text{test}}} \min_{y' \in \mathcal{Y}_{\text{train}}} |y - y'|_2.$$

(ii) The load-balancing factor is defined as

$$\nu := \max_{z \in \mathcal{Y}_{\text{train}}} \sum_{y \in \mathcal{Y}_{\text{test}}} \mathbb{1}\left\{\text{nn}(y) = z\right\},$$

where $\text{nn}(y) \in \mathcal{Y}_{\text{train}}$ denotes the nearest neighbor of y, with deterministic tie-breaking chosen to minimize ν.

The quantity β is the one-sided Hausdorff distance from the testing grid to the training grid and quantifies how well the testing domain is covered by the training grid. The load-balancing factor ν quantifies the maximum number of testing points whose predictions rely on a single training point. Intuitively, larger values of ν place a greater burden on individual training points to represent multiple unseen testing locations, increasing the risk of extrapolation error.

Given these quantities, the following result establishes a bound on the zero-shot super-resolution generalization error.

Theorem 30 (Zero-Shot Super-Resolution Generalization Bound)**.** *Let* G *denote the ground truth operator, and let* $\widehat{\mathrm{F}}_n$ *be any estimator trained on* n *samples* $\{(v_i, w_i)\}_{i=1}^n$*, where each* $w_i = \mathrm{G}(v_i)$ *is observed only on a discrete training grid* $\mathcal{Y}_{train} \subseteq \mathcal{Y}$*. Then, for any test grid* $\mathcal{Y}_{\text{test}} \supseteq \mathcal{Y}_{\text{train}}$*, we have*

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{test}) \leq \frac{|\mathcal{Y}_{train}|}{|\mathcal{Y}_{test}|} \cdot (2\nu - 1) \cdot \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{train}) + \left(1 - \frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}\right) \cdot 8c^2 \cdot \beta^{2\alpha}.$$

This bound decomposes the expected error on the higher-resolution test grid into two components. The first term is a statistical error that reflects the estimator's average error on the training grid, while the second term captures the error due to extrapolation to unseen points. In the special case where $\mathcal{Y}_{\text{train}} = \mathcal{Y}_{\text{test}}$, there is no extrapolation. Since every test point is observed during training, the second term vanishes. Moreover, each point is its own nearest neighbor, so $\nu = 1$, and the first term reduces to $\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}})$. This exactly recovers the estimator's expected error on the training grid.

The quantity ν is introduced to rule out pathological scenarios in which a small number of training points are responsible for extrapolating to a large number of testing points. To illustrate this, consider the case where $\mathcal{Y} = [0, 1]$ and the training grid is $\mathcal{Y}_{\text{train}} = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{1}{2}, 1\right\}$. Note that there are no other points between $1/2$ and 1. Suppose the testing grid satisfies $\mathcal{Y}_{\text{test}} \subseteq \mathcal{Y}_{\text{train}} \cup \mathcal{Y}_{\text{new}}$, where the set of new testing points $\mathcal{Y}_{\text{new}} \subseteq [1 - \varepsilon, 1]$ lies near the endpoint for some small $\varepsilon > 0$. In this setting, the coverage β is at most $\max\{\varepsilon, 1/N\}$, *a good geometric coverage*. However, suppose the estimator performs well at all points in $\mathcal{Y}_{\text{train}}$ except at the endpoint $y = 1$. When N is large, the error at this single point contributes negligibly to the average error on the training grid. Yet all points in $\mathcal{Y}_{\text{new}}$ have $y = 1$ as their nearest neighbor in the training grid, so extrapolation to the unseen region depends almost entirely on this poorly estimated point. In such cases, reliable extrapolation is unrealistic. The quantity ν captures this phenomenon by measuring how many testing points rely on each training point. In the example above, $\nu = |\mathcal{Y}_{\text{new}}|$, which can be arbitrarily large depending on the number of unseen testing points. This illustrates

why controlling ν is necessary for meaningfully bounding the extrapolation error.

Next, we apply the general bound in Theorem 30 to uniform grids over $[0,1)^d$. The proof is deferred to Appendix G.3.

Corollary 3 (Uniform Grid Version of Theorem 2)**.** *Suppose $\mathcal{Y}_{train}$ and $\mathcal{Y}_{test}$ are uniform grids over $[0,1)^d$ with resolutions N_1 and N_2 respectively along each direction, such that $N_2 = mN_1$ for some $m \in \mathbb{N}$. Then, the expected test error satisfies*

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{test}) \leq 2^{d+1}\, \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{train}}) + \left(1 - \frac{1}{m^d}\right) \cdot 8c^2 \cdot \left(\frac{\sqrt{d}}{N_1}\right)^{2\alpha}.$$

The 2^{d+1} dependence in Corollary 3 is somewhat loose, as it arises from applying the general bound in Theorem 30, which does not exploit the additional regularity of uniform grids. For uniform grids, the analysis can be refined to obtain the sharper bound

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{test}}) \leq 2\, \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{train}}) + \left(1 - \frac{1}{m^d}\right) 8c^2 \left(\frac{\sqrt{d}}{N_1}\right)^{2\alpha}.$$

This refinement is obtained simply by re-deriving the proof of Theorem 30 under the additional structure imposed by uniform grids. Since the argument closely parallels the general case, we provide only a proof sketch in Appendix G.4. Observing that $1 - 1/m^d \leq 1$, the bound further simplifies to

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{test}}) \leq 2\, \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{train}}) + 8c^2 d^\alpha N_1^{-2\alpha}.$$

This bound has a direct implication. For any $\varepsilon > 0$, one can first choose the training resolution N_1 sufficiently large so that $8c^2 d^\alpha N_1^{-2\alpha} \leq \varepsilon$, and then choose the sample size n large enough to ensure $\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{train}}) \leq \varepsilon$. Together, these yield

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\mathrm{test}}) \leq 2\varepsilon.$$

Thus, for uniform grids, zero-shot super-resolution is achievable with sufficiently large training resolution and sample size, *independently of how fine the testing grid is.*

8.5 On the Assumption of Hölder Continuity of Outputs

In this section, we examine when the assumption that the output functions are Hölder continuous is reasonable. We first draw on classical PDE regularity theory to justify settings

in which the ground-truth functions are Hölder smooth. We then establish conditions under which common operator-learning models produce Hölder-continuous predictions.

8.5.1 Hölder Smoothness of Ground Truth Output Functions

Consider a linear PDE of the form

$$\mathrm{L}\, u = f,$$

where $u, f : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^d$ is a bounded domain, and u satisfies homogeneous boundary conditions. The goal is to learn the solution operator G, which maps the input f (typically representing system specifications) to the corresponding solution u. Since G is as a partial inverse to the differential operator L, it can be expressed as an integral operator. In particular,

$$u(y) = (\mathrm{G}\, f)(y) = \int_\Omega g(y, x)\, f(x)\, dx,$$

where $g : \Omega \times \Omega \to \mathbb{R}$ is the Green's function associated with L [Hartmann, 2012, Chapter 1]. The following result states that the Hölder continuity of the Green's functions is sufficient to ensure that the solution $\mathrm{G}(f)$ is also Hölder continuous.

Proposition 2. *Suppose there exists $c > 0$ such that for almost every $x \in \Omega$, the function $y \mapsto g(y, x)$ belongs to $C^{0,\alpha}(\Omega)$ with Hölder constant at most c. Then, for every $v \in L^2(\Omega)$, the output $\mathrm{G}(v)$ lies in $C^{0,\alpha}(\Omega)$.*

The proof of Proposition 2 is deferred to Appendix G.5.

Next, let us consider the case where the operator of interest is non-linear. To that end, a widely used benchmark PDE in operator learning is the equation

$$(\mathrm{L}\, u)(x) := -\sum_{i=1}^{d} \sum_{j=1}^{d} \partial_i \Big(a^{ij}(x)\, \partial_j u(x) \Big),$$

referred to as the Darcy flow equation in Li et al. [2021]. When the goal is to learn the mapping $\mathrm{G} : f \mapsto u$, the solution can be expressed via an integral operator using the Green's function. However, Li et al. [2021] instead considers the problem of learning the operator $\mathrm{G}' : a \mapsto u$. While G' is not a linear operator anymore, regularity theory still provides useful information about $\mathrm{G}'(a)$ when the coefficients $a^{ij}(x)$ satisfy the uniform ellipticity condition. Precisely, if there exists $\lambda > 0$ such that

$$\sum_{i,j=1}^{d} a^{ij}(x)\, \xi_i \xi_j \geq \lambda |\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^d \text{ and a.e. } x \in \Omega,$$

and if $\|a^{ij}\|_{L^\infty} \leq \Lambda$ and $f \in L^\infty(\Omega)$, then the classical De Giorgi–Nash–Moser theorem guarantees that the solution u is Hölder continuous in the interior of Ω. Thus, $\mathrm{G}'(a) \in C^{0,\alpha}(\Omega)$, with the Hölder exponent depending only on λ, Λ, d, and $\|f\|_{L^\infty}$. For more details and a precise statement of the result, see [Gilbarg and Trudinger, 1977, Section 8.9].

The regularity of PDE solutions is a well-studied topic and typically depends on the specific structure of the equation. Since this is covered extensively in the literature, we refer the readers to standard references such as [Evans, 2022]. However, it is worth highlighting that much of this theory is developed in terms of weak solutions. That is, the solution functions u may not be classically differentiable but instead belong to Sobolev spaces $W^{k,p}(\Omega)$, where all weak derivatives up to order k exist and lie in $L^p(\Omega)$. When $k = 1$ and $p > d$, Morrey's inequality [Evans, 2022, Section 5.6.2] ensures that such functions are also Hölder continuous with exponent $\alpha = 1 - \frac{d}{p}$. More generally, Sobolev inequalities [Evans, 2022, Section 5.6.3] imply that $u \in C^{0,\alpha}(\Omega)$ for some $\alpha > 0$ whenever $k > d/p$.

8.5.2 Hölder Smoothness of Predicted Output Functions

We next discuss conditions under which common operator-learning models produce Hölder-continuous predictions.

8.5.2.1 Linear Operators

Consider a linear integral operator of the form

$$v \mapsto \mathrm{F}(v), \qquad \text{where} \qquad \mathrm{F}(v)(y) = \int_{\mathcal{X}} k(y,x)\, v(x)\, dx.$$

Proposition 3. *If $y \mapsto k(y,x) \in C^{0,\alpha}(\mathcal{Y})$ for almost every $x \in \mathcal{X}$ with Hölder constants uniformly bounded in x, then $\mathrm{F}(v) \in C^{0,\alpha}(\mathcal{Y})$ for every $v \in L^2(\mathcal{X})$.*

The proof is identical to that of Proposition 2 and is therefore deferred to Appendix G.5.

8.5.2.2 Neural Operators

Let $\mathcal{X} = \mathcal{Y}$. Then, for a given input function v, a single layer of a neural operator is defined as

$$\mathrm{N}(v)(y) = \sigma\Big(\int_{\mathcal{X}} k(y,x)\, v(x)\, dx + b(y) \Big)$$

Here, $b : \mathcal{Y} \to \mathbb{R}$ is a bias function and $\sigma : \mathbb{R} \to \mathbb{R}$ is the relu activation function. These layers are composed sequentially to get a multilayer neural operator. The following result

establishes that, given an squared integrable function as input, each layer of neural operator produces Hölder smooth functions.

Proposition 4. *Let $b \in C^{0,\alpha}(\mathcal{Y})$, and suppose that for each $x \in \mathcal{X}$, the function $y \mapsto k(y,x)$ belongs to $C^{0,\alpha}(\mathcal{Y})$, with a Hölder coefficient uniformly bounded in x. Then for every $v \in L^2(\mathcal{Y})$, we have $\mathrm{N}(v) \in C^{0,\alpha}(\mathcal{Y})$.*

We defer the proof of Proposition 4 to Appendix G.6.

Next, we show that a multilayer neural operator $v \mapsto \mathrm{F}(v)$, defined by

$$\mathrm{F}(v) = \mathrm{N}_L \circ \mathrm{N}_{L-1} \circ \cdots \circ \mathrm{N}_1(v),$$

also produces Hölder-continuous outputs. If each layer N_t is parameterized by (k_t, b_t) satisfying the assumptions of Proposition 4, then repeated application of Proposition 4 implies that $\mathrm{F}(v) \in C^{0,\alpha}(\mathcal{Y})$. To verify that these assumptions apply at every layer, it suffices to check that the output of each intermediate layer lies in $L^2(\mathcal{Y})$. Notice that the output of the first layer is $v_1 := \mathrm{N}_1(v)$, which belongs to $C^{0,\alpha}(\mathcal{Y})$ by Proposition 4. Since $\mathcal{Y}$ is bounded, we have

$$\|v_1\|^2_{L^2(\mathcal{Y})} = \int_{\mathcal{Y}} |v_1(y)|^2 \, dy \leq \sup_{y \in \mathcal{Y}} |v_1(y)|^2 \cdot \mathrm{vol}(\mathcal{Y}) < \infty.$$

The final step uses the fact that a continuous function on a bounded domain is also bounded. Thus $v_1 \in L^2(\mathcal{Y})$, and the same argument applies inductively to all subsequent layers.

We also note that, in some implementations of neural operators, a single layer is defined as

$$\mathrm{N}(v)(y) = \sigma\left(A\, v(y) + \int_{\mathcal{X}} k(y,x)\, v(x)\, dx + b(y)\right),$$

where $A \in \mathbb{R}$ is a scalar parameters for scalar-valued functions. To derive an analog of Proposition 4, we must additionally assume that the input function v lies in $C^{0,\alpha}(\mathcal{Y})$, rather than merely being square-integrable. Under this assumption, the multilayer extension follows immediately, as each layer maps functions in $C^{0,\alpha}(\mathcal{Y})$ to the same space, preserving Hölder continuity throughout the network.

Fourier Neural Operators. Finally, we conclude this section by considering the concrete example of the Fourier Neural Operator (FNO) [Li et al., 2021]. In the FNO architecture, the kernel is defined via a truncated Fourier series. That is,

$$k_{\mathrm{FNO}}(y,x) = \sum_{|m|_\infty \leq K} \lambda_m \, e^{2\pi \,\mathrm{i}\, m \cdot (y-x)},$$

where λ_m's are the learned parameters. For any y_1, y_2, we have

$$\begin{aligned}|k_{\mathrm{FNO}}(y_1, x) - k_{\mathrm{FNO}}(y_2, x)| &= \left| \sum_{|m|_\infty \le K} \lambda_m \, e^{2\pi \,\mathrm{i}\, m\cdot(y_1 - x)} - \sum_{|m|_\infty \le K} \lambda_m \, e^{2\pi \,\mathrm{i}\, m\cdot(y_2 - x)} \right| \\ &\le \sum_{|m|_\infty \le K} |\lambda_m| \cdot |e^{2\pi \,\mathrm{i}\, m\cdot x}| \cdot |e^{2\pi \,\mathrm{i}\, m\cdot y_1} - e^{2\pi \,\mathrm{i}\, m\cdot y_2}|\end{aligned}$$

Note that $|e^{2\pi \,\mathrm{i}\, m\cdot x}| \le 1$. Since the gradient of the function $y \mapsto e^{2\pi \,\mathrm{i}\, m\cdot y}$ is $2\pi \,\mathrm{i}\, m\, e^{2\pi \,\mathrm{i}\, m\cdot y}$, the mean value theorem implies that this function is Lipschitz with constant $|2\pi \,\mathrm{i}\, m|_2 \le 2\pi |m|_2$. Therefore,

$$|k_{\mathrm{FNO}}(y_1, x) - k_{\mathrm{FNO}}(y_2, x)| \le 2\pi \left(\sum_{|m|_\infty \le K} |\lambda_m| |m|_2 \right) |y_1 - y_2|_2.$$

In other words, k_{FNO} is uniformly Lipschitz in y with Lipschitz constant $2\pi \left(\sum_{|m|_\infty \le K} |\lambda_m| |m|_2 \right)$. Therefore, Proposition 4 implies that the output of FNO under appropriate assumptions are Hölder continuous.

8.6 When Inputs Are Available Only on a Discrete Grid

Thus far, we have assumed that each input function is available on the full continuum during both training and inference. In this section, we briefly discuss how our results extend to the more realistic setting in which input functions are observed only on discrete grids.

Suppose that the input functions v are accessible only on a discrete input grid $\mathcal{X}_{\mathrm{train}}$ during training, and on a possibly finer grid $\mathcal{X}_{\mathrm{test}} \supseteq \mathcal{X}_{\mathrm{train}}$ at test time. We define the expected errors when the estimator is evaluated on the training and testing grids as

$$\mathcal{E}_\mu\left(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{train}}, \mathcal{Y}_{\mathrm{train}}\right) := \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{train}}|} \sum_{y \in \mathcal{Y}_{\mathrm{train}}} \left(\widehat{\mathrm{F}}_n\left(v_{|\mathcal{X}_{\mathrm{train}}}\right)(y) - \mathrm{G}(v)(y) \right)^2 \right],$$

and

$$\mathcal{E}_\mu\left(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{test}}, \mathcal{Y}_{\mathrm{test}}\right) := \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y \in \mathcal{Y}_{\mathrm{test}}} \left(\widehat{\mathrm{F}}_n\left(v_{|\mathcal{X}_{\mathrm{test}}}\right)(y) - \mathrm{G}(v)(y) \right)^2 \right].$$

Note that $\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{train}}, \mathcal{Y}_{\mathrm{train}})$ is still a *test-time* error, not the empirical training error on the training dataset. The subscript "train" indicates only that the estimator is evaluated on the same input and output grids on which the training data were provided.

For zero-shot super-resolution in settings where input functions are accessible only on a discrete grid $\mathcal{X}$, we require that the estimator $\widehat{\mathrm{F}}_n$ satisfies a suitable form of discretization invariance, following the notion introduced in [Kovachki et al., 2021]. Specifically, assume

there exists a nested sequence of input grids $\{\mathcal{X}_k\}_{k\in\mathbb{N}}$ such that $\mathcal{X}_k \subseteq \mathcal{X}_{k+1}$ and $\lim_{k\to\infty} \mathcal{X}_k = \mathcal{X}$, for which the estimator satisfies

$$\lim_{k\to\infty} \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{y\in\mathcal{Y}_{\text{test}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \widehat{\mathrm{F}}_n\big(v_{|\mathcal{X}_k}\big)(y) \right)^2 \right] = 0.$$

The notion of discretization invariance used here is defined with respect to the discrete output grid $\mathcal{Y}_{\text{test}}$, rather than the continuous $L^2(\mathcal{Y})$ norm considered in Kovachki et al. [2023]. However, an inspection of the proof of Theorem 8 in Kovachki et al. [2023] shows that their argument first establishes uniform convergence at every point $y \in \mathcal{Y}$ (see Equation (50) therein), and then uses the boundedness of $\mathcal{Y}$ to extend this pointwise result to convergence in the continuum norm. Since our setting requires convergence only on a finite output grid, this weaker form of discretization invariance is sufficient. Finally, while the result of Kovachki et al. [2023] holds uniformly over all inputs v in a compact subset $\mathcal{A} \subseteq \mathcal{V}$, we formulate our condition in expectation with respect to a distribution μ. If μ is supported on a compact subset of $\mathcal{V}$, their uniform guarantee directly implies the condition above.

We now show that, in addition to the conditions in Section 8.4, discretization invariance is sufficient to guarantee zero-shot super-resolution when inputs are observed only on discrete grids. Since discretization invariance is a qualitative property, it does not by itself yield a convergence rate. To make this notion quantitative, we assume the existence of a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ with $\varepsilon_k \to 0$ as $k \to \infty$ such that

$$\mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{y\in\mathcal{Y}_{\text{test}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \widehat{\mathrm{F}}_n\big(v_{|\mathcal{X}_k}\big)(y) \right)^2 \right] \leq \varepsilon_k. \tag{8.1}$$

This condition provides a rate-controlled version of discretization invariance with respect to the discrete output grid $\mathcal{Y}_{\text{test}}$.

We then obtain the following bound.

Theorem 31. *Assume that $\mathcal{X}_{\text{train}} = \mathcal{X}_{k_1}$ and $\mathcal{X}_{\text{test}} = \mathcal{X}_{k_2}$ for some $k_1, k_2 \in \mathbb{N}$. For any estimator $\widehat{\mathrm{F}}_n$ satisfying (8.1), we have*

$$\begin{aligned}\mathcal{E}_\mu\left(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\text{test}}, \mathcal{Y}_{\text{test}}\right) \leq 2(\varepsilon_{k_1} + \varepsilon_{k_2}) + 2\frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}(2\nu - 1)\,\mathcal{E}_\mu\big(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}}\big) \\ + 16\left(1 - \frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}\right) c^2 \beta^{2\alpha}.\end{aligned}$$

In summary, Theorem 31 shows that, beyond the conditions in Section 8.4, discretization invariance of the estimator with respect to the discrete output norm is sufficient to ensure

zero-shot super-resolution when inputs are observed only on discrete grids. The additional error terms ε_{k_1} and ε_{k_2} quantify the loss incurred by approximating continuous inputs using discrete samples and vanish as the input grids are refined. In particular, when these discretization errors are small and the coverage and load-balancing conditions are satisfied, reliable cross-resolution generalization is achievable.

8.7 Experiments

In this section, we present empirical results illustrating the failure modes predicted by the impossibility result in Theorem 29. We demonstrate these failures in two simple settings. First, we consider the synthetic ground-truth operator used in the construction of our lower bound. Second, we study the inviscid Burgers equation, which is known to produce irregular solutions and therefore poses challenges for cross-resolution generalization. As our model of choice, we use Fourier Neural Operators (FNOs), which have been shown to be among the most effective architectures for zero-shot super-resolution across a wide range of empirical tasks. Since the success of neural operators in zero-shot super-resolution has already been demonstrated extensively in prior works [Li et al., 2021, Jiang et al., 2023, Luo et al., 2024, Yang et al., 2024, Yasuda and Onishi, 2025, Sinha et al., 2025], we focus exclusively on failure cases. Our experiments are intended primarily for illustration of our theoretical findings rather than exhaustive evaluation. For a more comprehensive empirical studies documenting the limitations of zero-shot super-resolution, we refer the reader to [Sakarvadia et al., 2025, Gao et al., 2025].

8.7.1 Synthetic Data: Lower-Bound Setup

The input functions v are Gaussian random fields sampled on a uniform grid of $[0, 1)$. Outputs are generated using the operator G_ω, defined by

$$(\mathrm{G}_\omega\, v)(y) = f_\omega(y) \int_0^1 x\, v(x)\, dx,$$

where $f_\omega(y) \in \{-1, +1\}$ is a Rademacher function defined on the output grid. This construction isolates resolution dependence: while the scalar inner product can be accurately learned from coarse data, correct prediction on a finer output grid requires resolving the unobserved sign pattern f_ω. We train models at resolution $N = 128$ and evaluate them at test resolutions $N = 128$, 256, and 512. As shown in Table 1, performance is excellent at the training resolution but deteriorates dramatically on finer grids. Figure 8.1 further illustrates this failure through residual plots comparing the predicted and ground-truth output

functions.

Test Resolution (N)	**Relative L^2 Error**
128	0.0066
256	0.9614
512	1.2331

Table 8.1: Test performance of a model trained at resolution $N = 128$ using 2056 samples. Errors are reported as relative L^2 norms, averaged over 640 random test inputs.

8.7.2 Inviscid Burgers Equation

We next evaluate zero-shot generalization on the one-dimensional inviscid Burgers equation,

$$\partial_t u(x,t) + \partial_x\Big(\tfrac{1}{2}u(x,t)^2\Big) = 0, \qquad x \in [0,1],\ t \in [0,T],$$

with periodic boundary conditions. Initial conditions are generated as random superpositions of m Fourier modes, $u_0(x) = \sum_{j=1}^{m} A_j \sin(2\pi k x + \phi_j) + B_j \cos(2\pi k x + \psi_j)$, where $A_j, B_j \sim \mathrm{Unif}[1.0, 5.0]$, fixed wavenumber $k = 8$, and phases $\phi_j, \psi_j \sim \mathrm{Unif}[0, 2\pi)$. This construction yields smooth random fields with moderate oscillations. Each initial condition is evolved to time $T = 0.2$ using a Godunov finite-volume scheme on a fine grid ($N_{\mathrm{hi}} = 512$). The resulting data are anti-aliased and downsampled to coarser grids for training and evaluation.

A Fourier Neural Operator is trained on $N = 128$-point grids using 640 samples and evaluated at test resolutions $N = 128, 256$, and 512. Table 8.2 reports relative L^2 errors averaged over 128 test samples. The model attains high accuracy on the training grid but fails to generalize to finer resolutions.

Test Resolution (N)	**Relative L^2 Error**
128	0.0578
256	0.1606
512	0.2047

Table 8.2: Test performance of a model trained at resolution $N = 128$ on smooth Burgers data. Errors are reported as relative L^2 norms averaged over 128 random test inputs.

Figure 8.2 compares the ground truth u, predictions $\hat{u}$, and residuals $r = \hat{u} - u$ across test resolutions. On the training grid, predictions coincide almost exactly with the reference solution. As the grid is refined, however, residuals increase and develop structured oscillations, particularly near the shock where u transitions sharply between positive and

negative values. This degradation in cross-resolution generalization in nonsmooth regions highlights the necessity of Hölder regularity or related smoothness assumptions for zero-shot super-resolution.

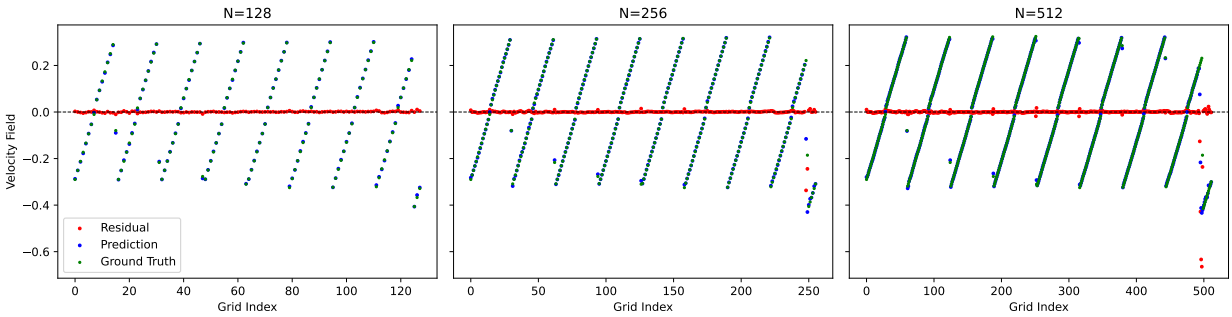


Figure 8.2: Ground truth u, predictions $\hat{u}$, and residuals $r = \hat{u} - u$ across test resolutions $N = 128, 256, 512$.

8.8 Discussion

Our work initiates a theoretical investigation of zero-shot super-resolution in operator learning, establishing both an impossibility result and sufficient conditions for its feasibility. Our impossibility result shows that zero-shot super-resolution is not guaranteed in general even for simple operators and highly expressive models. This suggest that zero-shot super-resolution should be understood as a non-trivial extrapolation problem rather than a consequence of discretization invariance or model capacity. Our positive results identify Hölder regularity as a key structural assumption under which zero-shot super-resolution becomes achievable. Finally, we also provide a generalization bound that explains the empirical success observed in smooth problem settings.

Two major directions remain open for future work. One natural extension is to establish tighter bounds for function classes characterized by different notions of regularity, such as Sobolev or Besov spaces. Second, understanding the trade-offs between strict zero-shot guarantees and multi-resolution training with limited higher-resolution samples, as suggested in by Sakarvadia et al. [2025], is also an important direction for developing reliable cross-resolution operator-learning methods.

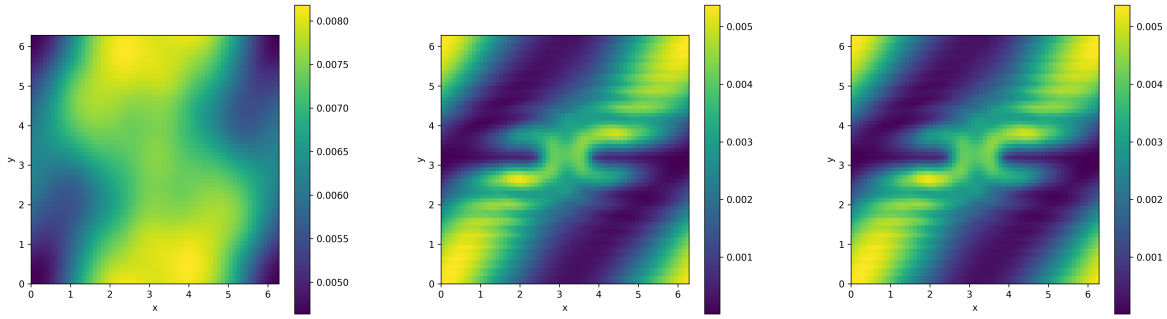
CHAPTER 9

Operator Learning for Schrödinger Equation: Unitarity, Error Bounds, and Time Generalization

In this chapter[1], we consider the problem of learning the evolution operator for the time-dependent Schrödinger equation, where the Hamiltonian may vary with time. Solving the time-dependent Schrödinger equation is of interest in various engineering applications, including material science [Liu et al., 2022, Li et al., 2020a] and the design of quantum computers [Mohammed et al., 2024, Chen et al., 2024]. In all but the simplest cases, solving this equation requires numerical methods, which are computationally expensive even for a single initial condition [Peskin and Moiseyev, 1993, van Dijk et al., 2011]. However, a single simulation is often insufficient for practical applications. For example, in qubit design, an essential requirement is that the system maintains coherence across a range of environmental conditions. This necessitates repeated simulations under varying initial conditions and system parameters, requiring significant computational resources [Nagele et al., 2023, Wu et al., 2014].

In fact, the need for repeated PDE solutions is widespread in engineering design, such as in simulating the Navier-Stokes equations for evaluating car or airfoil designs [Shahrokhi and Jahangirian, 2007, Eyi et al., 1994]. To address this, operator learning has emerged as a promising approach for the surrogate modeling of PDEs, allowing efficient computation in such cases where repeated evaluation is required [Azizzadenesheli et al., 2024, Augenstein et al., 2023]. Building on this idea, recent works have explored operator learning for accelerating solutions to the time-dependent Schrödinger equation [Mizera, 2023, Zhang et al., 2024, Shah et al., 2024]. These works, however, typically use general-purpose neural operators [Kovachki et al., 2023], most commonly the Fourier Neural Operator (FNO) [Li et al., 2021], without explicitly using the special structures of the Schrödinger evolution, such as linearity and unitary. In related fields, it has been demonstrated that incorporating known

[1]This chapter is based on: Yash Patel*, Unique Subedi*, and Ambuj Tewari (2025). *Operator Learning for the Schrödinger Equation: Unitarity, Error Bounds, and Time Generalization*. arXiv:2505.18288.



(a) Initial Wave (b) True Wave at $T = 0.1$ (c) Our Estimator's Prediction

Figure 9.1: Squared amplitude $|\psi(x)|^2$ of the initial wave, the true wave at $T = 0.1$, and the estimator's prediction for the barrier potential with double slits on $[0, 2\pi)^2$.

physical priors is often crucial for effective surrogate learning in data-scarce settings [Batzner et al., 2022, Merchant et al., 2023].

Thus, in this chapter, we propose a surrogate model for the time-dependent Schrödinger equation that exploits the fundamental structures of this equation. Specifically, let F be the true evolution operator of the Schrödinger equation that maps the initial wave function ψ to its evolved state at some fixed time $T > 0$. We introduce an active data collection strategy and a *linear estimator* $\widehat{\mathrm{F}}$ to approximate F and establish the following.

(i) **(Empirical Evaluation)** We evaluate $\widehat{\mathrm{F}}$ across a range of Hamiltonians, including hydrogen atoms, double-slit potentials (see Figure 9.1), an ion trap used in qubit design, and optical lattices. Our estimator consistently outperforms neural operators baselines such as the FNO and DeepONet by about 2 orders of magnitude, achieving relative errors about 10^{-2} times smaller than these baselines.

(ii) **(Preservation of Physical Laws)** We prove that the surrogate $\widehat{\mathrm{F}}$ preserves a weak form of unitarity, a fundamental property of the Schrödinger equation. Thus, our work aligns with broader efforts in physics-informed learning to incorporate known physical symmetries and conservation laws into model design and training [Li et al., 2024b, Richter-Powell et al., 2022].

(iii) **(Theoretical Guarantees)** We prove upper and lower bounds on the prediction error of $\widehat{\mathrm{F}}$, establishing tight rates in terms of the number of training samples, the accuracy of the PDE solver used to generate training data, and the Sobolev smoothness of the initial wave function. Instead of bounding expected error with respect to some distribution that is common in statistical learning theory, we establish a stronger error

guarantee that holds uniformly over smooth initial waves.

(iv) **(Time Generalization)** We establish time generalization bounds, showing that $\widehat{\mathrm{F}}$ trained on data up to time point T can extrapolate to future time points $T' > T$ when the potential is time-independent and sufficiently smooth. While time generalization has been studied empirically in prior works on the Schrödinger equation [Mizera, 2023, Shah et al., 2024], little is known theoretically about when such generalization is possible. To the best of our knowledge, this work provides the first mathematical guarantees for time generalization in operator learning.

9.1 Related Works

One of the earlier works in this area was by Mizera [2023], who used Fourier Neural Operators (FNOs) by Li et al. [2021] to estimate the evolution operator for simple quantum systems, including random potential functions and the double-slit potential. Additionally, they studied the time generalization ability of the learned operator by evolving the wave function beyond the time range included in the training dataset. Notably, their learned operator is more general as it maps from the initial wave function and potential function to the evolved state, rather than learning a propagator for a fixed Hamiltonian. A related work by Niarchos and Papageorgakis [2024] considers learning phases of amplitudes in scattering problems. While these works primarily study isolated quantum systems, Zhang et al. [2024] and Zhang et al. [2025a] used FNO-based architectures to study dissipative quantum systems, where the system interacts with the surrounding environment and can be driven by external fields. They also assessed the time generalization capacity of the learned propagator. Recently, Shah et al. [2024] used FNOs to learn the evolution operator of quantum spin systems and studied their ability for both single-step and multi-step time generalization.

We also note the related work on learning unitary transformations between large but finite-dimensional Hilbert spaces [Bisio et al., 2010, Hyland and Rätsch, 2017, Belov and Malyshkin, 2024], a problem that has been studied extensively in areas ranging from quantum state simulation [Johansson et al., 2012] to quantum computing [Huang et al., 2021]. There is also a growing body of work on learning linear operators from data in the context of PDE modeling [de Hoop et al., 2023, Mollenhauer et al., 2022, Subedi and Tewari, 2025a], which is most closely related to the perspective adopted in this work. Finally, our linear estimator shares some conceptual similarity with the exact diagonalization methods commonly used to numerically solve the Schrödinger equation [Lin et al., 1993]. However, unlike exact diagonalization, which explicitly diagonalizes the Hamiltonian matrix, our approach

defines an estimator for the solution operator in an arbitrary basis without performing any diagonalization. A more detailed discussion of related works is provided in Appendix H.1.

9.2 Preliminaries

Let $\mathbb{R}$ and $\mathbb{C}$ denote the real and complex numbers, respectively, while $\mathbb{N}$ and $\mathbb{Z}$ represent the natural numbers and integers. Define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For any $x \in \mathbb{R}^d$, the ℓ^p-norm is denoted by $|x|_p$. For $z \in \mathbb{C}$ with $z = a + bi$, define $|z| = \sqrt{a^2 + b^2}$. For $x, y \in \mathbb{R}^d$, the Euclidean inner product is denoted by $x \cdot y$. For two non-negative functions f and g defined on $\mathbb{N}$, we say $f \lesssim g$ if there exists a universal constant c and $n_0 \in \mathbb{N}$ such that $f(n) \leq cg(n)$ for all $n \geq n_0$. Equivalently, $f \gtrsim g$ means there exists a constant c and $n_0 \in \mathbb{N}$ such that $f(n) \geq cg(n)$ for all $n \geq n_0$.

Let $\Omega \subset \mathbb{R}^d$ be a bounded set. Define $L^2(\Omega) := \{u : \Omega \to \mathbb{C} \mid \int_\Omega |u(x)|^2 \, dx < \infty\}$ to be the set of squared-integrable functions on Ω. This is a Hilbert space with the inner product $\langle u, v \rangle_{L^2} = \int_\Omega u(x) \, \overline{v(x)} \, dx$, where $\overline{z} = a - bi$ is the complex conjugate of $z = a + bi$. The norm induced by this inner product is denoted by $\| \cdot \|_{L^2}$.

9.2.1 Time-Dependent Schrödinger Equation

The time-dependent Schrödinger equation states that the quantum system evolves as

$$\mathrm{i}\,\hbar \, \partial_t \psi(\cdot, t) = \mathrm{H}(t) \, \psi(\cdot, t).$$

Here, $\psi(\cdot, t) : \Omega \subseteq \mathbb{R}^d \to \mathbb{C}$ is the wave function at time point t, and $\mathrm{H}(t)$ is the Hamiltonian operator that can evolve over time. For a single particle system, we generally have $d \in \{1, 2, 3\}$, and the Hamiltonian can be written as $\mathrm{H}(t) = -\frac{\hbar^2}{2m}\Delta + V_t(\cdot)$, where $V_t(\cdot)$ is a time-dependent potential function and Δ is the Laplacian operator defined as $\Delta f := \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j^2}$ for Euclidean coordinates.

The solution operator, also referred to as the time evolution operator, of the Schrödinger equation takes the form $\mathcal{T}\left[\exp\left(-\frac{\mathrm{i}}{\hbar}\int_{t_0}^t \mathrm{H}(s)\, ds\right)\right]$. That is, given an initial wave $\psi(\cdot, t_0)$ at time t_0, the wave function at time point $t > t_0$ is given by $\psi(\cdot, t) = \mathcal{T}\left[\exp\left(-\frac{\mathrm{i}}{\hbar}\int_{t_0}^t \mathrm{H}(s)\, ds\right)\right]\psi(\cdot, t_0)$. Here, $\mathcal{T}$ denotes the time-ordering operator, which ensures that Hamiltonians at different time points are applied in the correct temporal sequence. This ordering is necessary because the time-varying Hamiltonians do not generally commute at different times. One can formally define this operator in terms of a Dyson series expansion [Sakurai and Napolitano, 2020, Chapter 2]. When the Hamiltonian is constant over time, $\mathrm{H}(t) = \mathrm{H}$, the evolution operator has a more familiar form $\exp\left(-\frac{\mathrm{i}(t-t_0)}{\hbar}\,\mathrm{H}\right)$. While the solu-

tion operator has a closed-form representation, it is generally impractical for computation, as evaluating operator exponentials requires computing higher-order powers of the Hamiltonian and summing an infinite series in its Taylor expansion.

9.2.2 Problem Formulation and Goal

Define $t_0 = 0$ without loss of generality. For a fixed time $T > 0$, we seek to learn the operator

$$\mathrm{F} := \mathcal{T}\left[\exp\left(-\frac{\mathrm{i}}{\hbar}\int_0^T \mathrm{H}(s)\,ds\right)\right].$$

Given a collection of n observed samples $\{(\psi_j(\cdot,0), \psi_j(\cdot,T))\}_{j=1}^n$, where each sample satisfies $\psi_j(\cdot,T) = \mathrm{F}\left(\psi_j(\cdot,0)\right)$, our goal is to construct an estimate $\widehat{\mathrm{F}}_n$ that approximates F accurately under a suitable metric. More precisely, we want to develop a data generation strategy and an estimation rule such that the error of the estimator,

$$\sup_{\psi\in\mathcal{V}} \|\widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi)|_{L^2},$$

is small. Here, $\mathcal{V}$ is some suitable subset of $L^2(\Omega)$ (consisting of, say, Sobolev-type smooth functions) that will be specified later. Note that our goal is to provide a *uniform error bound* over $\mathcal{V}$ rather than a bound on expected error that is more common in statistical learning theory. This stronger guarantee is possible in our setting because we can evaluate $\mathrm{F}(\psi)$ for any chosen initial wave function ψ, up to certain numerical accuracy, using a PDE solver. This flexibility allows us to actively select the most informative queries for constructing the estimator, rather than relying on the i.i.d. sampling typically used in statistical learning.

9.3 Data Collection Strategy and Estimator

In this section, we introduce our data generation strategy and define our estimation rule. Then, we establish the unitarity of the proposed estimator. For clarity of exposition, we assume that $\Omega = \mathbb{T}^d$, the d-dimensional torus [Grafakos, 2008, Chapter 3]. Extension to non-periodic domains is straightforward and is discussed in Section H.2. Throughout this paper, we identify $\mathbb{T}^d$ by $[0,1]^d$ equipped with periodic boundary conditions. Before moving forward, we first define the Sobolev space on $\mathbb{T}^d$. The theoretical guarantees established in this work apply only to initial waves that belong to these Sobolev spaces on $\mathbb{T}^d$ or their equivalent counterparts in more general domains Ω.

For any $k \in \mathbb{Z}^d$, define the function $\varphi_k : \mathbb{T}^d \to \mathbb{C}$ by $\varphi_k(x) := e^{2\pi i k\cdot x}$. Using these functions,

for any $s > 0$, we can define the Sobolev space as

$$\mathcal{H}^s(\mathbb{T}^d) := \left\{ f \in L^2(\mathbb{T}^d) \quad : \quad \sum_{k \in \mathbb{Z}^d} (1 + |k|_2^2)^s \, |\langle f, \varphi_k \rangle_{L^2}|^2 < \infty \right\}.$$

This space is equipped with the norm $\|f\|_{\mathcal{H}^s} := \sqrt{\sum_{k \in \mathbb{Z}^d} (1 + |k|_2^2)^s \, |\langle f, \varphi_k \rangle_{L^2}|^2}$. Although $\mathcal{H}^s(\mathbb{T}^d)$ is a Hilbert space, we will not rely on its Hilbertian properties in this work. When $s \in \mathbb{N}$, this definition of the Sobolev space is equivalent to its formulation based on derivatives. That is, $\mathcal{H}^s(\mathbb{T}^d) \simeq \left\{ f \in L^2(\mathbb{T}^d) \,\middle|\, \sum_{|\alpha|_1 \leq s} \|\mathrm{D}^\alpha f\|_{L^2}^2 < \infty \right\}$, where D^α denotes the differential operator. For more details on this equivalence, we refer the readers to [Taylor, 2011, Chapter 4.3].

9.3.1 Estimator

Given a sample size budget of n such that $n \geq 3^d$, define $K_n := (n^{\frac{1}{d}} - 1)/2$. The estimator is constructed by querying a PDE solver to obtain $w_k = \mathrm{P}(\varphi_k)$ for each $k \in \mathbb{Z}^d$ such that $|k|_\infty \leq K_n$, where P denotes the numerical PDE solver for the corresponding solution operator F of the Schrödinger equation. Since the number of such frequency indices satisfies $|\{k \in \mathbb{Z}^d : |k|_\infty \leq K_n\}| = (2 \lfloor K_n \rfloor + 1)^d \leq n$, the sample budget of n is not exceeded. Using the labeled dataset $\{(\varphi_k, w_k)\}_{|k|_\infty \leq K_n}$, we then define the estimator

$$\widehat{\mathrm{F}}_n := \sum_{k \in \mathbb{Z}^d \,:\, |k|_\infty \leq K_n} w_k \otimes \varphi_k. \tag{9.1}$$

Here, $w \otimes v$ denotes a rank-one operator, defined for any $u \in L^2(\mathbb{T}^d)$ as $(w \otimes v)(u) = w \langle u, v \rangle_{L^2}$. Note that this estimator naturally extends to general domains Ω. For general Ω, one can query the eigenfunctions of the Laplacian operator. Recall that φ_k are the eigenfunctions for $\mathbb{T}^d$. For more complex domains, the estimator can also be constructed using alternative domain-specific algebraic bases, such as orthogonal polynomials or wavelets. See Appendix H.2 for a more detailed discussion.

Remark. The estimator in Equation 9.1 is closely related to the approach of Subedi and Tewari [2025a], who proposed a similar data collection strategy and estimator to highlight the advantages of active data collection in operator learning. However, their error guarantees hold only in expectation over input samples drawn from a distribution, whereas we establish a uniform bound. While our results are stated for the F, the error rates in Section 9.4 apply to any bounded linear operator, making our work a strict generalization of Subedi and Tewari [2025a]. Furthermore, using the structure of the time-dependent Schrödinger equation, we establish additional properties such as weak unitarity and time generalization in this work,

which do not necessarily hold for general linear operators.

9.3.2 On Unitarity of the Estimator

A key property of the Schrödinger equation is the unitarity of F. More precisely, F is a surjective operator on $L^2(\mathbb{T}^d)$ that preserves inner products, meaning $\langle \mathrm{F}\, u, \mathrm{F}\, v\rangle_{L^2} = \langle u, v\rangle_{L^2}, \quad \forall u, v \in L^2(\mathbb{T}^d)$. A direct consequence of this is that the L^2-norm remains invariant under evolution, $\|\,\mathrm{F}(\psi)\|_{L^2}^2 = \|\psi\|_{L^2}^2$. This property is fundamental because, when interpreting $|\psi(x,t)|^2$ as the probability density of a particle's position, unitarity ensures that the total probability is conserved over time. Given this, it is natural to ask whether our estimator also satisfies unitarity.

Strictly speaking, $\widehat{\mathrm{F}}_n$ is not fully unitary. For instance, if φ_ℓ is a Fourier mode with $\|\ell\|_\infty > K_n$, then $\widehat{\mathrm{F}}_n\varphi_\ell = 0$ despite $\|\varphi_\ell\|_{L^2} = 1$. However, $\widehat{\mathrm{F}}_n$ satisfies a weaker form of unitarity, as captured in the following proposition (proof deferred to Appendix H.3).

Proposition 5. *Suppose the solver satisfies $\langle \mathrm{P}(v), \mathrm{P}(u)\rangle_{L^2} = \langle u, v\rangle_{L^2}$. Then, the following hold.*

(i) For all $u, v \in \mathrm{span}(\{\varphi_k : |k|_\infty \le K_n\})$, we have $\langle \widehat{\mathrm{F}}_n(u), \widehat{\mathrm{F}}_n(v)\rangle_{L^2} = \langle u, v\rangle_{L^2}$.

(ii) For any $u \in L^2(\mathbb{T}^d)$, we have $\|\widehat{\mathrm{F}}_n(u)\|_{L^2} \le \|u\|_{L^2}$.

The first property shows that $\widehat{\mathrm{F}}_n$ preserves inner products within the span of the Fourier modes used in its construction. However, the estimator may still not be fully unitary as the spans of φ_k and w_k for $|k|_\infty \le K_n$ may differ, violating the surjectivity requirement. Nonetheless, property (i) ensures that the estimator preserves the L^2 norm within this subspace. Although $\widehat{\mathrm{F}}_n$ may not preserve the L^2 norm outside this span, the second property shows that it always acts as a contraction. This property is crucial for establishing the time generalization behavior of the estimator in Section 9.5.

Finally, we note that the assumption $\langle \mathrm{P}(v), \mathrm{P}(v)\rangle_{L^2} = \langle u, v\rangle_{L^2}$ holds for many standard numerical solvers for the Schrödinger equation, such as the Crank-Nicolson and operator splitting methods.

9.4 Error Analysis and Convergence Rates

A meaningful guarantee for the estimator $\widehat{\mathrm{F}}_n$ requires a guarantee on the accuracy of the PDE solver P used to approximate F. To that end, we impose the following assumption on the solver's accuracy.

Assumption 5. *The learner has black-box access to* F *through an* ε*-accurate PDE solver* P*, which satisfies*

$$\sup_{k\in\mathbb{Z}^d} \| \mathrm{P}(\varphi_k) - \mathrm{F}(\varphi_k)\|_{L^2} \le \varepsilon.$$

We note that the assumption of ε-accuracy is only for inputs φ_k's, not for arbitrary initial wave functions.

9.4.1 Upper Bounds

Under this assumption, we establish the following upper bound on the error of the estimator.

Theorem 32 (Upper Bound)**.** *Under Assumption 5, the estimator defined in Equation* (9.1) *satisfies*

$$\|\widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi)\|_{L^2} \le \|\psi\|_{\mathcal{H}^s} \left(\varepsilon\, \gamma_n + 3^s\, n^{-\frac{s}{d}}\right)$$

for every wave function $\psi \in \mathcal{H}^s(\mathbb{T}^d)$*. Here,* $\gamma_n \lesssim 1$ *when* $2s > d$*,* $\gamma_n \lesssim \sqrt{\log n}$ *when* $2s = d$*, and* $\gamma_n \lesssim n^{\frac{1}{2}-\frac{s}{d}}$ *when* $2s < d$*.*

The term $\varepsilon\gamma_n$ is the irreducible error, which remains nonzero as long as $\varepsilon > 0$. The second term $3^s n^{-s/d}$ is the estimation error, which vanishes as the sample size $n \to \infty$. This result shows that the estimation error vanishes at the rate $n^{-s/d}$ for every s and d. In the special case where $s > d/2$, the estimation error of $n^{-s/d}$ is always faster than the Monte Carlo rate of $n^{-1/2}$. This improvement is due to our data collection strategy, as described in Section 9.3.1, rather than standard i.i.d. sampling, which can never yield better than $n^{-1/2}$ convergence rate (at least for metric losses rather than its square). Finally, when the PDE solver is exact ($\varepsilon = 0$), we obtain the upper bound $\le 3^s\, \|\psi\|_{\mathcal{H}^s}\, n^{-\frac{s}{d}}$ for every $s > 0$ regardless of d.

The proof of Theorem 32, provided in Appendix H.4, proceeds via a bias-variance type decomposition of the estimator's error. Each component of the decomposition is then bounded using the fact that $\psi \in \mathcal{H}^s(\mathbb{T}^d)$. It is worth noting that the assumption of Sobolev-type smoothness is implicitly present in many applied works on operator learning. See Appendix H.2.1 for a more detailed discussion.

9.4.2 Lower Bounds

A natural question that arises is whether these upper bounds are tight. This can be studied in two stages. First, can the bound be improved for the estimator defined in Section 9.3.1? Second, beyond this specific estimator, does the bound remain tight when considering all possible linear surrogates for the Schrödinger equation? Addressing the second question is

more subtle, as establishing a meaningful information-theoretic lower bound requires precisely specifying the information accessible to the learner. The class of linear operators the learner is allowed to consider must be well-defined. One could argue that the PDE solver P itself serves as the surrogate, or in the most extreme case, that F is the optimal surrogate. Given these nuances, we leave the broader question of optimality to future work and focus here on studying the tightness of the bounds for our specific estimator $\widehat{\mathrm{F}}_n$. The lower bound on the error of our estimator is presented in Theorem 33 and its proof is deferred to Appendix H.5.

Theorem 33 (Lower Bound)**.** *There exists a Hamiltonian* H *such that for any sample budget* n *and estimator* $\widehat{\mathrm{F}}_n$ *defined in* (9.1) *obtained by querying an* ε*-approximate PDE solver for* F*, we can find a wave function* ψ *with* $\|\psi\|_{\mathcal{H}^s} \leq 2$ *such that*

$$\|\widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi)\|_{L^2} \gtrsim \begin{cases} \varepsilon + n^{-\frac{s}{d}}, & \text{if } 2s \geq d, \\ \varepsilon\, n^{\frac{1}{2}\left(\frac{1}{2}-\frac{s}{d}\right)} + n^{-\frac{s}{d}} & \text{if } 2s < d. \end{cases}.$$

This lower bound shows that the rate $n^{-s/d}$ is tight for reducible error. Additionally, even accounting for irreducible error, this bound matches the upper bound when $s > d/2$. For $s = d/2$, there remains a gap of $\sqrt{\log n}$, and the exponent is half in the regime $s < d/2$. Nonetheless, the result conclusively establishes that the irreducible error accumulates and grows to ∞ as n grows in the regime $s < d/2$. This introduces a tradeoff between sample size and reducible error, which is undesirable. Ideally, we want the error to decrease monotonically as n increases.

The proof of Theorem 33 is subtle and relies on the careful construction of both the Hamiltonian H and an ε-approximate solver for the true evolution operator F. The main challenge lies in constructing a test function ψ that simultaneously satisfies three key properties: (i) it is a valid wave function with unit L^2-norm, (ii) it has bounded Sobolev norm, and (iii) it is sufficiently challenging that our estimator constructed from the training data incurs a large prediction error when applied to ψ.

9.4.3 Refined Upper Bound Under Stronger Assumptions on PDE Solver

From the proof of the lower bound in Appendix H.5, it is evident that establishing Theorem 33 relies on a somewhat unnatural and almost adversarial PDE solver. This raises the question of whether a tighter bound can be achieved for more realistic, non-adversarial

PDE solvers. A natural way to model such solvers is to assume that their errors behave as uncorrelated random noise.

Assumption 6. *For the Fourier modes φ_k, assume that the PDE solver satisfies $\mathrm{P}(\varphi_k) = \mathrm{F}(\varphi_k) + \delta_k$, where δ_k is a random variable in $L^2(\mathbb{T}^d)$ such that: (i) $\mathbb{E}[\|\delta_k\|_{L^2}^2] \leq \varepsilon^2$ and (ii) $\mathbb{E}[\langle \delta_k, \delta_\ell \rangle_{L^2}] = 0$ for all $k \neq \ell$.*

This assumption effectively places the problem as the fixed-design regression under uncorrelated, homoscedastic noise– a well-studied model in statistics. However, unlike in classical statistics, where the learner is given a fixed design matrix, our setting allows the learner to actively choose the design matrix. Under this stronger assumption on the solver, we establish the following improved upper bound on the estimator.

Theorem 34 (Improved Upper Bound)**.** *Under Assumption 6, the estimator defined in Equation* (9.1) *satisfies*

$$\mathbb{E}\left[\sup_{\|\psi\|_{\mathcal{H}^s} \leq c} \|\widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi)\|_{L^2}\right] \leq \varepsilon + 3^s \, c \, n^{-\frac{s}{d}}.$$

Here, the expectation is taken over the randomness introduced by the δ_k's in the estimator. Notably, the expectation is applied *after* the supremum over all wave functions in the Sobolev ball of radius c. This ensures that we are still bounding the error uniformly rather than in the mean-squared sense as is common in statistical learning theory. The expectation is required only because the uniform error is now a random variable. We defer the proof of Theorem 34 to Appendix H.6. Lastly, we want to point out that a straightforward adaptation of the proof of Theorem 34 can improve this result to a high-probability bound, provided the tails of δ_k's decay sufficiently fast. In particular, assuming that δ_k's are subgaussian in $L^2(\mathbb{T}^d)$ is sufficient for this improvement.

9.5 Time Generalization

In this section, we only consider the case where the Hamiltonian remains constant over time. Recall that the operator $\widehat{\mathrm{F}}_n$ is trained to predict the wave function at time $t = T$. We now analyze the error when using it to evolve the initial wave function over multiple time steps, i.e., at $t = T, 2T, 3T, \ldots$. For any $q \in \mathbb{N}$, the true wave function at time $t = qT$ is given by $\psi_{qT} = \exp\left(-\,\mathrm{i}\,/\hbar \cdot qT\ \mathrm{H}\right)\psi(\cdot, 0) = \mathrm{F}^q\left(\psi(\cdot, 0)\right)$. Here, we use the fact that $\mathrm{H}(s) = \mathrm{H}$ for all $s \geq 0$ and $\mathrm{F} = \exp\left(-\frac{\mathrm{i}}{\hbar} T\ \mathrm{H}\right)$. Now, our goal is to quantify the deviation of the estimated evolution $\widehat{\mathrm{F}}_n^q(\psi)$ from the exact solution $\mathrm{F}^q(\psi)$.

Theorem 35. *Suppose* P *satisfies* $\langle \mathrm{P}(u), \mathrm{P}(v)\rangle_{L^2} = \langle u, v\rangle_{L^2}$. *Let* $\gamma_n \lesssim 1$ *when* $2s > d$, $\gamma_n \lesssim \sqrt{\log n}$ *when* $2s = d$, *and* $\gamma_n \lesssim n^{\frac{1}{2}-\frac{s}{d}}$ *when* $2s < d$. *Then, for any* $\psi \in \mathcal{H}^s(\mathbb{T}^d)$, *the estimator* (9.1) *satisfies*

$$\|\widehat{\mathrm{F}}_n^q(\psi) - \mathrm{F}^q(\psi)\|_{L^2} \leq \left(\varepsilon\gamma_n + 3^s\, n^{-\frac{s}{d}}\right) \sum_{j=0}^{q-1} \|\,\mathrm{F}^j(\psi)\|_{\mathcal{H}^s}.$$

Theorem 35, whose proof is provided in Appendix H.7, establishes that the estimator $\widehat{\mathrm{F}}_n$ can be used to evolve the wave function beyond the time range covered in the training set. However, this requires the true evolution operator to be sufficiently regular. Specifically, the estimator can evolve the wave function for $q-1$ additional steps as long as $\mathrm{F}^j(\psi) \in \mathcal{H}^s(\mathbb{T}^d)$ for all $j \leq q-1$. This regularity requirement is natural, given that Theorem 32 already requires that the input belong to $\mathcal{H}^s(\mathbb{T}^d)$. However, instead of requiring $\widehat{\mathrm{F}}_n^j(\psi)$ to belong to $\mathcal{H}^s(\mathbb{T}^d)$, it is sufficient for $\mathrm{F}^j(\psi)$ to be in $\mathcal{H}^s(\mathbb{T}^d)$. We also note that the linearity of the estimator $\widehat{\mathrm{F}}_n$ plays a crucial role in the proof of Theorem 35. Equally important is the fact that $\widehat{\mathrm{F}}_n$ is a contraction in L^2 (property (ii) of Proposition 5), which ensures that the time generalization bound does not suffer from worse convergence rates in terms of sample size or an exponential dependence on q. Thus, it is unclear whether similar guarantees could be derived for generic neural network-based surrogates. In fact, our empirical results suggest otherwise: our estimator exhibits significantly smaller time extrapolation error at step $j = 16$ than the single-step prediction error ($j = 1$) observed for neural operator surrogates (see Tables 9.2 and 9.4).

Although Theorem 35 provides a time generalization bound, it is expressed in terms of the rather abstract quantity $\|\,\mathrm{F}^j(\psi)\|_{\mathcal{H}^s}$. Naturally, one may ask under what conditions this norm remains bounded. The following result bounds $\|\,\mathrm{F}^j(\psi)\|_{\mathcal{H}^s}$ in terms of the properties of the potential function.

Corollary 4. *Suppose the* ε*-approximate PDE solver satisfies* $\langle \mathrm{P}(u), \mathrm{P}(v)\rangle_{L^2} = \langle u, v\rangle_{L^2}$. *Let* $\gamma_n \lesssim 1$ *when* $2s > d$, $\gamma_n \lesssim \sqrt{\log n}$ *when* $2s = d$, *and* $\gamma_n \lesssim n^{\frac{1}{2}-\frac{s}{d}}$ *when* $2s < d$. *Then, we have:*

(i) If $V(x) = a \in \mathbb{R}$, *then* $\|\widehat{\mathrm{F}}_n^q(\psi) - \mathrm{F}^q(\psi)\|_{L^2} \leq \|\psi\|_{\mathcal{H}^s}\left(\varepsilon\gamma_n + 3^s\, n^{-\frac{s}{d}}\right) q$.

(ii) If $V \in C^\infty(\mathbb{T}^d)$ *is real-valued, then* $\exists c > 0$ *such that*

$$\|\widehat{\mathrm{F}}_n^q(\psi) - \mathrm{F}^q(\psi)\|_{L^2} \leq \|\psi\|_{\mathcal{H}^s}\left(\varepsilon\gamma_n + 3^s\, n^{-\frac{s}{d}}\right) \cdot c\, q(1 + T(q-1)).$$

(iii) If $V \in \mathcal{H}^r(\mathbb{T}^d)$ for $r \geq \max\{s, d/2\}$, then $\exists c > 0$ such that

$$\|\widehat{\mathrm{F}}_n^q(\psi) - \mathrm{F}^q(\psi)\|_{L^2} \leq \|\psi\|_{\mathcal{H}^s} \left(\varepsilon\gamma_n + 3^s\, n^{-\frac{s}{d}}\right) \cdot \frac{\exp\left(c\ \|V\|_{\mathcal{H}^r} \cdot qT\right) - 1}{\exp\left(c\ \|V\|_{\mathcal{H}^r} \cdot T\right) - 1}.$$

Corollary 4 shows that time generalization is possible when the potential function V is sufficiently smooth. The extrapolation penalty varies with the regularity of V: it is linear in q when V is constant, grows polynomially when V is infinitely differentiable, and becomes exponential in q when V has only limited Sobolev regularity. Recall that $C^\infty(\mathbb{T}^d)$ are functions for which derivatives of all orders exist and are continuous. Notably, the convergence rate with respect to the sample size n remains unaffected. Our empirical findings suggest that the exponential dependence on q may be conservative, and improving this dependence on q remains an open question. Note that the dependence on V's smoothness arises because, when V lacks regularity, the evolved wave function $\mathrm{F}^j(\psi)$ may lose Sobolev smoothness. Since $\widehat{\mathrm{F}}_n$ is recursively applied at each step, smoothness is needed to repeatedly invoke the one-step generalization guarantee. Without smoothness of V, time generalization would require a different estimator with small prediction error uniformly over all of $L^2(\mathbb{T}^d)$. However, such a stronger guarantee for one-step prediction likely requires a different structural assumption on V.

The proof of part (i) relies on the fact that φ_k's are eigenfunctions of the Hamiltonian when V is constant. Part (ii) uses the seminal result due to Bourgain [1999]. The dependence on q can be sharpened to $q(1+Tq)^\varepsilon$ for any $\varepsilon > 0$ using the refined estimate from Bourgain [1999]. For part (iii), we were unable to find a corresponding result in the literature for potentials V with limited Sobolev regularity (i.e., not C^∞). We therefore derive this estimate ourselves, adapting techniques from Bourgain [1999]. See Appendix H.8 for the proof of Corollary 4.

9.6 Experiments

9.6.1 Setup

We now compare the generalization error of our proposed estimator to that of the Fourier Neural Operator (FNO) [Li et al., 2021], U-Net Neural Operator (UNO) [Rahman et al., 2022], and DeepONet [Lu et al., 2021] surrogate models across several Hamiltonians of practical interest. We note that, while the time-evolution for a time-invariant Hamiltonian can theoretically be computed using a numerical solver for the time-independent Schrödinger equation to obtain eigenvalue-eigenfunction pairs $\{(E_k, \phi_k)\}_{|k|_\infty \leq K_n}$ (where we use ϕ in place of φ to explicitly note that ϕ need not be the Fourier modes) of Hamiltonian, fitting for any fu-

Table 9.1: Summary of potentials implemented for experiments. Potentials without an explicit t dependency are time-*independent*. Full descriptions of these potentials as well as values chosen for the free parameters are provided in H.9.

Potential Name	**Expression**	**Domain**
Free Particle	$V(x,y) = 0$	$\mathbb{T}^2$
Barrier	$V(x,y) = \begin{cases} V_0 & \text{at } x = \pi,\ y \notin [\pi \pm w] \\ 0 & \text{else} \end{cases}$	$\mathbb{T}^2$
Harmonic Oscillator	$V(x,y) = \frac{1}{2}m\omega^2[(x-\pi)^2 + (y-\pi)^2]$	$\mathbb{T}^2$
Random Field	$V(x,y) \sim \text{GRF}\,(0,\ \alpha(-\Delta + \beta\mathbf{I})^{-\gamma})$	$\mathbb{T}^2$
Paul Trap	$V(x,y,t) = \left(\frac{U_0 + V_0\cos(\omega t)}{r_0^2}\right)(x^2+y^2)$	$\mathbb{T}^2$
Shaken Lattice	$V(x,y,t) = V_0\cos[k(x - A\sin(\omega t))] + V_0\cos(ky)$	$\mathbb{T}^2$
Gaussian Pulse	$V(x,y,t) = V_0 \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right)\sum_{t_0} e^{-\frac{(t-t_0)^2}{2\sigma_t^2}}$	$\mathbb{T}^2$
Coulomb	$V(\theta,\phi) = -k\frac{e^2}{r^2}$	$\mathcal{S}^2$
Coulomb Dipole	$V(\theta,\phi) = V_0\cos(\theta)$	$\mathcal{S}^2$

ture queried initial condition $\psi(\cdot,0)$ the coefficients $\{\alpha_k\}$ such that $\psi(\cdot,0) = \sum_{|k|_\infty \le K_n} \alpha_k\phi_k(\cdot)$, and finally estimating $\psi(\cdot,T) = \sum_{|k|_\infty \le K_n} e^{-iE_kT/\hbar}\alpha_k\phi_k(\cdot)$, doing so is not feasible in all but the most trivial of Hamiltonians [van Dijk and Toyama, 2007, Leforestier et al., 1991]. In turn, learning a solution operator is of interest in both cases of time-invariant and time-dependent Hamiltonians, as we consider below.

In each of the experiments, we employed a standard second-order split-step pseudospectral method as the numerical solver P, where fields were solved in natural units, such that $\hbar = 1$ and $m = 1$ [Weideman and Herbst, 1986] . Experiments were conducted over $\mathbb{T}^2$ with a uniform discretization of 256×256, with the exception of the Coulomb and dipole potentials, where solutions were sought over $\mathcal{S}^2$, with $(\phi,\theta) \in [0,2\pi) \times [0,\pi]$ discretized in an equiangular grid of size 64×32. $K_n = 16$ was fixed across experiments, meaning the proposed estimator was fitted on $\mathcal{D} := \{(\varphi_k, \mathrm{P}(\varphi_k)\}$ for $k \in \{-16,\dots,0,\dots,16\}^2$, giving a total of $(2K_n+1)^2$ samples. For the Coulomb and dipole potentials, the estimator was fit on the spherical harmonics basis elements Y_ℓ^m for $\ell = 0,...,L_{\max}$ and $m = -\ell,...\ell$, where $L_{\max} = 10$, giving a total of $(L_{\max}+1)^2$ samples. For realism, we assumed such data were measured with noise, that is that measurements were instead made on $(\varphi_k + \varepsilon_{\text{in}}, \mathrm{P}(\varphi_k) + \varepsilon_{\text{out}})$ for $\varepsilon_{\text{in}}, \varepsilon_{\text{out}} := \sigma \cdot (Z_\Re + iZ_\Im)$ with $Z_\Re, Z_\Im \sim \mathcal{N}(0,1)$ and σ being the noise scale parameter. We considered various relative noise scales in the experiments presented below.

Initial conditions for test data were drawn from a Gaussian Random Field (GRF) by defining a field $\psi(\cdot,0) = \sum_{|k|_\infty \le N/2} c_k\varphi_k$, where $N/2$ is the Nyquist frequency and $c_k = Z\alpha^{1/2}(4\pi^2||k||_2^2 + \beta)^{-\gamma/2}$, where $Z \sim \mathcal{N}(0,1)$, $\alpha = 1$, $\beta = 1$, and $\gamma = 4$. These are samples from a Gaussian distribution on $L^2(\mathbb{T}^d)$ with mean 0 and covariance operator $\alpha(-\Delta + \beta\mathbf{I})^{-\gamma}$. Similar draws were made for the Coulomb and dipole potentials, with the expansion being

Table 9.2: Average relative errors across different Hamiltonians for a relative noise level of 0.1%, computed over 100 i.i.d. test samples, with standard deviations in parentheses. Note that, for the Coulomb and dipole potential, the FNO columns instead refer to SFNO models. Dashes for DeepONet and UNO indicate that they do not handle functions on a spherical domain.

	FNO	UNO	DeepONet	Linear
Barrier	5.146e-02 (1.897e-02)	2.79e-02 (7.088e-03)	1.733e-01 (6.926e-02)	**1.596e-03 (1.584e-05)**
Coulomb	5.173e-02 (1.733e-02)	—	—	**1.464e-03 (1.437e-05)**
Dipole	5.516e-02 (1.149e-02)	—	—	**1.462e-03 (1.906e-05)**
Free	1.65e-02 (1.094e-02)	1.398e-02 (8.162e-03)	1.582e-01 (8.435e-02)	**1.595e-03 (1.673e-05)**
Gaussian Pulse	5.448e-02 (2.072e-02)	9.535e-02 (8.224e-03)	2.055e-01 (6.288e-02)	**1.597e-03 (1.495e-05)**
Harmonic Oscillator	4.249e-02 (2.163e-02)	1.005e-01 (2.328e-02)	1.605e-01 (9.755e-02)	**1.598e-03 (1.845e-05)**
Paul Trap	1.179e-01 (4.435e-02)	9.955e-01 (1.055e-02)	7.573e-01 (6.236e-02)	**1.597e-03 (1.345e-05)**
Random	1.738e-02 (9.927e-03)	1.652e-02 (7.273e-03)	1.655e-01 (1.102e-01)	**1.594e-03 (1.659e-05)**
Shaken Lattice	7.918e-02 (2.003e-02)	2.093e-02 (1.143e-02)	2.032e-01 (1.080e-01)	**1.595e-03 (2.154e-05)**

over $\{Y_\ell^m\}$ instead of $\{\varphi_k\}$.

As Fourier Neural Operators are intended to be trained on data drawn i.i.d. from the test distribution, we generated a separate training dataset $\mathcal{D}' := \{(\psi_i(\cdot, 0), P(\varphi_i)\}$ identically to the test points, such that $|\mathcal{D}'| = |\mathcal{D}|$. FNOs were fitted with K_n modes using Adam [Kingma and Ba, 2014] for 20 epochs, where the complex fields were handled in the standard manner of representing the real and imaginary components as separate channels as in [Mizera, 2023]. This setup was identically repeated for the UNO and DeepONet. We also attempted to train these neural operators using the basis functions used to construct our linear estimator instead of i.i.d. samples. However, this resulted in all models collapsing to predicting near-zero fields on the test set. As there is currently no established active data-collection baseline for neural operators, and to ensure a fair comparison, we therefore only report results for FNO, UNO, and DeepONet trained on i.i.d. samples drawn from the test distribution.

For Coulomb and dipole potentials, we used the Spherical FNO proposed by Bonev et al. [2023]. Since no such extension to spherical domains exists, we exclude DeepOnet from this comparison.

9.6.2 Estimator Accuracy

We consider several Hamiltonians of interest from quantum mechanics, drawing examples from both classical settings and of recent research interest, summarized in 9.1 with full descriptions deferred to H.9.

As alluded to earlier, we consider various relative noise levels, sweeping over relative noises of 0.01% to 1%. We present the results for a relative noise level of 0.1% in 9.2 and defer the results over the remaining noise levels to H.10. From these results, we see that the

proposed estimator significantly outperforms alternative operator learning methods across all the Hamiltonians, both time-independent and time-dependent, and over both the Fourier and spherical harmonics bases, by leveraging the known linear structure of the true solution operator. Notably, as discussed in the experimental setup, the test samples were drawn over the full spectrum, i.e., with modes defined up to the Nyquist frequency. So, the test samples can be outside the span of modes used to define the estimator. If, however, such test points are restricted to be in the span of the basis elements used to define the estimator, we observe the perfect recovery; such results are provided in H.13.

9.6.3 Estimator Under Partial Observation

We now test for the robustness of the estimator to partial observation. Analogous to the noisy observation model described in 9.6.1, many practical settings involve only partial observation of the evolved state, for which reason we sought to characterize the relative robustness of the estimators in such a setup. To simulate partial observability, we assume the spectrum of the evolved state $\mathrm{P}(\varphi_i)$ has a random fraction of its modes zeroed out at training time. We drop these uniformly at random with a fixed probability across the Fourier modes for rectilinear potentials and similarly for the spherical harmonics coefficients for spherical potentials. The estimators were then fitted against this masked dataset and evaluated against a full, unmasked dataset, i.e. against a test dataset equivalent to that used in 9.6.2. The identical procedure was used to generate the data for the FNO and DeepONet models. We fixed the noise level to be at a relative level of 0.1% and again compared the performances using the relative errors.

The results for a mask probability of 10% are given in 9.3 and those for 20% deferred to H.11. We again see that the linear estimator robustly handles such partial measurement better than the alternative estimators considered.

9.6.4 Time Generalization

To assess time generalization of our estimator, we start with an initial wave $\psi(\cdot, 0)$ and evolve it iteratively using both the true flow and our learned operator. Specifically, for $j = 1, \ldots, q$, the true evolution gives $\mathrm{P}^j(\psi(\cdot, 0))$, and we generate noisy data by adding noise as described in 9.6.1, that is $\mathrm{P}^j(\psi(\cdot, 0)) + \varepsilon$. In parallel, starting from the noisy version of the initial condition $\psi(\cdot, 0) + \varepsilon$, we evolve our estimator to obtain $\widehat{\mathrm{F}}_n^j(\psi(\cdot, 0) + \varepsilon)$. At each time step $j = 1, \ldots, q$, we compute the relative error between the two. The test initial conditions are sampled from the GRF prior described earlier. Table 9.4 shows the average relative errors for a relative noise level of 0.1%, evaluated over 100 i.i.d. test samples. Results for a higher

Table 9.3: Average relative errors across different Hamiltonians for a masking probability of 10%, computed over 100 i.i.d. test samples, with standard deviations in parentheses. Note that, for the Coulomb and dipole potential, the FNO columns instead refer to SFNO models. Dashes for DeepONet and UNO indicate that they do not handle functions on a spherical domain.

	FNO	UNO	DeepONet	Linear
Barrier	1.616e-01 (2.333e-02)	2.746e-01 (7.726e-02)	3.258e-01 (9.531e-02)	**1.594e-03 (1.453e-05)**
Coulomb	2.200e-01 (4.145e-02)	—	—	**1.461e-03 (1.792e-05)**
Dipole	1.602e-01 (3.89e-02)	—	—	**1.463e-03 (1.597e-05)**
Free	8.848e-02 (5.676e-02)	1.377e-01 (3.91e-02)	3.775e-01 (1.204e-01)	**1.595e-03 (1.818e-05)**
Gaussian Pulse	1.219e-01 (7.07e-02)	1.879e-01 (3.594e-02)	3.021e-01 (8.951e-02)	**1.597e-03 (1.552e-05)**
Harmonic Oscillator	2.255e-01 (1.123e-01)	1.319e-01 (3.497e-02)	2.895e-01 (8.112e-02)	**1.598e-03 (1.477e-05)**
Paul Trap	2.030e-01 (6.04e-02)	1.607e-01 (3.037e-02)	4.804e-01 (5.122e-02)	**1.595e-03 (1.485e-05)**
Random	7.758e-02 (3.243e-02)	9.36e-02 (1.512e-02)	2.690e-01 (1.278e-01)	**9.358e-03 (6.039e-04)**
Shaken Lattice	3.179e-01 (6.896e-02)	1.636e-01 (1.917e-02)	2.863e-01 (1.021e-01)	**1.595e-03 (1.537e-05)**

Table 9.4: Average relative time-generalization errors across different Hamiltonians for a relative noise level of 0.1%, computed over 100 i.i.d. test samples, with standard deviations shown in parentheses.

Hamiltonian	$j = 1$	$j = 2$	$j = 4$	$j = 8$	$j = 16$
Barrier	1.592e-03 (1.190e-05)	1.805e-02 (3.739e-03)	1.823e-02 (3.964e-03)	1.692e-02 (3.349e-03)	1.641e-02 (3.363e-03)
Coulomb	1.465e-03 (1.748e-05)	1.468e-03 (1.438e-05)	1.462e-03 (1.374e-05)	1.465e-03 (1.627e-05)	1.461e-03 (1.715e-05)
Dipole	1.462e-03 (1.731e-05)	1.467e-03 (1.768e-05)	1.463e-03 (1.856e-05)	1.460e-03 (1.661e-05)	1.469e-03 (1.811e-05)
Free	1.591e-03 (1.221e-05)	1.591e-03 (1.383e-05)	1.593e-03 (1.241e-05)	1.588e-03 (1.274e-05)	1.591e-03 (1.385e-05)
Gaussian Pulse	1.592e-03 (1.232e-05)	1.546e-02 (5.556e-03)	1.799e-02 (6.853e-03)	1.852e-02 (6.885e-03)	1.988e-02 (7.362e-03)
Harmonic Oscillator	1.590e-03 (1.370e-05)	1.721e-03 (3.145e-05)	1.659e-03 (2.303e-05)	1.666e-03 (2.477e-05)	1.712e-03 (3.330e-05)
Paul Trap	1.591e-03 (1.499e-05)	1.511e-01 (2.023e-02)	4.536e-01 (4.893e-02)	6.344e-01 (4.913e-02)	6.670e-01 (4.594e-02)
Random Lattice	1.591e-03 (1.269e-05)	1.592e-03 (1.350e-05)	1.591e-03 (1.223e-05)	1.589e-03 (1.128e-05)	1.589e-03 (1.306e-05)
Shaken Lattice	1.592e-03 (1.382e-05)	5.507e-03 (6.975e-04)	5.516e-03 (6.573e-04)	3.740e-03 (2.924e-04)	6.269e-03 (5.130e-04)

noise level of 1% are deferred to Appendix H.12.

For the free, harmonic oscillator, and random potential, the error remains nearly constant across time steps, indicating long-term generalization. A similar trend is observed for the Coulomb and dipole potentials on the sphere. In contrast, the error increases sharply, by an order of magnitude at $j = 2$ for the barrier potential, which is likely due to its discontinuity. For time-dependent potentials such as the Paul trap and Gaussian pulse, the estimator incurs larger errors at later steps. Note that our time-generalization bounds do not apply to these time-varying Hamiltonians. Overall, the results show that our estimator generalizes well beyond the training time points for sufficiently smooth potentials. Furthermore, the empirical error growth is notably slower than the exponential bound suggested in part (iii) of Corollary 4. We leave this possible refinement for future work.

9.7 Discussion

In this work, we introduced a linear operator surrogate to estimate the evolution operator of the time-dependent Schrödinger equation. While our method provides rigorous theoretical guarantees, it is currently limited to a single particle system. A natural direction for future work is to extend the method to handle a system with N-interacting particles. A wave function for N-particle system must be symmetric for Bosons or antisymmetric for Fermions. Thus, one approach could be to generalize our data generation and the estimator to enforce these constraints.

Additionally, our method estimates the evolution operator for a fixed Hamiltonian. An interesting extension would be to develop a surrogate that takes both the initial wave and the potential as inputs and predicts the wave function at time T. Such a method would allow generalization across different system configurations, which can be used in applications such as qubit design. However, this requires moving beyond linear operators, as the ground truth operator mapping from $(\psi(\cdot, 0), V)$ to $\psi(\cdot, T)$ is nonlinear. Constructing such nonlinear estimators with neural networks is relatively straightforward, but providing rigorous theoretical guarantees presents a significant challenge.

CHAPTER 10

Future Directions

In this chapter, we conclude by discussing several open problems and promising future directions related to the work presented in this thesis. While some specific technical open problems were discussed within individual chapters, here we adopt a broader perspective and focus on high-level questions that deserve further attention from the community. Section 10.1 discusses open directions related to Part I of this dissertation, while Sections 10.2, 10.3, and 9.5 discuss questions arising from Part II.

10.1 Learnability, Uniform Convergence, and Empirical Risk Minimization

As discussed in Chapter 5, in classical learning problems, learnability is often equivalent to the property of uniform convergence. This equivalence is known to hold for binary and finite multiclass classification ($|\mathcal{Y}| < \infty$), as well as for scalar-valued regression in both batch [Ben-David, Cesa-Bianchi, and Long, 1995, Alon, Ben-David, Cesa-Bianchi, and Haussler, 1997] and online [Rakhlin, Sridharan, and Tewari, 2015a] settings. However, this equivalence breaks down for multiclass classification when $|\mathcal{Y}| = \infty$, a phenomenon first identified by Natarajan [1989b] in the batch setting and by Hanneke, Moran, Raman, Subedi, and Tewari [2023] in the online setting. In Chapter 5, we also establish such a separation between uniform convergence and learnability in a regression problem when the target space $\mathcal{Y}$ is an infinite-dimensional Hilbert space. While our separation holds in both batch and online settings, this separation is particularly concerning in the batch setting because, in the presence of such a separation, the empirical risk minimization (ERM) principle fails to be a valid learning rule [Daniely, Sabato, Ben-David, and Shalev-Shwartz, 2015]. Indeed, Brukhim, Carmon, Dinur, Moran, and Yehudayoff [2022] constructed a specific learning rule for multiclass problems with infinite labels that is not based on ERM. However, constructing a general learning principle that applies to arbitrary supervised learning problems remains an open problem in the batch setting. Such a general learning rule for arbitrary supervised

learning problems in the online setting was recently constructed by Raman, Subedi, and Tewari [2025], which forms the basis of Chapter 4 of this dissertation. In addition, providing a general characterization of batch learnability for arbitrary supervised learning problems, in the spirit of [Raman, Subedi, and Tewari, 2025], also remains an open problem.

10.2 A General Statistical Theory of Operator Learning

Beyond the study of isolated problems, it is natural to ask whether one can develop a general statistical theory of operator learning. As discussed earlier, the evolution of statistical learning theory was driven by precisely such a goal: to identify general conditions under which learning from data is possible and to quantify the associated rates [Vapnik and Chervonenkis, 1971, Natarajan, 1989b, Alon et al., 1997, Daniely and Shalev-Shwartz, 2014, Brukhim et al., 2022]. The central challenge in developing an analogous theory for operator learning is the identification of an appropriate complexity measure for the operator class $\mathcal{F}$ that meaningfully captures the difficulty of learning. Modern statistical learning theory relies heavily on covering numbers and related notions to quantify complexity [Ben-David et al., 1995, Van Der Vaart and Wellner, 1996, Alon et al., 1997, Geer, 2000]. Recent work by Reinhardt et al. [2024] has taken an important step in this direction by establishing estimation bounds for compact operator classes in terms of their covering numbers. However, covering-number-based analyses are unlikely to provide a complete foundation for operator learning, since many natural operator classes of practical interest are inherently non-compact, making such bounds vacuous. This suggest that new notions of complexity, beyond covering numbers, are needed for operator learning. Ideally, such a measure should characterize learnability for operator classes in the same way that VC dimension characterizes learnability in binary classification [Blumer et al., 1989]. More concretely, given a class $\mathcal{F}$, one may ask whether there exists a scale-sensitive complexity measure $\mathrm{C}_\gamma(\mathcal{F})$ such that $\mathcal{F}$ is learnable if and only if $\mathrm{C}_\gamma(\mathcal{F}) < \infty$ for every $\gamma > 0$. A promising direction is to develop suitable generalizations of the fat-shattering dimension, which characterizes learnability in scalar-valued regression [Bartlett et al., 1996, Alon et al., 1997]. The classical theory of regression suggests that an appropriate complexity measure $\mathrm{C}_\gamma(\mathcal{F})$ should not only determine whether learning is possible, but also provide tight bounds on the rate at which the associated worst-case excess risk converges to zero.

10.3 Active Data Collection in Operator Learning

As discussed in Chapter 7, the i.i.d.-based statistical (passive) model is likely not the best framework for studying operator learning for PDEs. Thus, consider an active learning setting where, given a sample budget of n, the learner can select *any* input functions $v_1, v_2, \ldots, v_n \in \mathcal{X}$ and obtain the corresponding labels $\mathrm{G}(v_i)$ for each $i \in [n]$. This is in contrast to passive settings, where the inputs v_i are drawn i.i.d. from some distribution over the input function space.

Such a framework, in which the learner can request labels for arbitrary inputs, is unrealistic in many traditional applications. For example, in human-centered datasets, it may not be feasible to request a label for an individual with an arbitrary feature vector, as such an individual may not exist in reality. However, this setting is entirely natural in operator learning, since a PDE solver can provide a solution for any admissible input function in the space $\mathcal{X}$. Therefore, the learner has no inherent reason to be restricted to i.i.d. samples. In fact, as generating training data often requires computationally expensive numerical solvers, the learner should ideally select inputs adaptively so that the computational cost of training is justified by improved predictive performance at evaluation time.

This seemingly small change in the data collection protocol can have substantial statistical consequences. As discussed in Chapter 7, Subedi and Tewari [2025a] showed that for classes of linear operators, active data collection can yield dramatically improved convergence rates compared to passive sampling. In particular, under suitable assumptions on the input distribution, active protocols can achieve arbitrarily fast convergence rates, whereas in the passive i.i.d. setting rates faster than n^{-1} are impossible. These results establish the benefits of active data collection over passive ones for operator learning.

A natural open problem is to extend such efficiency gains beyond linear operator classes. Many practically relevant operator learning problems involve nonlinear operators; for example, consider the solution operator Navier-Stokes equation. Developing a general theory of active data collection for nonlinear operator learning remains an important future direction.

10.4 Theory of Time Generalization

Many operator learning problems arising from time-dependent PDEs involve learning a family of operators indexed by time. More precisely, the ground-truth operator G_T maps an initial condition u_0 to the solution u_T at time $T > 0$. Although an operator estimator $\widehat{\mathrm{F}}$ is typically trained using data corresponding to a finite time horizon T, some applications may require predicting the system at future times $T' > T$ without additional retraining. This

naturally raises the question of *time generalization*: under what conditions can a learned operator extrapolate reliably to longer unseen time horizons?

In Chapter 9, we formalized this question and established time-extrapolation guarantees for the Schrödinger equation. In particular, we showed that reliable time generalization is possible when the underlying dynamics satisfy suitable structural conditions, such as time-invariance and sufficient smoothness of the potential. However, this result relies heavily on the linearity and unitarity of the Schrödinger solution operator. While these techniques extend in a relatively straightforward manner to the problem of learning other linear operators, developing a general theory of time generalization for nonlinear operator learning problems remains largely open.

Such a theory should identify necessary and sufficient conditions for time generalization across broad classes of time-dependent operators and provide quantitative bounds on extrapolation error as a function of the training horizon, testing horizon, operator-class complexity, and regularity of the underlying dynamics. Establishing such results would yield a principled understanding of when long-horizon prediction is possible and could guide the design of models capable of reliable time extrapolation.

APPENDIX A

A Characterization of Multioutput Learnability

A.1 Complexity Measures

In learning theory, the learnability of a function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is often characterized through a suitable notion of complexity that quantifies the difficulty of the corresponding learning task. In this section, we provide precise definitions of several complexity measures that are commonly used in learning theory.

A.1.1 Complexity Measures for Batch Learning

In binary classification, the Vapnik-Chervonenkis (VC) dimension of a function class characterizes its learnability.

Definition 31 (Vapnik-Chervonenkis Dimension)**.** *A set $S = \{x_1, \ldots, x_d\}$ is shattered by a binary function class $\mathcal{H} \subseteq \{-1, 1\}^d$ if for every $\sigma \in \{-1, 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $i \in [d]$, we have $h_\sigma(x_i) = \sigma_i$. The VC dimension of $\mathcal{H}$, denoted VC($\mathcal{H}$), is the size of the largest shattered set $S \subseteq \mathcal{X}$. If the size of the shattered set can be arbitrarily large, we say that VC($\mathcal{H}$) $= \infty$.*

The learnability of a multiclass function class is characterized by its Natarajan dimension.

Definition 32 (Natarajan Dimension)**.** *A set $S = \{x_1, \ldots, x_d\}$ is shattered by a multiclass function class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ if there exist two witness functions $f, g : S \to \mathcal{Y}$ such that $f(x_i) \neq g(x_i)$ for all $i \in [d]$, and for every $\sigma \in \{-1, 1\}^d$, there exists a function $h_\sigma \in \mathcal{H}$ such that for all $i \in [d]$, we have*

$$h_\sigma(x_i) = \begin{cases} f(x_i) & \text{if } \sigma_i = 1 \\ g(x_i) & \text{if } \sigma_i = -1 \end{cases} .$$

The Natarajan dimension of $\mathcal{H}$, denoted Ndim($\mathcal{H}$), is the size of the largest shattered set $S \subseteq \mathcal{X}$. If the size of the shattered set can be arbitrarily large, we say that Ndim($\mathcal{H}$) $= \infty$.

For real-valued regression problems, the learnability is characterized in terms of the fat-shattering dimension of a function class.

Definition 33 (Fat-Shattering Dimension)**.** *A real-valued function class $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ shatters points $S = \{x_1, x_2, \ldots, x_d\}$ at scale $1 > \gamma > 0$, if there exists witness functions $r : S \to [0,1]$ such that, for every $\sigma \in \{\pm 1\}^d$, there exists $g_\sigma \in \mathcal{G}$ such that $\forall i \in [d]$, $\sigma_i(g_\sigma(x_i) - r(x_i)) \geq \gamma$. The fat-shattering dimension of $\mathcal{G}$ at scale γ, denoted $fat_\gamma(\mathcal{G})$, is the size of the largest set that can be γ-shattered by $\mathcal{G}$. If the size of the shattered set can be arbitrarily large, then we say that $fat_\gamma(\mathcal{G}) = \infty$.*

We also define a general notion of complexity called Rademacher complexity that provides a sufficient and necessary condition of uniform convergence. Since uniform convergence implies learnability, we use Rademacher complexity to argue sufficient conditions for learnability.

Definition 34 (Empirical Rademacher Complexity)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. For a bounded loss function ℓ, define the loss class to be $\ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(f(x), y)\}$. If $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ be a set of i.i.d samples drawn from $\mathcal{D}$, then the empirical Rademacher complexity of $\ell \circ \mathcal{F}$ is defined as*

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \right) \right],$$

where $\sigma \in \{\pm 1\}^n$ is a sequence of n i.i.d. Rademacher random variables.

A.1.2 Complexity Measures in Online Learning

For the complexity measures below, it is useful to define a $\mathcal{Z}$-valued binary tree [Rakhlin et al., 2015b]. A binary tree $\mathcal{T}$ of depth d is $\mathcal{Z}$-valued if each of its internal nodes are labelled by elements of $\mathcal{Z}$. Such a tree can be identified by a sequence $(\mathcal{T}_1, ..., \mathcal{T}_d)$ labelling functions $\mathcal{T}_i : \{\pm 1\}^{i-1} \to \mathcal{Z}$ which provide labels for each internal node. A path of length d is given by a sequence $\sigma = (\sigma_1, ..., \sigma_d) \in \{\pm 1\}^d$. Then, $\mathcal{T}_i(\sigma_1, ..., \sigma_{i-1})$ gives the label of node following the path $(\sigma_1, ..., \sigma_{i-1})$ starting from the root, going "right" if $\sigma_j = +1$ and "left" if $\sigma_j = -1$. Note that, $\mathcal{T}_1 \in \mathcal{Z}$ is the label for the root node. For brevity, we slightly abuse notation by letting $\mathcal{T}_i(\sigma_1, ..., \sigma_{i-1}) = \mathcal{T}_i(\sigma_{<i})$, but it is understood that $\mathcal{T}_i$ only depends on the prefix $(\sigma_1, ..., \sigma_{i-1})$. We are now ready to formally define complexity measures in the online setting.

When $\mathcal{Y} = \{-1, +1\}$ is binary, the Littlestone Dimension (Littlestone [1987]) tightly characterizes the online learnability of a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to the 0-1 loss.

Definition 35 (Littlestone Dimension). *Let $\mathcal{T}$ denote a complete binary tree of depth d whose internal nodes are labeled by elements $\mathcal{X}$ and two edges from parent to child nodes are labeled by -1 and $+1$. The tree is shattered by a binary hypothesis class $\mathcal{G} \subseteq \{-1,+1\}^{\mathcal{X}}$ if for every $\sigma \in \{-1,+1\}^d$, there exists a hypothesis $g_\sigma \in \mathcal{G}$ such that the root to leaf path $(x_1,\ldots,x_d)$ obtained by taking left when $\sigma_t = -1$ and right when $\sigma_t = +1$ satisfies $g_\sigma(x_t) = \sigma_t$ for all $1 \le t \le d$. The Littlestone Dimension of $\mathcal{G}$, denoted Ldim($\mathcal{G}$), is the maximal depth of the complete binary tree shattered by $\mathcal{G}$. If $\mathcal{G}$ can shatter a tree of arbitrary depth, we say that Ldim($\mathcal{G}$) $= \infty$.*

For finite label spaces $\mathcal{Y}$, the Multiclass Littlestone Dimension [Daniely et al., 2011] tightly characterizes the online learnability of a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to the 0-1 loss.

Definition 36 (Multiclass Littlestone Dimension). *Let $\mathcal{T}$ denote a $\mathcal{X}$-valued binary tree of depth d whose edges are labelled by elements from $\mathcal{Y}$, such that the edges from a single parent to its child-nodes are each labeled with a different label. The tree $\mathcal{T}$ is shattered by a function class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if, for every path $\sigma \in \{\pm 1\}^d$, there is a function $h_\sigma \in \mathcal{H}$ such that $h_\sigma(\mathcal{T}_i(\sigma_{<i})) = y(\sigma_i)$, where $y(\sigma_i)$ is the label of the edge between nodes $(\mathcal{T}_i(\sigma_{<i}), \mathcal{T}_{i+1}(\sigma_{<i+1}))$. The Multiclass Littlestone Dimension (MCLdim) of $\mathcal{H}$, denoted MCLdim($\mathcal{H}$), is the maximal depth of a complete binary tree that is shattered by $\mathcal{H}$. If MCLdim $= \infty$, then there exists shattered trees of arbitrarily large depth.*

When $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, the sequential fat-shattering dimension [Rakhlin et al., 2015a] at scale γ characterizes the learnability of $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ with respect to the absolute loss d_1.

Definition 37 (Sequential Fat-Shattering Dimension). *Let $\mathcal{T}$ denote a $\mathcal{X}$-valued binary tree of depth d. The tree $\mathcal{T}$ is γ-shattered by a function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ if there exists an $\mathbb{R}$-valued binary tree $\mathcal{R}$ of depth d such that for all $\sigma \in \{\pm 1\}^d$, there exists $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$,*

$$\sigma_t(h_\sigma(\mathcal{T}_t(\sigma_{<t})) - \mathcal{R}_t(\sigma_{<t})) \ge \gamma$$

The tree $\mathcal{R}$ is called the witness to shattering. The sequential fat shattering dimension of $\mathcal{H}$ at scale γ, denoted $fat^{seq}_\gamma(\mathcal{H})$, is the maximal depth of a complete binary tree that is γ-shattered by $\mathcal{H}$. If there exists γ-shattered trees of arbitrarily large depth, then $fat^{seq}_\gamma(\mathcal{H}) = \infty$.

Beyond both finite and bounded label spaces, the sequential Rademacher complexity [Rakhlin et al., 2015a] provides a useful tool for giving sufficient conditions for learnability.

Definition 38 (Sequential Rademacher Complexity). *Let $\sigma = \{\sigma_i\}_{i=1}^T$ be a sequence of independent Rademacher random variables. Let $\mathcal{T}$ be a $\mathcal{Z}$-valued binary tree of depth d. The sequential Rademacher complexity of a function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ on $\mathcal{T}$ is defined as*

$$\mathfrak{R}_T^{seq}(\mathcal{H}; \mathcal{T}) = \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{h \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} \sigma_t h(\mathcal{T}_t(\sigma_{<t})) \right].$$

Then, the worst-case sequential Rademacher complexity is defined as $\mathfrak{R}_T^{seq}(\mathcal{H}) = \sup_{\mathcal{T}} \mathfrak{R}_T^{seq}(\mathcal{H}; \mathcal{T})$.

A.2 Natarajan Dimension Characterizes Batch Multilabel Learnability

A multilabel classification problem where labels (i.e. bitstrings) in $\mathcal{Y}$ are of length K can also be viewed as multiclass classification on the target space with 2^K labels. Given this observation, the Natarajan dimension of the function class $\mathcal{F}$ continues to characterize the multilabel learnability with respect to any loss function ℓ satisfying the identity of indiscernibles.

Theorem 36 (Ben-David et al. [1995]). *Let ℓ be any loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is agnostic learnable with respect to ℓ in the batch setting if and only if $Ndim(\mathcal{F}) < \infty$.*

The proof in Ben-David et al. [1995] involves arguments based on growth function. Here, we provide proof that uses realizable and agnostic learnability due to Hopkins et al. [2022].

Proof. (of sufficiency) We first show that the finiteness of $\mathrm{Ndim}(\mathcal{F})$ is sufficient for learnability. Suppose $\mathrm{Ndim}(\mathcal{F}) < \infty$. Then, we know that $\mathcal{F}$ is agnostic learnable with respect to 0-1 loss [Ben-David et al., 1995]. Since the target space $\mathcal{Y}$ as well as the range space of $\mathcal{F}$ is finite, for every loss ℓ satisfying the identity of indiscernibles, there exists an $a > 0$ such that $a\,\ell(h(x), y) \leq \mathbb{1}\{h(x) \neq y\}$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and function $h \in \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{D}$ be a realizable distribution to $\mathcal{F}$ with respect to ℓ. Since $\ell(y_1, y_2) = 0$ if and only if $\mathbb{1}\{y_1 \neq y_2\} = 0$, the distribution $\mathcal{D}$ is also realizable with respect to 0-1 loss. Since $\mathcal{F}$ is learnable with respect to 0-1 loss, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for a sufficiently large $S \sim \mathcal{D}^n$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, we have $\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] \leq a\,\epsilon$. Using the inequality stated above pointwise, the predictor h also satisfies $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq \epsilon$. Therefore, $\mathcal{A}$ is also a realizable algorithm with respect to ℓ. Since ℓ satisfies the identity

of indiscernible, Lemma 1 guarantees the existence of agnostic PAC learner $\mathcal{B}$ for $\mathcal{F}$ with respect to ℓ. ■

Proof. (of necessity) Suppose $\mathcal{F}$ is learnable with respect to ℓ. Since the target space is finite, there must exist a constant $b > 0$ such that $\mathbb{1}\{h(x) \neq y\} \leq b\,\ell(h(x), y)$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and any function $h \in \mathcal{Y}^{\mathcal{X}}$. Let $\mathcal{D}$ be a realizable distribution with respect to ℓ. Due to the 0 alignment property, $\mathcal{D}$ is also realizable with respect to 0-1 loss. Since $\mathcal{F}$ is learnable with respect to ℓ loss, there exists a learning algorithm $\mathcal{A}$ with the following property: for any $\epsilon, \delta > 0$, for a sufficiently large $S \sim \mathcal{D}^n$, the algorithm outputs a predictor $h = \mathcal{A}(S)$ such that, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$, we have $\mathbb{E}_{\mathcal{D}}[\ell(h(x), y)] \leq b\,\epsilon$. In particular, using the inequality stated above pointwise, we obtain $\mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] \leq \epsilon$. Therefore, $\mathcal{F}$ is learnable with respect to 0-1 loss in the realizable setting. As the finiteness of the Natarajan dimension is necessary for the learnability of $\mathcal{F}$ under the 0-1 loss [Natarajan, 1989a], we must have $\mathrm{Ndim}(\mathcal{F}) < \infty$. ■

A.3 Proofs for Batch Multioutput Regression

A.3.1 Proof of Sufficiency in Theorem 3

Proof. We first prove that the agnostic learnability of each $\mathcal{F}_k$ is sufficient for the agnostic learnability of $\mathcal{F}$. As in the classification setting, the proof here is based on a reduction. That is, given oracle access to agnostic learners $\mathcal{A}_k$ for each $\mathcal{F}_k$ with respect to $\psi_k \circ d_1$ loss, we construct an agnostic learner $\mathcal{A}$ for $\mathcal{F}$ with respect to loss ℓ.

Denote $\mathcal{D}_k$ to be the marginal distribution of $\mathcal{D}$ restricted to $\mathcal{X} \times \mathcal{Y}_k$. Let us use $m_k(\epsilon, \delta)$ to denote the sample complexity of $\mathcal{A}_k$. Then, for all $k \in [K]$, the marginal samples $S_k = \{(x_i, y_i^k)\}_{i=1}^n$ with scalar-valued targets are iid samples form $\mathcal{D}_k$. For each $k \in [K]$, define $g_k = \mathcal{A}_k(S_k)$ to be the predictor returned by algorithm $\mathcal{A}_k$ when trained on S_k. Since $\mathcal{A}_k$ is an agnostic learner for $\mathcal{F}_k$, we have that for sample size $n \geq \max_k m_k(\frac{\epsilon}{K}, \frac{\delta}{K})$, with probability at least $1 - \delta/K$ over samples $S_k \sim \mathcal{D}_k^n$,

$$\mathop{\mathbb{E}}_{\mathcal{D}_k}[\psi_k \circ d_1(g_k(x), y^k)] \leq \inf_{f_k \in \mathcal{F}_k} \mathop{\mathbb{E}}_{\mathcal{D}_k}[\psi_k \circ d_1(f_k(x), y^k)] + \frac{\epsilon}{K}.$$

Summing these risk bounds over all k coordinates and union bounding over the success probabilities, we get that with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^n$,

$$\sum_{k=1}^{K} \mathop{\mathbb{E}}_{\mathcal{D}_k}[\psi_k \circ d_1(g_k(x), y^k)] \leq \sum_{k=1}^{K} \inf_{f_k \in \mathcal{F}_k} \mathop{\mathbb{E}}_{\mathcal{D}_k}[\psi_k \circ d_1(f_k(x), y^k)] + \epsilon.$$

Using the fact that the sum of infimums over individual coordinates is at most the overall infimum of sums followed by the linearity of expectation, we can write the expression above as

$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\sum_{k=1}^{K}\psi_k \circ d_1(g_k(x), y^k)\right] \leq \inf_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{\mathcal{D}}\left[\sum_{k=1}^{K}\psi_k \circ d_1(f_k(x), y^k)\right] + \epsilon.$$

This shows that the learning rule that runs $\mathcal{A}_k$ on marginal samples S_k and concatenates the resulting scalar-valued predictors to get a vector-valued predictor is an agnostic learner for $\mathcal{F}$ with respect to loss ℓ with sample complexity at most $\max_k m_k(\epsilon/K, \delta/K)$. This completes our proof of sufficiency. ■

A.3.2 Equivalence of d_1 and $\psi \circ d_1$ Learnability in Batch Regression

In this section, we provide proof of Lemma 3, which establishes the equivalence of d_1 and $\psi \circ d_1$ learnability in scalar-valued batch regression.

Proof. (of Lemma 3) To prove sufficiency first, let $\mathcal{G}$ be agnostically learnable with respect to d_1. This implies that the fat-shattering dimension of $\mathcal{G}$ is finite at every scale [Bartlett et al., 1996] and uniform convergence holds over the loss class $d_1 \circ \mathcal{G}$. Since ψ is a Lipschitz function, a simple application of Talagrand's contraction lemma on Rademacher complexity [Bartlett and Mendelson, 2003] implies that uniform convergence holds over the loss class $\psi \circ d_1 \circ \mathcal{G}$ as well. Thus, $\mathcal{G}$ is learnable with respect to $\psi \circ d_1$ via ERM.

Next, we show that if $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ is learnable with respect to $\psi \circ d_1$, then $\mathcal{G}$ is learnable with respect to d_1. Since the fat-shattering dimension of $\mathcal{G}$ characterizes d_1 learnability of $\mathcal{G}$, it suffices to show that $\mathcal{G}$ being learnable with respect to $\psi \circ d_1$ implies $\mathrm{fat}_\gamma(\mathcal{G}) < \infty$ for every $\gamma \in (0,1)$.

Suppose, for the sake of contradiction, $\mathcal{G}$ is learnable with respect to $\psi \circ d_1$ but there exists a scale $\gamma \in (0,1)$ such that $\mathrm{fat}_\gamma(\mathcal{G}) = \infty$. Then, for every $d \in \mathbb{N}$, there exists $X = \{x_1, \ldots, x_d\} \subseteq \mathcal{X}$ and a witness function $r : \mathcal{X} \to [0,1]$ such that for every $\sigma \in \{-1,1\}^d$, there exists a $g_\sigma \in \mathcal{G}$ such that $\sigma_i(g_\sigma(x_i) - r(x_i)) \geq \gamma$ for all $i \in [d]$. Define $\mathcal{G}_X = \{g_\sigma \in \mathcal{G} \mid \sigma \in \{-1,1\}^d\}$ be the set of functions that shatters X. Define $\mathcal{H} = \{-1,1\}^X$ to be a set of all functions from X to $\{-1,1\}$. By definition of $\mathcal{H}$, we must have $\mathrm{VC}(\mathcal{H}) = d$. We use an agnostic learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ to construct an agnostic learner for $\mathcal{H}$ whose sample complexity, for large enough d, is smaller than the known lower bound for VC classes. Since $\mathrm{fat}_\gamma(\mathcal{G}) = \infty$, d can be made arbitrarily large and thus we derive a contradiction.

Let $\mathcal{A}$ be the promised agnostic learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ with sample complexity $m(\epsilon, \delta)$. For all $f \in [0,1]^X$, define a threshold function $h_f : X \to \{-1, 1\}$ as $h_f(x) = 2\,\mathbb{1}\{f(x) \geq r(x)\} - 1$. Let $\mathcal{D}$ be an arbitrary distribution on $X \times \{-1, 1\}$ and $\mathcal{D}_X$ be its marginal on X.

Algorithm 9 Agnostic PAC learner for $\mathcal{H}$

Require: Agnostic learner $\mathcal{A}$ for $\mathcal{G}$, unlabeled samples $S_U \sim \mathcal{D}_X$, and another independent labeled samples $S_L \sim \mathcal{D}$

1: Define $S_{\text{aug}} = \{(S_U, g^\alpha(S_U)) \mid g \in \mathcal{G}_X\}$, all possible augmentations of S_U by α-discretization of functions in $\mathcal{G}_X$ for $\alpha \leq \gamma/2$.

2: Run $\mathcal{A}$ over all possible augmentations to get

$$C(S_U) := \left\{\mathcal{A}\Big(S\Big) \mid S \in S_{\text{aug}}\right\}.$$

3: Define $C_{\pm 1}(S_U) = \{h_f \mid f \in C(S_U)\}$, a thresholded class of $C(S_U)$.

4: Return the predictor in $C_{\pm 1}(S_U)$ with the lowest empirical 0-1 risk over S_L.

We now show that Algorithm 9 is an agnostic learner for $\mathcal{H}$. Consider $d \gg S_U + S_L$. Then, $|C_{\pm 1}(S_U)| = |S_{\text{aug}}| \leq (2/\alpha)^{|S_U|}$ can be much smaller than 2^d. Let $h^\star := \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}]$ be the optimal hypothesis for $\mathcal{D}$. Note that, by definition of shattering, for every $h \in \mathcal{H}$, there exists a $g \in \mathcal{G}_X$ such that $h(x) = h_g(x)$ for all $x \in X$. In particular, there must exist $g^\star \in \mathcal{G}_X$ such that $h^\star(x) = h_{g^\star}(x) := 2\,\mathbb{1}\{g^\star(x) \geq r(x)\} - 1$ for all $x \in X$. Let g^α denote the α-discretization of g as defined in Equation (2.1) for some $\alpha \leq \gamma/2$. Now, consider a sample $(S_U, g^{\star,\alpha}(S_U)) \in S_{\text{aug}}$. Let $\hat{g} = \mathcal{A}((S_U, g^{\star,\alpha}(S_U)))$ be the predictor returned by the algorithm when run on a sample labeled by $g^{\star,\alpha}$. Define $h_{\hat{g}} = 2\,\mathbb{1}\{\hat{g}(x) \geq r(x)\} - 1$ to be its thresholded function. Then, using the triangle inequality on the indicator function, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x) \neq y\}] \leq \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h^\star(x) \neq y\}] + \mathop{\mathbb{E}}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}]. \tag{A.1}$$

Note that $\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\} = \mathbb{1}\{h_{\hat{g}}(x) \neq h_{g^\star}(x)\} \leq \mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\}$. To see why the last inequality is true, we only have to consider the case where the indicator on the left is 1, otherwise, the inequality is trivial. Recall that $\mathbb{1}\{h_{\hat{g}}(x) \neq h_{g^\star}(x)\} = 1$ whenever $\hat{g}(x)$ and $g^\star(x)$ lie on the opposite side of witness $r(x)$. Since $g^\star$ has to be at least γ away from the witness $r(x)$, we obtain $\mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\} = 1$. Next, using the fact that $\alpha \leq \gamma/2$, we have $\mathbb{1}\{|\hat{g}(x) - g^\star(x)| \geq \gamma\} \leq \mathbb{1}\{|\hat{g}(x) - g^{\star,\alpha}(x)| \geq \gamma/2\}$ because discretization can decrease

the distance between these functions by at most $\gamma/2$. Furthermore, using monotonicity of ψ, we get $\mathbb{1}\{|\hat{g}(x)-g^{\star,\alpha}(x)| \geq \gamma/2\} \leq \mathbb{1}\{\psi(|\hat{g}(x)-g^{\star,\alpha}(x)|) \geq \psi(\gamma/2)\}$. Combining everything, we get a pointwise inequality

$$\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\} \leq \mathbb{1}\{\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|) \geq \psi(\gamma/2)\} \leq \frac{1}{\psi(\gamma/2)}\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|).$$

Using this inequality gives an upperbound on the risk of $h_{\hat{g}}$, namely

$$\mathop{\mathbb{E}}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}] \leq \frac{1}{\psi(\gamma/2)} \mathop{\mathbb{E}}_{\mathcal{D}}[\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|)]. \tag{A.2}$$

Since $\hat{g} = \mathcal{A}((S_U, g^{\star,\alpha}(S_U)))$, we can use the algorithm's guarantee to get a further upperbound on the expectation above. In particular, if $|S_U| \geq m(\frac{\epsilon\psi(\gamma/2)}{4}, \delta/2)$, then with probability at least $1-\delta/2$ over sampling $S_U \sim \mathcal{D}_X$, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}_X}[\psi(|\hat{g}(x) - g^{\star,\alpha}(x)|)] \leq \inf_{g\in\mathcal{G}} \mathop{\mathbb{E}}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] + \frac{\epsilon\psi(\gamma/2)}{4}.$$

Note that $\inf_{g\in\mathcal{G}} \mathbb{E}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] \leq \mathbb{E}_{\mathcal{D}_X}[\psi(|g^\star(x) - g^{\star,\alpha}(x)|)] \leq \psi(\alpha)$, where the last step uses the fact that $|g^\star(x) - g^{\star,\alpha}(x)| \leq \alpha$ and ψ is monotonic. Using L-Lipschitzness of ψ and the fact that $\psi(0) = 0$, we get $\psi(\alpha) \leq L\alpha$. Picking $\alpha = \min(\gamma/2, \frac{\epsilon\psi(\gamma/2)}{4L})$, we get $\inf_{g\in\mathcal{G}} \mathbb{E}_{\mathcal{D}_X}[\psi(|g(x) - g^{\star,\alpha}(x)|)] \leq \frac{\epsilon\psi(\gamma/2)}{4}$. Plugging this back to the inequality in the display above, we get to $\mathbb{E}_{\mathcal{D}_X}[\psi(|\hat{g}(x)-g^{\star,\alpha}(x)|)] \leq \frac{\epsilon\psi(\gamma/2)}{2}$. Using this guarantee on (A.2), we obtain

$$\mathop{\mathbb{E}}_{\mathcal{D}_X}[\mathbb{1}\{h_{\hat{g}}(x) \neq h^\star(x)\}] \leq \frac{\epsilon}{2}.$$

This bound applied to (A.1) yields

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x) \neq y\}] \leq \inf_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h(x) \neq y\}] + \frac{\epsilon}{2}.$$

Thus, we have shown the existence of a predictor $h_{\hat{g}} \in C_{\pm 1}(S_U)$ that achieves agnostic PAC bounds for $\mathcal{H}$. Let $\hat{h}$ be the predictor returned by step 4 of the algorithm. Next, we show that for sufficiently large S_L, the predictor $\hat{h}$ also attains agnostic PAC bounds. Recall that by Hoeffding's Inequality and union bound, with probability at least $1-\delta/2$, the empirical risk of every hypothesis in $C_{\pm 1}(S_L)$ on a sample of size $\geq \frac{8}{\epsilon^2}\log\frac{4|C_{\pm 1}(S_U)|}{\delta}$ is at most $\epsilon/4$ away from its true error. So, if $|S_L| \geq \frac{8}{\epsilon^2}\log\frac{4|C_{\pm 1}(S_U)|}{\delta}$, then with probability at least $1-\delta/2$, the

empirical risk of the predictor $h_{\hat{g}}(x)$ is

$$\frac{1}{|S_L|}\sum_{(x,y)\in S_L}\mathbb{1}\{h_{\hat{g}}(x)\neq y\}\leq \mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h_{\hat{g}}(x)\neq y\}]+\frac{\epsilon}{4}\leq \inf_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h(x)\neq y\}]+\frac{3\epsilon}{4},$$

where the last inequality follows from the risk guarantee of $h_{\hat{g}}$ established above. Since $\hat{h}$ is the empirical risk minimizer over S_L, we must have

$$\frac{1}{|S_L|}\sum_{(x,y)\in S_L}\mathbb{1}\{\hat{h}(x)\neq y\}\leq \frac{1}{|S_L|}\sum_{(x,y)\in S_L}\mathbb{1}\{h_{\hat{g}}(x)\neq y\}\leq \inf_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h(x)\neq y\}]+\frac{3\epsilon}{4}.$$

Finally, as the population risk of $\hat{h}$ is at most $\epsilon/4$ away from its empirical risk, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{\hat{h}(x)\neq y\}]\leq \inf_{h\in\mathcal{H}}\mathop{\mathbb{E}}_{\mathcal{D}}[\mathbb{1}\{h(x)\neq y\}]+\epsilon,$$

which is the agnostic PAC guarantee for $\mathcal{H}$. Applying union bounds, the entire process, running algorithm $\mathcal{A}$ on the dataset augmented by $g^{\star,\alpha}$ and the ERM in step 4, succeeds with probability $1-\delta$. This establishes that the Algorithm 9 is an agnostic PAC learner for $\mathcal{H}$. The sample complexity of Algorithm 9 is the number of samples required for Algorithm $\mathcal{A}$ to succeed and the ERM in step 4 to succeed. Thus, the overall sample complexity of Algorithm 9, denoted $m_{\mathcal{H}}(\epsilon,\delta)$, can be bounded as

$$\begin{aligned}
m_{\mathcal{H}}(\epsilon,\delta) &\leq m_{\mathcal{A}}\left(\frac{\epsilon\,\psi(\gamma/2)}{4},\frac{\delta}{2}\right)+\frac{8}{\epsilon^2}\log\frac{4|C_{\pm 1}(S_U)|}{\delta}\\
&\leq m_{\mathcal{A}}\left(\frac{\epsilon\,\psi(\gamma/2)}{4},\frac{\delta}{2}\right)\left(1+\frac{8}{\epsilon^2}\log\left(\frac{2}{\min(\gamma/2,\epsilon\psi(\gamma/2)/4)}\right)\right)+\frac{8}{\epsilon^2}\log\frac{4}{\delta}
\end{aligned}$$

where the second inequality follows because $|C_{\pm 1}(S_U)|=|S_{\text{aug}}|\leq(2/\alpha)^{|S_U|}$ and we need $|S_U|$ to be of size $m_{\mathcal{A}}\left(\frac{\epsilon\,\psi(\gamma/2)}{4},\frac{\delta}{2}\right)$. We also use the fact that $\alpha=\min(\frac{\gamma}{2},\frac{\epsilon\psi(\frac{\gamma}{2})}{4})$.

However, it is well known [Shalev-Shwartz and Ben-David, 2014, Theorem 6.8] that the sample complexity of learning $\mathcal{H}$ in agnostic setting is

$$C\frac{d+\log(2/\delta)}{\epsilon^2}$$

for some $C>0$. Thus, we must have $m_{\mathcal{H}}(\epsilon,\delta)\geq C(d+\log(2/\delta))/\epsilon^2$. However, this is a contradiction because d can be arbitrarily large but $m_{\mathcal{H}}(\epsilon,\delta)$ must have a finite upper bound for every fixed ϵ,δ. Therefore, the function class $\mathcal{G}$ cannot be learnable with respect to $\psi\circ d_1$ whenever there exists a scale $\gamma\in(0,1)$ such that $\text{fat}_\gamma(\mathcal{G})=\infty$. ■

A.4 Rademacher Based Proof for Batch Regression

To show that the learnability of each $\mathcal{F}_k$ with respect to d_1 is sufficient for the learnability of $\mathcal{F}$ with respect to ℓ_p norms for $p \geq 1$, we use the fact that $\ell_p(f(x), y)$ is a K-Lipschitz in its first argument with respect to $\|\cdot\|_\infty$ norm, that is $|\ell_p(f(x), y) - \ell_p(g(x), y)| \leq K \|f(x) - g(x)\|_\infty$, and use the following bound on the Rademacher complexity of the loss class $\ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(f(x), y)\} \mid f \in \mathcal{F}\}$.

Lemma 11 (Foster and Rakhlin [2019]). *Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multioutput function class. For any $\delta \in (0, 1)$, there exists a constants $0 < c < 1$ and $C > 0$ such that*

$$\mathfrak{R}_n(\ell_p \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{C}{\sqrt{n}} \sum_{k=1}^{K} \int_\alpha^1 \sqrt{fat_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} d\epsilon \right\}.$$

The result presented here is in fact the intermediate result in Foster and Rakhlin [2019], and we provide a sketch of how their argument can be adapted to our setting.

Proof. Note that for $f, g \in \mathcal{F}$, we have $|\ell_p(f(x), y) - \ell_p(g(x), y)| \leq |\ell_p(f(x), g(x))| \leq \ell_1(f(x), g(x)) \leq K \|f(x) - g(x)\|_\infty$. Furthermore, we have that $|f_k(x) - y^k| \leq 1$, so we obtain $|\ell_p(f(x), y)| \leq K$. Define the normalized ℓ_p loss as $\bar{\ell}_p(f(x), y) := \ell_p(f(x), y)/K$. By standard chaining argument, we know that

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = K\, \mathfrak{R}_n(\bar{\ell} \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^1 \sqrt{\log \mathcal{N}_2(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n)} d\epsilon \right\}.$$

Since a cover with $||\cdot||_\infty$ norm is also a cover with respect to $||\cdot||_2$ norm, we have that $\log \mathcal{N}_2(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n) \leq \log \mathcal{N}_\infty(\bar{\ell} \circ \mathcal{F}, \epsilon, n)$. Since $\bar{\ell}_p(f(x), y)$ is $1-$Lipschitz with respect to $||\cdot||_\infty$ norm, following Lemma 1 of Foster and Rakhlin [2019], we obtain $\log \mathcal{N}_\infty(\bar{\ell}_p \circ \mathcal{F}, \epsilon, n) \leq \sum_{k=1}^K \log \mathcal{N}_\infty(\mathcal{F}_k, \epsilon, n)$.

A result due to Rudelson and Vershynin [2006] states that for any $\delta \in (0, 1)$, there exists constants $0 < c_k < 1$ and $C_k > 0$ such that

$$\log \mathcal{N}_\infty(\mathcal{F}_k, \epsilon, n) \leq C_k \operatorname{fat}_{c_k \epsilon}(\mathcal{F}_k)\, \log^{1+\delta}(en/\epsilon).$$

Picking $C = \max_k C_k$ and $c = \min_k c_k$, we obtain the contraction inequality

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12\, C}{\sqrt{n}} \int_\alpha^1 \sqrt{\sum_{k=1}^{K} \operatorname{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} d\epsilon \right\}.$$

Using $\sqrt{\sum_{k=1}^K \operatorname{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)} \leq \sum_{k=1}^K \sqrt{\operatorname{fat}_{c\epsilon}(\mathcal{F}_k) \log^{1+\delta}\left(\frac{e\, n}{\epsilon}\right)}$ yields the desired con-

traction inequality.

■

With Lemma 11 in our repertoire, the sufficiency proof is a routine uniform convergence argument.

Proof. (of sufficiency in Theorem 5) Suppose each restriction $\mathcal{F}_k$ is learnable with respect to d_1. Then, we know that for all $k \in [K]$ and for all $1 > \gamma > 0$, we have $\text{fat}_\gamma(\mathcal{F}_k) < \infty$ [Bartlett et al., 1996], [Anthony and Bartlett, 1999, Chatper 19]). Using Lemma 11, for $\delta = 1/2$, we can find constants c, C such that

$$\mathfrak{R}_n(\ell_p \circ \mathcal{F}) \leq K \inf_{\alpha > 0} \left\{ 4\alpha + \frac{C}{\sqrt{n}} \sum_{k=1}^{K} \int_\alpha^1 \sqrt{\text{fat}_{c\epsilon}(\mathcal{F}_k) \log^{3/2}\left(\frac{e\,n}{\epsilon}\right)} d\epsilon \right\}.$$

Fix $\alpha > 0$. The second term inside infimum vanishes as $n \to \infty$, yielding $\mathfrak{R}_n(\ell \circ \mathcal{F}) \leq 4\alpha K$. As $\alpha > 0$ is arbitrary, the Rademacher complexity $\mathfrak{R}_n(\ell \circ \mathcal{F})$ goes to 0 as $n \to \infty$. This argument can be readily turned into non-asymptotic bounds on $\mathfrak{R}_n(\ell_p \circ \mathcal{F})$ if the precise form of $\text{fat}_\gamma(\mathcal{F}_k)$ as a function of γ is known. Since the empirical Rademacher complexity vanishes, uniform convergence holds over the loss class $\ell_p \circ \mathcal{F}$ and thus $\mathcal{F}$ is learnable with respect to ℓ_p via empirical risk minimization. ■

A.5 Online Multilabel Learnability with respect to Hamming Loss

In this section, we provide the proof of Theorem 7.

Proof. We first prove that the online learnability of each restriction is sufficient for the online learnability of ℓ_H.

Part 1: Sufficiency. Our proof is based on a reduction: given oracle access to online learners $\{\mathcal{A}_k\}_{k=1}^K$ for $\{\mathcal{F}_k\}_{k=1}^K$ with respect to $\ell_{0\text{-}1}$, we construct an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H. In fact, similar to the batch setting, the online multilabel learning algorithm $\mathcal{A}$ is simple: in each round $t \in [T]$, receive x_t, query the predictions $\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t)$, and finally predict the concatenation $\hat{y}_t = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. Once the true label $y_t = (y_t^1, ..., y_t^K)$ is revealed, update each online learner $\mathcal{A}_k$ by passing (x_t, y_t^k) for $k \in [K]$. It suffices to show that the expected regret of $\mathcal{A}$ is sublinear in T with respect to ℓ_H. By Definition 4, we have that for all $k \in [K]$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\} - \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^{T} \mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq R_k(T)$$

where $R_k(T)$ is some sublinear function in T. Summing the regret bounds across all $k \in [K]$ splitting up the expectations, and using linearity of expectation, we get that $\mathbb{E}\left[\sum_{k=1}^K \sum_{t=1}^T \mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] - \mathbb{E}\left[\sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq \sum_{k=1}^K R_k(T)$. Noting that $\sum_{k=1}^K \inf_{f_k \in \mathcal{F}_k} \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\} \leq \inf_{f \in \mathcal{F}} \sum_{k=1}^K \sum_{t=1}^T \mathbb{1}\{f_k(x_t) \neq y_t^k\}$, swapping the order of summations, and using the definition of ℓ_H we have that,

$$\mathbb{E}\left[\sum_{t=1}^T \ell_H(\mathcal{A}(x_t), y_t)\right] - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell_H(f(x_t), y_t)\right] \leq \sum_{k=1}^K R_k(T),$$

where $\mathcal{A}(x_t) = (\mathcal{A}_1(x_t), ..., \mathcal{A}_K(x_t))$. This concludes the proof of this direction since $\sum_{k=1}^K R_k(T)$ is still a sublinear function in T.

Part 2: Necessity. Next we prove that if $\mathcal{F}$ is online learnable with respect to ℓ_H, then each $\mathcal{F}_k$ is online learnable with respect to $\ell_{0\text{-}1}$. Namely, given oracle access to an online learner $\mathcal{A}$ for $\mathcal{F}$ with respect to ℓ_H, we construct an online learner $\mathcal{B}$ for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. A similar reduction can be used to construct online learners for each restriction $\mathcal{F}_k$. Similar to the batch setting, the online learning algorithm $\mathcal{B}$ is simple: in each round $t \in [T]$, receive x_t, query $\hat{y}_t = \mathcal{A}(x_t)$ and predict $\hat{y}_t^1 = \mathcal{A}_1(x_t)$. Once the true label y_t^1 is revealed, update $\mathcal{A}$ by passing (x_t, y_t) where $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$ and $\{\sigma_t^i\}_{i=2}^K$ is an i.i.d sequence of Rademacher random variables.

It suffices to show that the expected regret of $\mathcal{B}$ is sublinear in T with respect to $\ell_{0\text{-}1}$. As previously mentioned, we assume that the sequence $(x_1, y_1^1), ..., (x_T, y_T^1)$ is chosen by an oblivious adversary, and thus is not random. Let $y_t = (y_t^1, \sigma_t^2, ..., \sigma_t^K)$. By Definition 4, we have that,

$$\mathbb{E}\left[\sum_{t=1}^T \ell_H(\mathcal{A}(x_t), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell_H(f(x_t), y_t)\right] \leq R(T, 2^K)$$

where the expectation is over both the randomness of $\mathcal{A}(x_t)$ and $(\sigma_t^2, ..., \sigma_t^K)$ and $R(T, 2^K)$ is a sub-linear function of T. Splitting up the expectation, using the definition of the Hamming loss, and by the linearity of expectation, we have that

$$\sum_{t=1}^T \sum_{k=1}^K \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \sum_{k=1}^K \mathbb{E}\left[\mathbb{1}\{f_k(x_t) \neq y_t^k\}\right] \leq R(T, 2^K).$$

Next, observe that for every $t \in [T]$, for every $k \in \{2, ..., K\}$, the randomness of $y_t^k = \sigma_t^k$ implies $\mathbb{E}\left[\mathbb{1}\{\mathcal{A}_k(x_t) \neq y_t^k\}\right] = \mathbb{E}\left[\mathbb{1}\{f(x_t) \neq y_t^k\}\right] = \frac{1}{2}$. Thus,

$$\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}\{\mathcal{A}_1(x_t) \neq y_t^1\} + \frac{K-1}{2}\right] - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}\{f_1(x_t) \neq y_t^1\} + \frac{K-1}{2}\right] \leq R(T, 2^K).$$

Canceling constant factors gives, $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{A}_1(x_t) \neq y_t^1\}\right] - \inf_{f_1 \in \mathcal{F}_1} \sum_{t=1}^{T} \mathbb{1}\{f_1(x_t) \neq y_t^1\} \leq R(T, 2^K)$, showing that $\mathcal{B}$ is an online agnostic learner for $\mathcal{F}_1$ with respect to $\ell_{0\text{-}1}$. ■

A.6 MCLdim Characterizes Online Multilabel Learnability

In this section, we show that the MCLdim characterizes the online learnability of a multilabel function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with respect to any loss ℓ that satisfies the identity of indiscernibles. Theorem 37 makes this more precise.

Theorem 37. *Let ℓ be any loss function satisfying the identity of indiscernibles. A function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ is online learnable with respect to ℓ if and only if MCLdim$(\mathcal{F}) < \infty$.*

Proof. (of sufficiency) We first show that finiteness of MCLdim is sufficient for online learnability. The proof follows exactly like the proof of Lemma 5. We include it here again for completeness sake. Let ℓ be any loss function satisfying the identity of indiscernibles and $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a multilabel function class such that MCLdim$(\mathcal{F}) = d < \infty$. Since $\mathcal{F}$ has finite MCLdim, the deterministic Multiclass Standard Optimal Algorithm for $\mathcal{F}$, hereinafter denoted MCSOA$(\mathcal{F})$, achieves mistake-bound d in the realizable setting [Daniely et al., 2011]. Therefore, following the same procedure as in Daniely et al. [2011], we can construct a finite set of experts $\mathcal{E}$ of size $|\mathcal{E}| = \sum_{j=0}^{d} \binom{T}{j} |\text{im}(\mathcal{F})|^j \leq (2^K T)^d$ such that for any (oblivious) sequence of instances $x_1, ..., x_T$, for any function $f \in \mathcal{F}$, there exists an expert $E_f \in \mathcal{E}$, such that $f(x_t) = E(x_t)$ for all $t \in [T]$. Finally, running the celebrated Randomized Exponential Weights Algorithm (REWA) using $\mathcal{E}$ as the set of experts and the scaled loss function $\frac{\ell}{B} \in [0, 1]$ guarantees that for any labelled sequence $(x_1, y_1), ..., (x_T, y_T)$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \ell(E(x_t), y_t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \\ &\leq O\left(B\sqrt{T \ln(|\mathcal{E}|)}\right) \leq O\left(B\sqrt{dTK \ln(T)}\right)\end{aligned}$$

where $\hat{y}_t$ is the prediction of REWA in the t'th round. Thus, running REWA over the set of experts $\mathcal{E}$ using $\frac{\ell}{B}$ gives an online learner for $\mathcal{F}$ with respect to ℓ. ■

Proof. (of necessity) To prove necessity, we need to show that if $\mathcal{F}$ is online learnable with respect to ℓ, then MCLdim$(\mathcal{F}) < \infty$. To do so, we show that if $\mathcal{F}$ is online learnable with respect to ℓ, then $\mathcal{F}$ is online learnable with respect to $\ell_{0\text{-}1}$ in the realizable setting. Let $\mathcal{A}$

be an online learner for $\mathcal{F}$ with respect to ℓ. Then, by definition,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{A}(x_t),y_t)-\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)\right]\leq R(T,2^K)$$

where $R(T,2^K)$ is a sublinear function in T. Since ℓ satisfies the identity of indiscernibles, in the realizable setting, $\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)=0$. Therefore, under realizability,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{A}(x_t),y_t)\right]\leq R(T,2^K).$$

Because there are only a finite number of inputs to ℓ, there must exist a universal constant a such that $a\ell_{0\text{-}1}\leq\ell$. Substituting in gives that,

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{0\text{-}1}(\mathcal{A}(x_t),y_t)\right]\leq\frac{R(T,2^K)}{a}.$$

Since a is a universal constant that does not depend on T, $\frac{R(T,2^K)}{a}$ is still a sublinear function in T, implying that $\mathcal{A}$ is also a realizable online learner for $\mathcal{F}$ with respect to $\ell_{0\text{-}1}$. This completes the proof as MCLdim characterizes realizable learnability and so we must have $\text{MCLdim}(\mathcal{F})<\infty$. ■

A.7 Proofs for Bandit Online Multilabel Classification

In this section, we provide proofs for the characterization of online multilabel classification under bandit feedback.

Proof. (of Theorem 10) The proof of Theorem 10 is nearly identical to the proof of Theorem 6. The only difference is that in Algorithm 4, we now need use the bandit Expert's algorithm EXP4 from Auer et al. [2002] instead of REWA. Similar to REWA, based on Theorem 2.3 in Daniely and Helbertal [2013] and the fact that A, B and P are independent, EXP4 guarantees that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell(\mathcal{P}(x_t),y_t)\right]\leq\mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\ell(E(x_t),y_t)\right]+eM\mathbb{E}\left[\sqrt{2T|\mathcal{Y}|\ln(|\mathcal{E}_B|)}\right],$$

where $\mathcal{P}(x_t)$ denotes EXP4's prediction in round t. The remaining proof for deriving the upper bound

$$\mathbb{E}\left[\inf_{E\in\mathcal{E}_B}\sum_{t=1}^{T}\ell(E(x_t),y_t)\right]\leq\inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\ell(f(x_t),y_t)+\frac{cT}{T^\beta}\overline{R}(T^\beta,|\mathcal{Y}|)$$

is identical to that in Theorem 6, so we omit it here. Putting these pieces together gives the stated guarantee. ■

Proof. (of Theorem 9) Let $c = \frac{\max_{r \neq t} \ell(r,t)}{\min_{r \neq t} \ell(r,t)}$. We first show necessity: if $\mathcal{F}$ is bandit online learnable with respect to ℓ, then each restriction $\mathcal{F}_k$ is online learnable with respect to $\ell_{0\text{-}1}$. This follows trivially from the fact that if $\mathcal{A}$ is a bandit online learner for $\mathcal{F}$, then $\mathcal{A}$ is also an online learner for $\mathcal{F}$ under full-feedback. Thus, by Theorem 8, online learnability of $\mathcal{F}$ with respect to ℓ implies online learnability of restriction $\mathcal{F}_k$ with respect to the 0-1 loss.

We now focus on showing sufficiency: if for all $k \in [K]$, $\mathcal{F}_k$ is online learnable with respect to 0-1 loss, then $\mathcal{F}$ is *bandit* online learnable with respect to loss ℓ. Since $|\mathcal{Y}| = 2^K < \infty$ and ℓ is a c-subadditive, by Theorem 10, it suffices to show that there exists a realizable online learner for $\mathcal{F}$ with respect to ℓ. However, using Theorem 8, online learnability of each restriction $\mathcal{F}_k$ with respect to 0-1 the loss implies (agnostic) online learnability of $\mathcal{F}$ with respect to ℓ. Since an agnostic online learner is trivially a realizable online learner, the proof is complete. ■

A.8 Equivalence of d_1 and $\psi \circ d_1$ Online Learnability

Proof. (of Lemma 6) Since ψ is a Lipschitz function, the proof of sufficiency follows immediately from Corollary 5 in Rakhlin et al. [2015a], a contraction Lemma for the sequential Rademacher complexity. Thus, we focus on proving necessity - if $\mathcal{G}$ is online learnable with respect to $\psi \circ d_1$, then $\mathcal{G}$ is online learnable with respect to d_1.

Since the sequential fat shattering dimension of $\mathcal{G}$ characterizes d_1 learnability [Rakhlin et al., 2015a], it suffices to show that $\mathcal{G}$ being online learnable with respect to $\psi \circ d_1$ implies $\text{fat}^{\text{seq}}_\gamma(\mathcal{G}) < 0$ for every $\gamma \in (0,1)$. Like in the batch setting, we prove this via contradiction.

Suppose, for the sake of contradiction, $\mathcal{G}$ is online learnable with respect to $\psi \circ d_1$ but there exists a scale $\gamma \in (0,1)$ such that $\text{fat}^{\text{seq}}_\gamma(\mathcal{G}) = \infty$. Then, for every $T \in \mathbb{N}$, there exists a $\mathcal{X}$-valued binary tree $\mathcal{T}$ and a $[0,1]$-valued binary witness tree $\mathcal{R}$ both of depth T such that for all $\sigma \in \{-1,1\}^T$, there exists $g_\sigma \in \mathcal{G}$ such that for all $t \in [T]$, $\sigma_t(g_\sigma(\mathcal{T}_t(\sigma_{<t})) - \mathcal{R}_t(\sigma_{<t})) \geq \gamma$. Without loss of generality, assume that for any path $\sigma \in \{-1,1\}^T$, the set of instances $\{\mathcal{T}_t(\sigma_{<t})\}_{t=1}^T$ are distinct. This is true because we can construct a γ-shattered tree of much bigger depth and prune away repeated instances along a path to get a tree of depth T. Define $\mathcal{G}_T = \{g_\sigma \in \mathcal{G} | \sigma \in \{-1,1\}^T\}$ to be the set of functions that shatter $\mathcal{T}$ with witness $\mathcal{R}$. Let $X \subseteq \mathcal{X}$ denote the set of examples that label the internal nodes of $\mathcal{T}$. Consider the binary hypothesis class $\mathcal{H} = \{-1,1\}^X$ which contains all possible functions from X to $\{-1,1\}$. By definition of $\mathcal{H}$, we must have $\text{Ldim}(\mathcal{H}) \geq T$. Therefore, $\mathcal{T}$, with left and right edges labeled

by -1 and $+1$ respectively, is shattered by $\mathcal{H}$. Let $\mathcal{T}_{\pm}$ denote such a tree. Note that for all $t \in [T]$, we have $\mathcal{T}_t(\sigma_{<t}) = \mathcal{T}_{\pm,t}(\sigma_{<t})$. Since $\mathrm{Ldim}(\mathcal{H}) \geq T$, any realizable online learner for $\mathcal{H}$ must make at least $\frac{T}{2}$ mistakes in expectation for an adversary that plays according to a root-to-leaf path in $\mathcal{T}_{\pm}$ chosen *uniformly at random*. However, using an agnostic online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$, we construct a realizable online learner for $\mathcal{H}$ that achieves a *sublinear* regret bound when an adversary plays according to a root-to-leaf path in $\mathcal{T}_{\pm}$ chosen uniformly at random. Since $\mathrm{fat}^{\mathrm{seq}}_{\gamma}(\mathcal{G}) = \infty$, T can be made arbitrarily large, eventually giving us a contradiction.

To that end, let $\mathcal{A}$ be an online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$ with regret $R_{\mathcal{A}}(T)$. Let $\sigma \sim \{-1, 1\}^T$ denote a sequence of T i.i.d Rademacher random variables and $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$ the associated sequence of labeled instances determined by traversing $\mathcal{T}_{\pm}$ using σ. Note that $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$ is a sequence of labeled instances corresponding to a root-to-leaf path in $\mathcal{T}_{\pm}$ chosen *uniformly at random*. By construction of $\mathcal{H}$, there exists a $h^\star_\sigma \in \mathcal{H}$ such that $h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})) = \sigma_t$ for all $t \in [T]$ and therefore the stream is realizable by $\mathcal{H}$. Let $g^\star_\sigma \in \mathcal{G}$ be the function at the end of the root-to-leaf path corresponding to σ in $\mathcal{T}$, the original tree shattered by $\mathcal{G}$.

We now use $\mathcal{A}$ to construct an agnostic online learner for $\mathcal{H}$ with sublinear regret on the stream $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^T$. Our algorithm is very similar to realizable-to-agnostic conversion in Theorem 6. Namely, we construct a finite set of experts, each of which uses $\mathcal{A}$ to make predictions, but only updates $\mathcal{A}$ on certain rounds. Finally, we run REWA using this set of Experts over our stream. For completeness' sake, we provide the full description below.

For any bitstring $b \in \{0,1\}^T$, let $\phi : \{t : b_t = 1\} \to \mathrm{im}(\mathcal{G}^\alpha)$ denote a function mapping time points where $b_t = 1$ to elements in the discretized image space $\mathrm{im}(\mathcal{G}^\alpha)$. Let $\Phi_b : (\mathrm{im}(\mathcal{G}^\alpha))^{\{t:b_t=1\}}$ denote all such functions ϕ. For every $g \in \mathcal{G}$, let $\phi^g_b \in \Phi_b$ be the mapping such that for all $t \in \{t : b_t = 1\}$, $\phi^g_b(t) = g^\alpha(\mathcal{T}_t(\sigma_{<t}))$. Let $|b| = |\{t : b_t = 1\}|$. For every $b \in \{0,1\}^T$ and $\phi \in \Phi_b$, define an Expert $E_{b,\phi}$. Expert $E_{b,\phi}$, formally presented in Algorithm 10, uses $\mathcal{A}$ to make predictions in each round. However, $E_{b,\phi}$ only updates $\mathcal{A}$ on those rounds where $b_t = 1$, using ϕ to produce a labeled instance $(\mathcal{T}_t(\sigma_{<t}), \phi(t))$. For every $b \in \{0,1\}^T$, let $\mathcal{E}_b = \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$ denote the set of all Experts parameterized by functions $\phi \in \Phi_b$. If b is the all zeros bitstring, then $\mathcal{E}_b$ is empty. Therefore, we actually define $\mathcal{E}_b = \{E_0\} \cup \bigcup_{\phi \in \Phi_b} \{E_{b,\phi}\}$, where E_0 is the expert that never updates $\mathcal{A}$. Note that $1 \leq |\mathcal{E}_b| \leq (\frac{2}{\alpha})^{|b|}$.

With this notation in hand, we are now ready to present Algorithm 11, our main online learner $\mathcal{Q}$ for $\mathcal{H}$ with respect to 0-1 loss. The analysis is similar to the one before, but we include it below for completeness sake.

Our goal now is to show that $\mathcal{Q}$ enjoys sublinear expected regret. There are three main sources of randomness: the randomness involved in sampling B, the internal randomness

Algorithm 10 Expert(b, ϕ)

Require: Independent copy of Online Learner $\mathcal{A}$ for $\psi \circ d_1$

for $t = 1, \ldots, T$ **do**
  Receive example $\mathcal{T}_{\pm,t}(\sigma_{<t})$
  Predict $\hat{y}_t = 2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \geq \mathcal{R}_t(\sigma_{<t})\} - 1$
  Receive $y_t = \sigma_t$
  if $b_t = 1$ **then**
    Update $\mathcal{A}$ by passing $(\mathcal{T}_{\pm,t}(\sigma_{<t}), \phi(t))$
  end if
end for

Algorithm 11 Online learner $\mathcal{Q}$ for $\mathcal{H}$ with respect to 0-1 loss

Require: Parameters $0 < \beta < 1$ and $0 < \alpha < \frac{\gamma}{2}$

Let $B \in \{0,1\}^T$ such that $B_t \overset{\text{iid}}{\sim} \text{Bernoulli}\left(\frac{T^\beta}{T}\right)$
Construct the set of experts $\mathcal{E}_B = \{E_0\} \cup \bigcup_{\phi \in \Phi_B} \{E_{B,\phi}\}$ according to Algorithm 10.
Run REWA $\mathcal{P}$ using $\mathcal{E}_B$ and the 0-1 loss over the stream

$$(\mathcal{T}_{\pm,1}(\sigma_{<1}), \sigma_1), \ldots, (\mathcal{T}_{\pm,T}(\sigma_{<T}), \sigma_T).$$

of each independent copy of the online learner $\mathcal{A}$, and the internal randomness of REWA. Let B, A and P denote the random variable associated with these sources of randomness respectively. By construction, A, B, and P are independent.

Using Theorem 21.11 in Shalev-Shwartz and Ben-David [2014] and the fact that A, B and P, are independent, REWA guarantees,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \mathbb{1}\{E(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \\
&\leq \mathbb{E}\left[\inf_{E \in \mathcal{E}_B} \sum_{t=1}^{T} \mathbb{1}\{E(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] + \mathbb{E}\left[\sqrt{2T \ln(|\mathcal{E}_B|)}\right].
\end{aligned}$$

In the last step, we used the fact that for all $b \in \{0,1\}^T$ and $g \in \mathcal{G}$, $E_{b,\phi_b^g} \in \mathcal{E}_b$.

It now suffices to upperbound $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right]$. We use the same nota-

tion used to prove Theorem 6, but for the sake of completeness, we restate it here. Given an online learner $\mathcal{A}$ for $\psi \circ d_1$, an instance $x \in \mathcal{X}$, and an ordered sequence of labeled examples $L \in (\mathcal{X} \times [0,1])^*$, let $\mathcal{A}(x|L)$ be the random variable denoting the prediction of $\mathcal{A}$ on the instance x after running and updating on L. For any $b \in \{0,1\}^T$, $g^\alpha \in \mathcal{G}^\alpha$, and $t \in [T]$, let $L^g_{b_{<t}} = \{(\mathcal{T}_{\pm,t}(\sigma_{<i}), g^\alpha(\mathcal{T}_{\pm,t}(\sigma_{<i}))) : i < t \text{ and } b_i = 1\}$ denote the *subsequence* of the sequence of labeled instances $\{(\mathcal{T}_{\pm,t}(\sigma_{<i}), g^\alpha(\mathcal{T}_{\pm,t}(\sigma_{<i})))\}_{i=1}^{t-1}$ where $b_i = 1$. Using this notation, we can write

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{E_{B,\phi_B^{g^\star_\sigma}}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big\{2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq \sigma_t\Big\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big\{2\,\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))\Big\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{|\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}) - g^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))| \geq \gamma\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{|\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}) - g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))| \geq \frac{\gamma}{2}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))) \geq \psi(\frac{\gamma}{2})\right\}\right] \\
&\leq \frac{1}{\psi(\frac{\gamma}{2})}\mathbb{E}\left[\sum_{t=1}^{T} \psi \circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}), g^{\star,\alpha}_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]
\end{aligned}
$$

The first inequality follows from γ-shattering. Indeed, if $2\mathbb{1}\{\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}}) \geq \mathcal{R}_t(\sigma_{<t})\} - 1 \neq h^\star_\sigma(\mathcal{T}_{\pm,t}(\sigma_{<t}))$, then $\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L^{g^\star_\sigma}_{B_{<t}})$ and $g^\star_\sigma$ must lie on opposite sides of the witness $\mathcal{R}_t(\sigma_{<t})$. The second inequality stems from the choice of $\alpha < \frac{\gamma}{2}$. The third inequality follows from the monotonicity of ψ. The last inequality follows from Markov's. Now, we can continue

like before.

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]$$
$$=\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\frac{\mathbb{P}\left[B_t=1\right]}{\mathbb{P}\left[B_t=1\right]}\right]$$
$$=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t=1\}\right]$$

To see the last equality, note that the prediction $\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g^\star})$ only depends on bitstring $(B_1,\ldots,B_{t-1})$, the string $(\sigma_1,...,\sigma_{t-1})$, and the internal randomness of A, all of which are independent of B_t. Thus, we have

$$\mathbb{E}\left[\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t=1\}\right]$$
$$=\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{g_\sigma^\star}),y_t)\right]\mathbb{E}\left[\mathbb{1}\{B_t=1\}\right]$$
$$=\mathbb{E}\left[\psi_1\circ d_1(\mathcal{A}_1(x_t|L_{B_{<t}}^{g_\sigma^\star}),y_t)\right]\mathbb{P}[B_t=1]$$

as needed. Continuing onwards,

$$\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g_\sigma^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]$$
$$=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t=1}^{T}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g_\sigma^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\mathbb{1}\{B_t=1\}\right]$$
$$=\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g_\sigma^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]$$
$$=\frac{T}{T^\beta}\mathbb{E}\left[\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(\mathcal{A}(\mathcal{T}_{\pm,t}(\sigma_{<t})|L_{B_{<t}}^{g_\sigma^\star}),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\middle|B\right]\right]$$
$$\leq\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(g_\sigma^\star(\mathcal{T}_{\pm,t}(\sigma_{<t})),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))+R_{\mathcal{A}}(|B|)\right]$$

The last inequality follows from the fact that $\mathcal{A}$ is an online learner for $\psi\circ d_1$ with regret bound $R_{\mathcal{A}}(T)$ and is updated using a stream labeled by $g^{\star,\alpha}$ only when $B_t=1$. Now, we

can upperbound:

$$\begin{aligned}
&\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi\circ d_1(g^\star(\mathcal{T}_{\pm,t}(\sigma_{<t})),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t})))\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\
&\qquad\leq\frac{T}{T^\beta}\mathbb{E}\left[\sum_{t:B_t=1}\psi(\alpha)\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\
&\qquad\leq\frac{T}{T^\beta}\mathbb{E}\left[\alpha L|B|\right]+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]\\
&\qquad=\alpha LT+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]
\end{aligned}$$

The first two inequalities follow from the fact that ψ is monotonic, L-Lipschitz, $\psi(0)=0$, and $d_1(g^\star(\mathcal{T}_{\pm,t}(\sigma_{<t})),g_\sigma^{\star,\alpha}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\leq\alpha$. Putting things together, we find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\neq\sigma_t\}\right]&\leq\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{E_{B,\phi_B^{g_\sigma^\star}}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\neq\sigma_t\}\right]+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]\\
&\leq\frac{\alpha LT+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})}+\mathbb{E}\left[\sqrt{2T\ln(|\mathcal{E}_B|)}\right]\\
&\leq\frac{\alpha LT+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})}+\mathbb{E}\left[\sqrt{2T|B|\ln(\frac{2}{\alpha})}\right].
\end{aligned}$$

where the last inequality follows from the fact that that $|\mathcal{E}_B|\leq(\frac{2}{\alpha})^{|B|}$. By Jensen's inequality, we further get that, $\mathbb{E}\left[\sqrt{2T|B|\ln(\frac{2}{\alpha})}\right]\leq\sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}$, which implies that

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\neq\sigma_t\}\right]\leq\frac{\alpha LT+\frac{T}{T^\beta}\mathbb{E}\left[R_{\mathcal{A}}(|B|)\right]}{\psi(\frac{\gamma}{2})}+\sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}.$$

Next, by Lemma 4, there exists a concave sublinear function $\overline{R}_{\mathcal{A}}(|B|)$ that upperbounds $R_{\mathcal{A}}(|B|)$. By Jensen's inequality, we obtain $\mathbb{E}[\overline{R}_{\mathcal{A}}(|B|)]\leq\overline{R}_{\mathcal{A}}(T^\beta)$, which yields

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\neq\sigma_t\}\right]\leq\frac{\alpha LT+\frac{T}{T^\beta}\overline{R}_{\mathcal{A}}(T^\beta)}{\psi(\frac{\gamma}{2})}+\sqrt{2T^{\beta+1}\ln(\frac{2}{\alpha})}.$$

Picking $\alpha=\frac{1}{LT}$ and $\beta\in(0,1)$, gives that $\mathcal{Q}$ enjoys sublinear expected regret

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t}))\neq\sigma_t\}\right]\leq\frac{1}{\psi(\frac{\gamma}{2})}+\frac{T}{\psi(\frac{\gamma}{2})T^\beta}\overline{R}_{\mathcal{A}}(T^\beta)+\sqrt{4T^{\beta+1}\ln(LT)}.$$

Since $\overline{R}_{\mathcal{A}}(T^\beta)$ is sublinear in T^β, $\mathcal{Q}$ is a realizable online learner for $\mathcal{H}$ with *sublinear* regret. Thus, for a sufficiently large T, $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{Q}(\mathcal{T}_{\pm,t}(\sigma_{<t})) \neq \sigma_t\}\right] < \frac{T}{2}$. This is a contradiction because $\{(\mathcal{T}_{\pm,t}(\sigma_{<t}), \sigma_t)\}_{t=1}^{T}$ is a realizable sequence of instances corresponding to a root-to-leaf path in $\mathcal{T}_\pm$ chosen *uniformly at random* and thus any realizable online learner must suffer expected regret at least $\frac{T}{2}$. Thus, if there exists a scale $\gamma > 0$ such that $\mathrm{fat}_\gamma^{\mathrm{seq}}(\mathcal{G}) = \infty$, there cannot exist an online learner for $\mathcal{G}$ with respect to $\psi \circ d_1$. ■

APPENDIX B

Online Learning with Set-Valued Feedback

B.1 Relationships Between Combinatorial Dimensions

B.1.1 Proof of (i) in Theorem 14.

Fix $p \geq 2$ and $\gamma \in (0, \frac{1}{p}]$. We first prove $\mathrm{MS}_\gamma(\mathcal{H}) \leq \mathrm{SL}(\mathcal{H})$. Let $\mathcal{T}$ be a $\Pi(\mathcal{Y})$-ary tree of depth $d_\gamma = \mathrm{MS}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$. For each internal node v in $\mathcal{T}$, keep the outgoing edges indexed by $\{\delta_y\}_{y\in\mathcal{Y}}$, where δ_y is a Dirac measure with point mass on y, and remove all other edges. Let A_y be the set labeling the outgoing edge from v indexed by δ_y. Since $\delta_y(A_y) \leq 1-\gamma$, we have $y \notin A_y$. Changing the index of edges from δ_y to y for all the remaining outgoing edges, we obtain a $\mathcal{Y}$-ary tree of depth d_γ. Repeating this process of pruning and reindexing recursively for every internal node, a $\Pi(\mathcal{Y})$-ary tree shattered by $\mathcal{H}$ can be transformed into a $\mathcal{Y}$-ary tree of the same depth shattered by $\mathcal{H}$. Thus, we must have $\mathrm{MS}_\gamma(\mathcal{H}) \leq \mathrm{SL}(\mathcal{H})$ for all $\gamma \in (0, \frac{1}{p}]$. For $\gamma = 0$, the shattering condition gives $\delta_y(A_y) < 1$, which implies that $y \notin A_y$. The rest of the arguments are identical to case $\gamma \in (0, \frac{1}{p}]$ presented above. Therefore, $\mathrm{MS}_\gamma(\mathcal{H}) \leq \mathrm{SL}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{p}]$.

We now prove $\mathrm{SL}_p(\mathcal{H}) \leq \mathrm{MS}_\gamma(\mathcal{H})$ for $\gamma \in [0, \frac{1}{p}]$. Let $\mathcal{T}$ be a $[p]$-ary tree shattered by $\mathcal{H}$. We expand $\mathcal{T}$ to obtain a $\Pi(\mathcal{Y})$-ary tree of depth d at scale $\frac{1}{p}$. Let v be the root node in $\mathcal{T}$, and $A_1, ..., A_p$ be the labels on the outgoing edges from v. To transform $\mathcal{T}$ to a $\Pi(\mathcal{Y})$-ary tree, we construct an outgoing edge for each measure. Fix a measure $\mu \in \Pi(\mathcal{Y})$. There must be an $A_\mu \in \{A_1, ..., A_p\}$ such that $\mu(A_\mu) \leq 1 - \frac{1}{p}$. Suppose, for the sake of contradiction, this is not true. That is, $\mu(A_i) > 1 - \frac{1}{p}$ for all $A_1, ..., A_p$, which further implies that $\mu(A_i^c) < \frac{1}{p}$. Since $\bigcap_{i=1}^p A_i = \emptyset$, we have $\mathcal{Y} = \bigcup_{i=1}^p A_i^c$ and thus

$$\mu(\mathcal{Y}) = \mu\left(\bigcup_{i=1}^p A_i^c\right) \leq \sum_{i=1}^p \mu(A_i^c) < 1,$$

which contradicts the fact that μ is a probability measure. Therefore, for every μ, there exists a $A_\mu \in \{A_1, ..., A_p\}$ such that $\mu(A_\mu) \leq 1 - \frac{1}{p}$. For every measure $\mu \in \Pi(\mathcal{Y})$, add an

outgoing edge from v indexed by μ and labeled by A_μ. Pick the sub-tree in $\mathcal{T}$ following the outgoing edge from v labeled by A_μ and append it to the newly constructed outgoing edge from v indexed by μ. Remove the two original outgoing edges from v indexed by elements of $[p]$ and their corresponding subtree. Upon repeating this process recursively for every internal node v in $\mathcal{T}$, we obtain a $\Pi(\mathcal{Y})$-ary tree that is $\frac{1}{p}$-shattered by $\mathcal{H}$. Thus, we have $\mathrm{MS}_{\frac{1}{p}}(\mathcal{H}) \geq \mathrm{SL}_p(\mathcal{H})$. Using monotonicity of MSdim, we therefore conclude that $\mathrm{MS}_\gamma(\mathcal{H}) \geq \mathrm{SL}_p(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{p}]$.

B.1.2 Proof of (ii) in Theorem 14.

Let $p = \mathrm{H}(\mathcal{S}(\mathcal{Y})) < \infty$. Given $p \geq 2$ and (i), it suffices to show that $\mathrm{SL}_p(\mathcal{H}) \geq \mathrm{MS}_\gamma(\mathcal{H}) \geq \mathrm{SL}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{p}]$. We first show that $\mathrm{MS}_\gamma(\mathcal{H}) \geq \mathrm{SL}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{p}]$

Consider a $\mathcal{Y}$-ary tree $\mathcal{T}$ of depth $d = \mathrm{SL}(\mathcal{H})$ shattered by $\mathcal{H}$. Let v be the root node of $\mathcal{T}$, and $\{A_y\}_{y\in\mathcal{Y}}$ be the sequence of sets labeling the outgoing edges from v. Since $p < \infty$, there must be a subsequence $\{A_{y_i}\}_{i=1}^p \subset \{A_y\}_{y\in\mathcal{Y}}$ such that $\cap_{i=1}^p A_{y_i} = \emptyset$. We keep the edges labeled by sets $\{A_{y_i}\}_{i=1}^p$ and remove all other edges, and repeat this process for every internal node v in $\mathcal{T}$. The subsequence of length p may not be unique, but choosing arbitrarily is permissible. Upon repeating this process recursively for every internal node in the tree $\mathcal{T}$, we obtain a tree $\mathcal{T}'$ of width p such that the sets labeling the p outgoing edges from any internal node are mutually disjoint.

Next, we expand $\mathcal{T}'$ to obtain a $\Pi(\mathcal{Y})$-ary tree of depth d at scale $\frac{1}{p}$. Let v be the root node in $\mathcal{T}'$, and $\{A_{y_i}\}_{i=1}^p$ be the labels on the outgoing edges from v. To transform $\mathcal{T}'$ to a $\Pi(\mathcal{Y})$-ary tree, we now construct an outgoing edge for each measure. Fix a measure $\mu \in \Pi(\mathcal{Y})$. There must be an $i \in [p]$ such that $\mu(A_{y_i}) \leq 1 - \frac{1}{p}$. Suppose, for the sake of contradiction, this is not true. That is, $\mu(A_{y_i}) > 1 - \frac{1}{p}$ for all $i \in [p]$, which further implies that $\mu(A_{y_i}^c) < \frac{1}{p}$. Since $\cap_{i=1}^p A_{y_i} = \emptyset$, we have $\mathcal{Y} = \cup_{i=1}^p A_{y_i}^c$ and thus

$$\mu(\mathcal{Y}) = \mu\left(\cup_{i=1}^p A_{y_i}^c\right) \leq \sum_{i=1}^p \mu(A_{y_i}^c) < \sum_{i=1}^p \frac{1}{p} < 1,$$

which contradicts the fact that μ is a probability measure. Therefore, for every μ, there exists a $y_\mu \in \{y_i\}_{i=1}^p$ such that $\mu(A_{y_\mu}) \leq 1 - \frac{1}{p}$. For every measure $\mu \in \Pi(\mathcal{Y})$, add an outgoing edge from v indexed by μ and labeled by A_{y_μ}. Pick the sub-tree in $\mathcal{T}'$ following the outgoing edge from v indexed by y_μ and append it to the newly constructed outgoing edge from v indexed by μ. Remove p remaining outgoing edges from v indexed by $y \in \{y_i\}_{i=1}^p$. Upon repeating this process for every internal node v in $\mathcal{T}'$, we obtain a $\Pi(\mathcal{Y})$-ary tree that is $\frac{1}{p}$-shattered by $\mathcal{H}$. Thus, we have $\mathrm{MS}_{\frac{1}{p}}(\mathcal{H}) \geq \mathrm{SL}(\mathcal{H})$. Using monotonicity of MSdim, we

therefore conclude that $\mathrm{MS}_\gamma(\mathcal{H}) \geq \mathrm{SL}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{p}]$.

We now prove that $\mathrm{SL}_p(\mathcal{H}) \geq \mathrm{MS}_\gamma(\mathcal{H})$. Suppose $\mathcal{T}$ is a $\Pi(\mathcal{Y})$-ary tree γ-shattered by $\mathcal{H}$ according to Definition 12. Let v be the root node of $\mathcal{T}$. Let A_y be the set labeling the outgoing edge from v indexed by δ_y. Since $\delta_y(A_y) \leq 1 - \gamma$, we have that $y \notin A_y$. Therefore, $\bigcap_{y \in \mathcal{Y}} A_y = \emptyset$. Since $p < \infty$, there must be a subsequence $\{A_{y_i}\}_{i=1}^p \subset \{A_y\}_{y \in \mathcal{Y}}$ such that $\bigcap_{i=1}^p A_{y_i} = \emptyset$. Keep the outgoing edges indexed by $\{\delta_{y_i}\}_{i=1}^p$ and remove all other edges along with their subtrees. For each $i \in [p]$, change the index δ_{y_i} to i. The root node v should now have p outgoing edges, where each edge is indexed by a unique element $i \in [p]$ and labeled by the set A_{y_i} such that $\bigcap_{i=1}^p A_{y_i} = \emptyset$. Repeat this process recursively on the subtrees following the p reindexed edges results into a SL_p tree of depth d_γ shattered by $\mathcal{H}$. Thus, $\mathrm{SL}_p(\mathcal{H}) \leq \mathrm{MS}_\gamma(\mathcal{H})$ for $\gamma \in (0, \frac{1}{p}]$. The case when $\gamma = 0$ follows similarly.

B.2 Deterministic Learnability in the Realizable Setting

B.2.1 Upperbounds

Proof. (of upperbound in Theorem 16) We first show that Algorithm 7 is a mistake-bound algorithm that makes at most $\mathrm{SL}(\mathcal{H})$ mistakes on any realizable stream. To show this, we argue that (1) every time Algorithm 7 makes a mistake, the SLdim of the version space goes down by 1 and (2) if the SLdim of the current version space is 0, then there is a prediction strategy such that the algorithm does not make any further mistakes.

Let $t \in [T]$ be a round where Algorithm 7 makes a mistake, that is $\hat{y}_t \notin S_t$, and $\mathrm{SL}(\mathcal{H}) > 0$. We show that the SL goes down by at least 1, that is $\mathrm{SL}(V_t) \leq \mathrm{SL}(V_{t-1}) - 1$. For the sake of contradiction, assume that $\mathrm{SL}(V_t) > \mathrm{SL}(V_{t-1}) - 1$. As $\mathrm{SL}(V_t) \leq \mathrm{SL}(V_{t-1})$, we must have $\mathrm{SL}(V_t) = \mathrm{SL}(V_{t-1}) =: m$. Since the SL did not go down and the algorithm made a mistake, the min-max prediction strategy implies that for every $y \in \mathcal{Y}$, there exists $A_y \in \mathcal{S}(\mathcal{Y})$ such that $y \notin A_y$ and $\mathrm{SL}(V_{t-1}(A_y)) = m$. Next, construct a $\mathcal{Y}$-ary tree $\mathcal{T}$ with x_t labeling the root node. For every $y \in \mathcal{Y}$, label the outgoing edge indexed by y with the set A_y. Append the $\mathcal{Y}$-ary tree of depth m associated with version space $V_{t-1}(A_y)$ to the edge indexed by y. Note that the depth of tree $\mathcal{T}$ must be $m + 1$, thus implying $\mathrm{SL}(V_{t-1}) = m + 1$, which is a contradiction. Therefore, it must be the case that $\mathrm{SL}(V_t) \leq \mathrm{SL}(V_{t-1}) - 1$.

Let $t^\star \in [T]$ be round when the algorithm makes its $\mathrm{SL}(\mathcal{H})^{th}$ mistake. If $t^\star$ does not exist, the algorithm makes at most $\mathrm{SL}(\mathcal{H}) - 1$ mistakes. So, without loss of generality, consider the case when $t^\star$ exists. It now suffices to show that the algorithm makes no further mistakes. We have already shown that $\mathrm{SL}(V_{t^\star}) = 0$. Next, we show that for any $t > t^\star$, there must exist $y \in \mathcal{Y}$ such that for all $A \in \mathcal{S}_t(\mathcal{Y})$ we have $y \in A$. Suppose, for the sake of contradiction, this is not true. That means, for all $y \in \mathcal{Y}$, there exists $A_y \in \mathcal{S}_t(\mathcal{Y})$ such that $y \notin A_y$.

Consider a tree with x_t in the root node, and every edge indexed by $y \in \mathcal{Y}$ is labeled with the set A_y. As $A_y \cap \{h(x_t) \mid h \in V_{t-1}\} \neq \emptyset$, for every y, there exists a hypothesis h_y such that $h_y(x_t) \in A_y$. By definition of SL, this implies that $\mathrm{SL}(V_{t-1}) \geq 1$, which contradicts the fact that $\mathrm{SL}(V_{t^\star}) = 0$. Thus, there must be a prediction strategy $y \in \mathcal{Y}$ such that for any set $S_t \in \mathcal{S}_t(\mathcal{Y})$ that the adversary can reveal, $y \in S_t$. With the prediction strategy in step 4, the algorithm makes no further mistakes. ■

B.2.2 Lowerbounds

Proof. (of lowerbound in Theorem 16) We now show that for any deterministic learner, there exists a realizable stream where the learner makes at least $\mathrm{SL}(\mathcal{H}) = d$ mistakes. The stream is obtained by traversing the Set Littlestone tree of depth d, adapting to the algorithm's prediction. Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued, $\mathcal{Y}$-ary tree of depth d that is shattered by $\mathcal{H}$. Let $(f_1, ..., f_d)$ be the sequence of edge-labeling functions $f_t : \mathcal{Y}^t \to \mathcal{S}(\mathcal{Y})$ associated with $\mathcal{T}$. Consider the stream $\{(\mathcal{T}_1(\hat{y}_{<t}), f_t(\hat{y}_{\leq t}))\}_{t=1}^d$, where $\mathcal{T}_1(\hat{y}_{<1})$ is the root node of the tree, and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_d)$ is algorithm's prediction on rounds $1, 2, \ldots, d$. Note that we can use the learner's prediction on round t to generate the true feedback for round t because the learner is deterministic and its prediction on any instance can be simulated apriori. Since we have $\hat{y}_t \notin f_t(\hat{y}_{\leq t})$ for all $t \in [d]$ by the definition of the tree, the algorithm makes at least d mistake in the stream above. Finally, the stream considered above is realizable because there exists $h_{\hat{y}}$ such that $h_{\hat{y}}(\mathcal{T}_t(\hat{y}_{<t})) \in f_t(\hat{y}_{\leq t})$ for all $t \in [d]$. This completes our proof. ■

B.3 Randomized Learnability in the Realizable Setting

B.3.1 Upperbounds

B.3.1.1 Fixed-scale Randomized Learner

We give a fixed-scale learner in the realizable setting and prove a guarantee on its expected number of mistakes. In particular, we show that the expected mistake bound of Algorithm 12, for any fixed input scale $\gamma > 0$, is at most $\gamma T + \mathrm{MS}_\gamma(\mathcal{H})$ on any realizable stream.

Lemma 12 (Fixed-scale Randomized Learning Guarantee)**.** *For any $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$, $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$, and any input scale $\gamma > 0$, the expected cumulative loss of Algorithm 12, on any realizable stream, is $\leq \gamma T + \mathrm{MS}_\gamma(\mathcal{H})$.*

Proof. We show that given any target accuracy $\varepsilon > 0$, the expected cumulative loss of Algorithm 12 is at most $d_\varepsilon + \varepsilon T$ on any realizable stream, where $d_\varepsilon = \mathrm{MS}_\varepsilon(\mathcal{H})$. In fact, we

Algorithm 12 Randomized Standard Optimal Algorithm (RSOA)

Require: $\mathcal{H}$, Target accuracy $\varepsilon > 0$

Initialize $V_0 = \mathcal{H}$
for $t = 1, \ldots, T$ **do**
  Receive unlabeled example $x_t \in \mathcal{X}$.
  For each $A \in \mathcal{S}(\mathcal{Y})$, define $V_{t-1}(A) := \{h \in V_{t-1} \mid h(x_t) \in A\}$.
  Let $\mathcal{S}_t(\mathcal{Y}) := \{A \in \mathcal{S}(\mathcal{Y}) : A \cap \{h(x_t) \mid h \in V_{t-1}\} \neq \emptyset\}$.
  if $\mathrm{MS}_\varepsilon(V_{t-1}) = 0$ **then**
    Let $\hat{\mu}_t \in \Pi(\mathcal{Y})$ be such that for all $A \in \mathcal{S}_t(\mathcal{Y})$ we have $\hat{\mu}_t(A) > 1 - \varepsilon$.
  else
    Compute

$$\hat{\mu}_t = \operatorname*{arg\,min}_{\mu \in \Pi(\mathcal{Y})} \max_{\substack{A \in \mathcal{S}(\mathcal{Y}) \\ \mu(A) \leq 1-\varepsilon}} \mathrm{MS}_\varepsilon(V_{t-1}(A)).$$

  end if
  Predict $\hat{y}_t \sim \hat{\mu}_t$.
  Receive feedback S_t and update $V_t = V_{t-1}(S_t)$.
end for

show that Algorithm 12 achieves an even *stronger* guarantee, namely that on any realizable sequence $\{(x_t, S_t)\}_{t=1}^T$, Algorithm 12 computes distributions $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that

$$\sum_{t=1}^{T} \mathbb{1}\{\hat{\mu}_t(S_t^c) \geq \varepsilon\} \leq d_\varepsilon. \tag{B.1}$$

From here, it follows that $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{\hat{y}_t \notin S_t\}\right] \leq d_\varepsilon + \varepsilon T$. To see this, observe that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\hat{y}_t \notin S_t\}\right] &= \sum_{t=1}^{T} \mathbb{P}\left[\hat{y}_t \notin S_t\ \right] \\
&= \sum_{t=1}^{T} \mathbb{P}\left[\hat{y}_t \notin S_t\ \right] \mathbb{1}\{\hat{\mu}_t(S_t^c) \geq \varepsilon\} + \mathbb{P}\left[\hat{y}_t \notin S_t\ \right] \mathbb{1}\{\hat{\mu}_t(S_t^c) < \varepsilon\} \\
&\leq \sum_{t=1}^{T} \mathbb{1}\{\hat{\mu}_t(S_t^c) \geq \varepsilon\} + \varepsilon T \\
&\leq d_\varepsilon + \varepsilon T
\end{aligned}$$

We now show that the outputs of Algorithm 12 satisfy Equation (B.1). It suffices to show that (1) on any round where $\hat{\mu}_t(S_t) \leq 1 - \varepsilon$ and $\mathrm{MS}_\varepsilon(V_{t-1}) > 0$, we have $\mathrm{MS}_\varepsilon(V_t) \leq \mathrm{MS}_\varepsilon(V_{t-1}) - 1$, and (2) if $\mathrm{MS}_\varepsilon(V_{t-1}) = 0$ then there is always a distribution $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that $\mathbb{P}\left[\hat{y}_t \notin S_t\right] \leq \varepsilon$.

Let $t \in [T]$ be a round where $\hat{\mu}_t(S_t) \leq 1 - \varepsilon$ and $\mathrm{MS}_\varepsilon(V_{t-1}) > 0$. For the sake contradiction,

suppose that $\mathrm{MS}_\varepsilon(V_t) = \mathrm{MS}_\varepsilon(V_{t-1}) = d$. Then, by the min-max computation in line (4) of Algorithm 12, for every measure $\mu \in \Pi(\mathcal{Y})$, there exists a subset $\mathcal{A}_\mu \in \mathcal{S}(\mathcal{Y})$ such that $\mu(A_\mu) \leq 1 - \varepsilon$ and $\mathrm{MS}_\varepsilon(V_{t-1}(A_\mu)) = d$. Now construct a tree $\mathcal{T}$ with x_t labeling the root node. For each measure $\mu \in \Pi(\mathcal{Y})$, construct an outgoing edge from x_t indexed by μ and labeled by A_μ. Append the tree of depth d associated with the version space $V_{t-1}(A_\mu)$ to the edge indexed by μ. Note that the depth of $\mathcal{T}$ must be $d + 1$. Therefore, by definition of MSdim, we have that $\mathrm{MS}_\varepsilon(V_{t-1}) = d + 1$, a contradiction. Thus, it must be the case that $\mathrm{MS}_\varepsilon(V_t) \leq \mathrm{MS}_\varepsilon(V_{t-1}) - 1$.

Now, suppose $t \in [T]$ is a round such that $\mathrm{MS}_\varepsilon(V_{t-1}) = 0$. We show that there always exist a distribution $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that for all $A \in \mathcal{S}_t(\mathcal{Y})$, we have $\hat{\mu}_t(A) \geq 1 - \varepsilon$. Since we are in the realizable setting, it must be the case that $S_t \in \mathcal{S}_t(\mathcal{Y})$. Therefore, $\hat{\mu}_t(S_t) \geq 1 - \varepsilon$ and $\mathbb{P}\left[\hat{y}_t \notin S_t\right] \leq \varepsilon$ as needed. To see why such a $\hat{\mu}_t$ must exist, suppose for the sake of contradiction that it does not exist. Then, for all $\mu \in \Pi(\mathcal{Y})$, there exists a set $A_\mu \in \mathcal{S}_t(\mathcal{Y})$ such that $\mu(A_\mu) \leq 1 - \varepsilon$. As before, consider a tree with root node labeled by x_t. For each measure $\mu \in \Pi(\mathcal{Y})$, construct an outgoing edge from x_t indexed by μ and labeled by A_μ. Since $A_\mu \cap \{h(x_t) \mid h \in V_{t-1}\} \neq \emptyset$, there exists a hypothesis h_μ such that $h_\mu(x_t) \in A_\mu$. By definition of MSdim, this implies that $\mathrm{MS}_\varepsilon(V_{t-1}) \geq 1$, which contradicts the fact that $\mathrm{MS}_\varepsilon(V_{t-1}) = 0$. Thus, there must be a distribution $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that for any set $A \in \mathcal{S}_t(\mathcal{Y})$, we have $\hat{\mu}_t(A) \geq 1 - \varepsilon$. Since this is precisely the distribution that Algorithm 12 plays in step (3) and since $\mathrm{MS}_\varepsilon(V_{t'}) \leq \mathrm{MS}_\varepsilon(V_{t-1})$ for all $t' \geq t$, the algorithm no longer suffers expected loss more than ε. This completes the proof of Lemma 12. ■

We point out that Filmus et al. [2023] also considers a randomized online learner in the realizable setting that shares similarities with Algorithm 12. In particular, their algorithm also maintains a version space and optimizes over probability distributions. However, they only consider binary classification and use a different complexity measure. Moreover, the idea of optimizing over probability distributions on a measurable space should also remind the reader of the generic min-max algorithm proposed by Rakhlin et al. [2012a].

B.3.1.2 Multi-scale Randomized Learner

The RSOA (Algorithm 12) runs at a fixed, pre-determined scale $\gamma \in [0, 1]$. In this section, we upgrade this result by adapting the technique from Daskalakis and Golowich [2022] to give a randomized, *multi-scale* online learner (Algorithm 14) in the realizable setting. Lemma 13 presents the main result, which bounds the expected cumulative loss of Algorithm 14 on any realizable data stream and gives the upperbound stated in Theorem 17.

Lemma 13 (Multi-scale Randomized Online Learner)**.** *For any $\mathcal{S}(\mathcal{Y}) \subseteq \sigma(\mathcal{Y})$ and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, the expected cumulative loss of Algorithm 14 on any realizable stream is at most*

$$C \inf_{\gamma \in [0,1]} \left\{ \gamma T + \int_{\gamma}^{1} \mathrm{MS}_{\eta}(\mathcal{H}) \, d\eta \right\},$$

for some universal constant $C > 0$.

We highlight that the guarantee given by Lemma 13 is analogous to Dudley's integral entropy bound in batch setting and also matches Theorem 1 in Daskalakis and Golowich [2022]. Compared to Lemma 12, the upperbound given by Lemma 13 can be significantly better. For example, when the Measure Shattering dimension exhibits growth $\mathrm{MS}_{\gamma}(\mathcal{H}) \approx \gamma^{-p}$ for some $p \in (0, 1)$, the bound given by Lemma 13 is constant $O(1)$, while the bound given by Lemma 12 scales according to $T^{\frac{p}{(1+p)}}$.

The main algorithmic idea needed to obtain the guarantee in Lemma 13 is to figure out how to make predictions using more than one scale. At a high-level, our multi-scale learner uses a sequence of N scales $\{\gamma_i\}_{i=1}^{N}$, where $\gamma_i = \frac{1}{2^i}$, to compute a sequence of measures $\{\mu_t^i\}_{i=1}^{N} \subset \Pi(\mathcal{Y})$ in each round $t \in [T]$. Then, our multi-scale learner uses the Measure Selection Procedure, defined in Algorithm 13, to carefully select one of the measures $\hat{\mu}_t \in \{\mu_i^t\}_{i=1}^{N}$ and makes a prediction $\hat{y}_t \sim \hat{\mu}_t$.

Algorithm 13 Measure Selection Procedure (MSP)

Require: Sequence of measures $\mu_1, \ldots, \mu_N$, valid sets $\mathcal{S} \subseteq \sigma(\mathcal{Y})$

if there exists $m \in \mathbb{N}$ such that for all $2 \leq i \leq m$:

$$\sup_{A \in \mathcal{S}} |\mu_i(A^c) - \mu_{i-1}(A^c)| \leq 2\gamma_{i-1} \quad \text{but} \quad \inf_{A \in \mathcal{S}} |\mu_m(A^c) - \mu_{m+1}(A^c)| \geq 2\gamma_m$$

then
    return m.
else
    return N.
end if

Once the true label set is revealed, the multi-scale learner updates its self in the exact same way as RSOA. Algorithm 14 formalizes the idea above.

We now prove Lemma 13, which closely follows the analysis by Daskalakis and Golowich [2022].

Proof. Fix a $N \in \mathbb{N}$. Our first goal is to show that on any realizable stream, the expected cumulative loss of Algorithm 14 is at most

Algorithm 14 Multi-scale Online Learner (MSOL)

Require: $\mathcal{H}$, number of scales N

Initialize: $V_0 = \mathcal{H}$, $\gamma_i = \frac{1}{2^i}$ for $i \in [N]$
for $t = 1, \ldots, T$ **do**
  Receive unlabeled example $x_t \in \mathcal{X}$.
  For each $A \in \mathcal{S}(\mathcal{Y})$, define $V_{t-1}(A) := \{h \in V_{t-1} \mid h(x_t) \in A\}$.
  Let $\mathcal{S}_t(\mathcal{Y}) := \{A \in \mathcal{S}(\mathcal{Y}) : A \cap \{h(x_t) \mid h \in V_{t-1}\} \neq \emptyset\}$.
  if $\mathrm{MS}_{\gamma_N}(V_{t-1}) = 0$ **then**
    Let $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that $\hat{\mu}_t(A) > 1 - \gamma_N$ for all $A \in \mathcal{S}_t(\mathcal{Y})$.
  else
    for $i = 1, \ldots, N$ **do**
      if $\mathrm{MS}_{\gamma_i}(V_{t-1}) = 0$ **then**
        Let $\mu_t^i \in \Pi(\mathcal{Y})$ such that $\mu_t^i(A) > 1 - \gamma_i$ for all $A \in \mathcal{S}_t(\mathcal{Y})$.
      else
        Let
$$\mu_t^i = \operatorname*{arg\,min}_{\mu \in \Pi(\mathcal{Y})} \max_{\substack{A \in \mathcal{S}(\mathcal{Y}) \\ \mu(A) \leq 1 - \gamma_i}} \mathrm{MS}_{\gamma_i}(V_{t-1}(A)).$$
      end if
    end for
    Compute $m_t = \mathrm{MSP}(\{\mu_t^i\}_{i=1}^N, \mathcal{S}_t(\mathcal{Y}))$ and let $\hat{\mu}_t = \mu_t^{m_t}$.
  end if
  Predict $\hat{y}_t \sim \hat{\mu}_t$.
  Receive feedback $S_t \in \mathcal{S}_t(\mathcal{Y})$ and update $V_t = V_{t-1}(S_t)$.
end for

$$\gamma_N T + 16 \sum_{i=1}^{N} \gamma_i \cdot \mathrm{MS}_{\gamma_i}(\mathcal{H}),$$

where $\gamma_i = \frac{1}{2^i}$. To that end, let $\{(x_t, S_t)\}_{t=1}^{T}$ denote the realizable stream that is to be observed by the learner. For all $t \in [T+1]$, define the potential function

$$\Phi_t = (T + 1 - t)\gamma_N + 16 \sum_{i=1}^{N} \gamma_i \, \mathrm{MS}_{\gamma_i}(V_{t-1}).$$

It suffices to show that $\Phi_t - \Phi_{t+1} \geq \hat{\mu}_t(S_t^c)$ for all $t \in [T]$. To see why this is sufficient, observe that summing over all $t \in [T]$ gives

$$\sum_{t=1}^{T} \hat{\mu}_t(S_t^c) \leq \sum_{t=1}^{T} (\Phi_t - \Phi_{t+1}) = \Phi_1 - \Phi_{T+1} \leq T\gamma_N + 16 \sum_{i=1}^{N} \gamma_i \, \mathrm{MS}_{\gamma_i}(\mathcal{H})$$

where the inequality follows from the fact that $\Phi_{T+1} \geq 0$ and $V_0 = \mathcal{H}$. Finally, noting that $\mathbb{E}_{\hat{y}_t \sim \hat{\mu}_t} [\mathbb{1}\{\hat{y}_t \notin S_t\}] = \hat{\mu}_t(S_t^c)$ gives $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\hat{y}_t \notin S_t\}\right] \leq T\gamma_N + 16 \sum_{i=1}^{N} \gamma_i \, \mathrm{MS}_{\gamma_i}(\mathcal{H})$ as desired.

The rest of this proof is dedicated to showing that $\Phi_t - \Phi_{t+1} \geq \hat{\mu}_t(S_t^c)$ for all $t \in [T]$. Fix a $t \in [T]$. Using the definition of Φ_t, we need to show that

$$\gamma_N + 16 \sum_{i=1}^{N} \gamma_i (\mathrm{MS}_{\gamma_i}(V_{t-1}) - \mathrm{MS}_{\gamma_i}(V_t)) \geq \hat{\mu}_t(S_t^c). \tag{B.2}$$

If $\hat{\mu}_t(S_t^c) < \gamma_N$, then Inequality B.2 holds since for all $t \in [T]$ and $i \in [N]$, $\mathrm{MS}_{\gamma_i}(V_{t-1}) \geq \mathrm{MS}_{\gamma_i}(V_t)$. Thus, we focus on the case where $\hat{\mu}_t(S_t^c) \geq \gamma_N$.

Suppose $\hat{\mu}_t(S_t^c) \geq \gamma_N$. Then, $\mathrm{MS}_{\gamma_N}(V_{t-1}) \geq 1$, the for-loop on line 5(a) runs, and the measure $\hat{\mu}_t = \mu_t^{m_t}$ computed on line 5(b) is used to make a prediction. This is because when $\mathrm{MS}_{\gamma_N}(V_{t-1}) = 0$, we are guaranteed the existence of a measure $\hat{\mu}_t \in \Pi(\mathcal{Y})$ such that $\hat{\mu}_t(S_t^c) < \gamma_N$ (see proof of Theorem 17) and by line 4, this would have precisely been the measure the learner uses to make its prediction.

We now show that when $\hat{\mu}_t(S_t^c) \geq \gamma_N$, there exists an index $j \in [N]$ such that $\gamma_j \geq \frac{\hat{\mu}_t(S_t^c)}{16}$ and $\mu_t^j(S_t^c) \geq \gamma_j$. This implies Inequality (2), because if $\mu_t^j(S_t^c) \geq \gamma_j$, then $\mathrm{MS}_{\gamma_j}(V_{t-1}) \geq 1$, and $\mathrm{MS}_{\gamma_j}(V_t) < \mathrm{MS}_{\gamma_j}(V_{t-1})$, which follows from the definition of MSdim, and the min-max prediction strategy in step 5(a:ii). Then, we can compute

$$\gamma_N + 16 \sum_{i=1}^{N} \gamma_i (\mathrm{MS}_{\gamma_i}(V_{t-1}) - \mathrm{MS}_{\gamma_i}(V_t)) \geq 16\gamma_j (\mathrm{MS}_{\gamma_j}(V_{t-1}) - \mathrm{MS}_{\gamma_j}(V_t)) \geq \hat{\mu}_t(S_t^c),$$

which matches the guarantee of Inequality B.2. Accordingly, the rest of the proof will focus on showing the existence of such an index $j \in [N]$. To do so, let $k \in \mathbb{N}$ denote the smallest natural number such that $\hat{\mu}_t(S_t^c) \geq \gamma_k = \frac{1}{2^k}$. By definition of k, we have that $\hat{\mu}(S_t^c) < \gamma_{k-1} = 2\gamma_k$. Note that $k \neq N+1$ since that would imply that $\hat{\mu}(S_t^c) < \frac{1}{2^N} = \gamma_N$ which contradicts the fact that $\hat{\mu}_t(S_t^c) \geq \gamma_N$. Thus, it must be the case that $k \in \{1, ..., N\}$. Let $m_t = \text{MSP}(\{\mu_t^i\}_{i=1}^N, \mathcal{S}_t(\mathcal{Y}))$ denote the index output by MSP in round t. We consider two subcases: (1) $k \in \{m_t + 1, ..., N\}$ and (2) $k \in \{1, ..., m_t\}$.

Case I. Suppose $k \in \{m_t + 1, ..., N\}$. Then, we show that $j = m_t + 1$. That is, $\gamma_{m_t+1} \geq \frac{\hat{\mu}_t(S_t^c)}{16}$ and $\mu_t^{m_t+1}(S_t^c) \geq \gamma_{m_t+1}$ Recall that $\hat{\mu}_t(S_t^c) = \mu_t^{m_t}(S_t^c)$. Since $m_t < N$, by definition, we have that $\inf_{A \in \mathcal{S}_t(\mathcal{Y})} |\mu_t^{m_t}(A) - \mu_t^{m_t+1}(A)| \geq 2\gamma_{m_t}$. This implies that $|\mu_t^{m_t}(S_t^c) - \mu_t^{m_t+1}(S_t^c)| \geq 2\gamma_{m_t}$. Moreover, we have that $\mu_t^{m_t}(S_t^c) = \hat{\mu}_t(S_t^c) < 2\gamma_k \leq 2\gamma_{m_t+1} = \gamma_{m_t}$. Combining the two inequalities, we get that $\mu_t^{m_t+1}(S_t^c) \geq \gamma_{m_t} > \gamma_{m_t+1}$. Since $\hat{\mu}_t(S_t^c) < 2\gamma_{m_t+1}$, we also obtain $\gamma_{m_t+1} \geq \frac{\hat{\mu}_t(S_t^c)}{2} > \frac{\hat{\mu}_t(S_t^c)}{16}$. This completes this case.

Now, suppose that $k \in \{1, ..., m_t\}$. Then we know that $\mu_t^{m_t}(S_t^c) = \hat{\mu}_t(S_t^c) \geq \gamma_k \geq \gamma_{m_t}$. We further break this case down into two subcases: (a) $k \in \{m_t - 3, m_t - 2, ..., m_t\}$ and (b) $k \in \{1, ..., m_t - 4\}$.

Case II(a). Consider the case where $k \in \{m_t - 3, m_t - 2, ..., m_t\}$. We show that $j = m_t$. We know that $\hat{\mu}_t(S_t^c) < 2\gamma_k = 2\frac{1}{2^k} = 16\gamma_{k+3} \leq 16\gamma_{m_t}$. This implies that $\gamma_{m_t} \geq \frac{\hat{\mu}_t(S_t^c)}{16}$. Since we have that $\mu_t^{m_t}(S_t^c) = \hat{\mu}_t(S_t^c) \geq \gamma_{m_t}$, this completes the proof that $j = m_t$.

Case II(b). Consider the case where $k \in \{1, ..., m_t - 4\}$. Here, we will show that $j = k + 1$. Observe that,

$$|\mu_t^{m_t}(S_t^c) - \mu_t^{k+3}(S_t^c)| \leq \sum_{i=k+3}^{m_t-1} |\mu_t^i(S_t^c) - \mu_t^{i+1}(S_t^c)| \leq \sum_{i=k+3}^{m_t-1} 2\gamma_i$$
$$\leq 2\sum_{i=k+3}^{\infty} \frac{1}{2^i} = 4\gamma_{k+3} = \frac{\gamma_k}{2},$$

where the second inequality follows from the definition of $m_t = \text{MSP}(\{\mu_t^i\}_{i=1}^N, \mathcal{S}_t(\mathcal{Y}))$. This implies that $\mu_t^{m_t}(S_t^c) - \mu_t^{k+3}(S_t^c) \leq \frac{\gamma_k}{2}$. Since $\mu_t^{m_t}(S_t^c) \geq \gamma_k$, we get that $\mu_t^{k+3}(S_t^c) \geq \frac{\gamma_k}{2} = 4\gamma_{k+3} \geq \gamma_{k+3}$. Finally, recall that $\hat{\mu}_t(S_t^c) < 2\gamma_k = 16\gamma_{k+3}$, implying that $\gamma_{k+3} \geq \frac{\hat{\mu}_t(S_t^c)}{16}$ as desired. This completes the subcase.

Overall, we have shown that when $\hat{\mu}_t(S_t^c) \geq \gamma_N$, there exists an index $j \in [N]$ such that $\gamma_j \geq \frac{\hat{\mu}_t(S_t^c)}{16}$ and $\mu_t^j(S_t^c) \geq \gamma_j$. This means that for all $t \in [T]$, $\Phi_t - \Phi_{t+1} \geq \hat{\mu}_t(S_t^c)$ and therefore the expected cumulative loss of Algorithm 14 is at most $\gamma_N T + \sum_{i=1}^N \gamma_i \cdot \text{MS}_{\gamma_i}(\mathcal{H})$, as needed.

Our next goal is to show that if $\gamma^* = \inf_{\gamma > 0}\{\gamma T + \int_\gamma^1 \text{MS}_\eta(\mathcal{H}) d\eta\}$, then setting $N = \left\lceil \frac{1}{\log 2\gamma^*} \right\rceil$ gives that

$$\gamma_N\, T + 16 \sum_{i=1}^{N} \gamma_i \cdot \mathrm{MS}_{\gamma_i}(\mathcal{H}) \leq C \inf_{\gamma > 0} \left\{ \gamma T + \int_{\gamma}^{1} \mathrm{MS}_{\eta}(\mathcal{H}) d\eta \right\}$$

for some constant $C > 0$. However, this follows from the fact that when $N = \left\lceil \frac{1}{\log 2\gamma^*} \right\rceil$, $\gamma_N \leq 2\gamma^*$ and the fact that $16 \sum_{i=1}^{N} \gamma_i \cdot \mathrm{MS}_{\gamma_i}(\mathcal{H})$ is, up to a constant factor, the appropriate lower Reimann sum such that $16 \sum_{i=1}^{N} \gamma_i \cdot \mathrm{MS}_{\gamma_i}(\mathcal{H}) \leq C \int_{\gamma^*}^{1} \mathrm{MS}_{\eta}(\mathcal{H}) d\eta$. ■

B.3.2 Lowerbounds

In this section, we prove the lowerbound given in Theorem 17. Fix $\gamma > 0$. Let $\mathcal{H}$ and $\mathcal{S}(\mathcal{Y})$ be such that $\mathrm{MS}_{\gamma}(\mathcal{H}) = d_{\gamma}$. By definition of MSdim, there exists a $\mathcal{X}$-valued, $\Pi(\mathcal{Y})$-ary tree $\mathcal{T}$ of depth d_{γ} shattered by $\mathcal{H}$. Let $(f_1, ..., f_d)$ be the sequence of edge-labeling functions $f_t : \Pi(\mathcal{Y})^t \to \mathcal{S}(\mathcal{Y})$ associated with $\mathcal{T}$. Let $\mathcal{A}$ be any randomized learner for $\mathcal{H}$. Our goal will be to use $\mathcal{T}$ and its edge-labeling functions $(f_1, ..., f_d)$ to construct a hard realizable stream for $\mathcal{A}$ such that on every round, $\mathcal{A}$ makes a mistake with probability at least γ. This stream is obtained by traversing $\mathcal{T}$, adapting to the sequence of distributions output by $\mathcal{A}$.

To that end, for every round $t \in [d_{\gamma}]$, let $\hat{\mu}_t$ denote the distribution that $\mathcal{A}$ computes before making its prediction $\hat{y}_t$. Consider the stream $\{(\mathcal{T}_t(\hat{\mu}_{<t}), f_t(\hat{\mu}_{\leq t}))\}_{t=1}^{d_{\gamma}}$, where $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_{d_{\gamma}})$ denotes the sequence of distributions output by $\mathcal{A}$. This stream is obtained by starting at the root of $\mathcal{T}$, passing $\mathcal{T}_1$ to $\mathcal{A}$, observing the distribution $\hat{\mu}_1$ computed by $\mathcal{A}$, passing the label $f_t(\hat{\mu}_{\leq 1})$ to $\mathcal{A}$, and then finally moving along the edge labeled by $\hat{\mu}_1$. This process then repeats $d_{\gamma} - 1$ times until the bottom of $\mathcal{T}$ is reached. Note that we can observe and use the distribution computed by $\mathcal{A}$ on round t to generate the true feedback because a randomized algorithm *deterministically* maps a sequence of labeled instances to a distribution. Moreover the stream is realizable since by the definition of shattering, there exists a $h_{\hat{\mu}} \in \mathcal{H}$ such that $h_{\hat{\mu}}(\mathcal{T}_t(\hat{\mu}_{<t})) \in f_t(\hat{\mu}_{\leq t})$ for all $t \in [d_{\gamma}]$.

Now, we are ready to show that this stream is difficult for $\mathcal{A}$. By definition of the tree, for all $t \in [d_{\gamma}]$, we have that $\hat{\mu}_t(f_t(\hat{\mu}_{\leq t})) \leq 1 - \gamma$. Therefore, since $\mathcal{A}$ receives $f_t(\hat{\mu}_{\leq t})$ as feedback on round t, we have that $\mathbb{P}\left[\mathcal{A}(\mathcal{T}_t(\hat{\mu}_{<t})) \notin f_t(\hat{\mu}_{\leq t})\right] = \mathbb{P}_{\hat{y}_t \sim \hat{\mu}_t}\left[\hat{y}_t \notin f_t(\hat{\mu}_{\leq t})\right] = 1 - \hat{\mu}_t(f_t(\hat{\mu}_{\leq t})) \geq \gamma$ for all $t \in [d_{\gamma}]$. Summing over all $t \in [d_{\gamma}]$ gives that

$$\mathbb{E}\left[\sum_{t=1}^{d_{\gamma}} \mathbb{1}\{\mathcal{A}(\mathcal{T}_t(\hat{\mu}_{<t})) \notin f_t(\hat{\mu}_{\leq t})\}\right] = \sum_{t=1}^{d_{\gamma}} \mathbb{P}\left[\mathcal{A}(\mathcal{T}_t(\hat{\mu}_{<t})) \notin f_t(\hat{\mu}_{\leq t})\right] \geq \gamma\, d_{\gamma}.$$

This shows that $\mathcal{A}$ makes at least γd_{γ} mistakes in expectation on the realizable stream $\{(\mathcal{T}_t(\hat{\mu}_{<t}), f_t(\hat{\mu}_{\leq t}))\}_{t=1}^{d_{\gamma}}$. Since our choice of γ and the randomized algorithm $\mathcal{A}$ was arbitrary, this holds true for any $\gamma > 0$ and any randomized online learner. This completes the proof.

B.4 Agnostic Learnability

B.4.1 Agnostic Upperbound

Proof. (of (i) in Theorem 18) Let $(x_1, S_1), \ldots, (x_T, S_T)$ be the data stream. Let $h^\star = \arg\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \notin S_t\}$ be an optimal function in hind-sight. For a target accuracy $\varepsilon > 0$, let $d_\varepsilon = \mathrm{MS}_\varepsilon(\mathcal{H})$. Given time horizon T, let $L_T = \{L \subset [T]; |L| \leq d_\varepsilon\}$ denote the set of all possible subsets of $[T]$ with size at most d_ε. For every $L \in L_T$ define an expert E_L such that

$$E_L(x_t) := \mathrm{RSOA}_\varepsilon(x_t \mid L_{<t}),$$

where $L_{<t} = L \cap \{1, 2, \ldots, t-1\}$ and $\mathrm{RSOA}_\varepsilon(x_t \mid L_{<t})$ is the prediction of the Randomized Standard Optimal Algorithm (RSOA), defined as Algorithm 12, running at scale ε that has updated on labeled examples $\{(x_i, S_i)\}_{i \in L_{<t}}$. Let $\mathcal{E} = \bigcup_{L \in L_T} \{E_L\}$ denote the set of all Experts parameterized by subsets $L \in L_T$. Note that $|\mathcal{E}| = \sum_{i=0}^{d_\varepsilon} \binom{T}{i} \leq T^{d_\varepsilon}$. Finally, given our set of experts $\mathcal{E}$, we run the Randomized Exponential Weights Algorithm (REWA), denoted hereinafter as $\mathcal{P}$, over the stream $(x_1, S_1), ..., (x_T, S_T)$ with a learning rate $\eta = \sqrt{2\ln(|\mathcal{E}|)/T}$. Let B denote the random variable associated with the internal randomness of the RSOA. Then, conditioned on B, Theorem 21.11 of Shalev-Shwartz and Ben-David [2014] tells us that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\{\mathcal{P}(x_t) \notin S_t\} \mid B\right] &\leq \inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \mathbb{1}\{E(x_t) \notin S_t\} + \sqrt{2T\ln(|\mathcal{E}|)} \\
&\leq \inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \mathbb{1}\{E(x_t) \notin S_t\} + \sqrt{2d_\varepsilon T \ln(T)},
\end{aligned}$$

where the second inequality follows because $|\mathcal{E}| \leq T^{d_\varepsilon}$. Taking expectations on both sides of the inequality above, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(x_t) \notin S_t\}\right] \leq \mathbb{E}\left[\inf_{E \in \mathcal{E}} \sum_{t=1}^{T} \mathbb{1}\{E(x_t) \notin S_t\}\right] + \sqrt{2d_\varepsilon T \ln(T)},$$

Here, we have an expectation on the right-hand side because the Expert predictions are random. Define $R^\star = \{t \in [T] \mid h^\star(x_t) \in S_t\}$ to be the part of the stream realizable by $h^\star$.

Note that the set $R^\star$ is not random because the adversary is oblivious. Then, we have

$$\begin{aligned}\inf_{E\in\mathcal{E}}\sum_{t=1}^{T}\mathbb{1}\{E(x_t)\notin S_t\} &= \inf_{E\in\mathcal{E}}\left(\sum_{t\in R^\star}\mathbb{1}\{E(x_t)\notin S_t\}+\sum_{t\notin R^\star}\mathbb{1}\{E(x_t)\notin S_t\}\right)\\ &\le \inf_{E\in\mathcal{E}}\sum_{t\in R^\star}\mathbb{1}\{E(x_t)\notin S_t\}+\sum_{t\notin R^\star}\mathbb{1}\{h^\star(x_t)\notin S_t\}\\ &= \inf_{E\in\mathcal{E}}\sum_{t\in R^\star}\mathbb{1}\{E(x_t)\notin S_t\}+\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\{h(x_t)\notin S_t\},\end{aligned}$$

where the first inequality above follows because $\mathbb{1}\{h^\star(x_t)\notin S_t\}=1$ for all $t\in R^\star$. Thus, the expected cumulative loss of $\mathcal{P}$ is

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mathcal{P}(x_t)\notin S_t\}\right]\le\inf_{h\in\mathcal{H}}\sum_{t=1}^{T}\mathbb{1}\{h(x_t)\notin S_t\}+\mathbb{E}\left[\inf_{E\in\mathcal{E}}\sum_{t\in R^\star}\mathbb{1}\{E(x_t)\notin S_t\}\right]+\sqrt{2d_\varepsilon T\ln(T)} \tag{B.3}$$

Thus, it suffices to show that the second term on the right side of the inequality above is $\le d_\varepsilon+\varepsilon T$.

To do so, we need some more notation. Let us define $\hat{\mu}_t=\mu\text{-RSOA}_\varepsilon(x_t\mid L)$ to be the measure returned by RSOA_ε, as described in step 4 and 5 of Algorithm 12, for x_t given that the algorithm has been updated on examples of the time points $t\in L$. We say that μ-RSOA_ε makes a mistake on round t if $\mathbb{1}\{\hat{\mu}_t(S_t^c)\ge\varepsilon\}=1$. With this notion of the mistake, Equation (B.1) tells us that RSOA_ε, run and updated on any realizable sequence, makes at most d_ε mistakes. Since μ-$\text{RSOA}_\varepsilon(x\mid L)$ is a deterministic mapping from the past examples to a probability measure in $\Pi(\mathcal{Y})$, we can procedurally define and select a sequence of time points in $R^\star$ where μ-RSOA_ε, had it run exactly on this sequence of time points, would make mistakes at each time point. To that end, let

$$\tilde{t}_1=\min\left\{t\in R^\star:\hat{\mu}_t(S_t^c)\ge\varepsilon\text{ where }\hat{\mu}_t=\mu\text{-RSOA}_\varepsilon\Big(x_t\big|\,\{\}\Big)\right\}$$

be the earliest time point in $R^\star$, where a fresh, unupdated copy of μ-RSOA_ε makes a mistake, if it exists. Given $\tilde{t}_1$, we recursively define $\tilde{t}_i$ for $i>1$ as

$$\tilde{t}_i=\min\left\{t\in R^\star:\hat{\mu}_t(S_t^c)\ge\varepsilon\text{ where }\hat{\mu}_t=\mu\text{-RSOA}_\varepsilon\Big(x_t\big|\,\{\tilde{t}_1,\ldots,\tilde{t}_{i-1}\}\Big)\text{ and }t>\tilde{t}_{i-1}\right\}$$

if it exists. That is, $\tilde{t}_i$ is the earliest timepoint after $\tilde{t}_{i-1}$ in $R^\star$ where μ-RSOA_ε having updated only on the sequence $(x_{\tilde{t}_1},S_{\tilde{t}_1}),...,(x_{\tilde{t}_{i-1}},S_{\tilde{t}_{i-1}}))$ makes a mistake. We stop this process when we reach an iteration where no such time point in $R^\star$ can be found where μ-RSOA_ε makes

a mistake.

Using the definitions above, let $\tilde{t}_1, \tilde{t}_2...$, denote the sequence of timepoints in $R^\star$ selected via this recursive procedure. Define $L^\star = \{\tilde{t}_1, \tilde{t}_2..., \}$ and let $E_{L^\star}$ be the expert parametrized by the set of indices $L^\star$. The expert $E_{L^\star}$ exists because $R^\star$ is a part of the stream that is realizable to $h^\star$ and Equation (B.1) implies that $|L^\star| \leq d_\varepsilon$. By definition of the expert, we have $E_{L^\star}(x_t) = \mathrm{RSOA}_\varepsilon(x_t \mid L^\star_{<t})$ for all $t \in [T]$. Let us define $\hat{\mu}^\star_t = \mu\text{-RSOA}_\varepsilon(x_t \mid L^\star_{<t})$. Then, we have

$$
\begin{aligned}
&\inf_{E\in\mathcal{E}} \sum_{t\in R^\star} \mathbb{1}\{E(x_t) \notin S_t\} \\
&\leq \sum_{t\in R^\star} \mathbb{1}\{E_{L^\star}(x_t) \notin S_t\} \\
&= \sum_{t\in R^\star} \mathbb{1}\{\mathrm{RSOA}_\varepsilon(x_t \mid L^\star_{<t}) \notin S_t\} \left(\mathbb{1}\{\hat{\mu}^\star_t(S^c_t) < \varepsilon\} + \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) \geq \varepsilon\}\right) \\
&\leq \sum_{t\in R^\star} \mathbb{1}\{\mathrm{RSOA}_\varepsilon(x_t \mid L^\star_{<t}) \notin S_t\}\, \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) < \varepsilon\} + \sum_{t\in R^\star} \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) \geq \varepsilon\} \\
&\leq \sum_{t\in R^\star} \mathbb{1}\{\mathrm{RSOA}_\varepsilon(x_t \mid L^\star_{<t}) \notin S_t\}\, \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) < \varepsilon\} + d_\varepsilon,
\end{aligned}
$$

where the last inequality follows from the definition of $L^\star$ and the fact that $|L^\star| \leq d_\varepsilon$. Since

$$
\mathbb{E}\left[\mathbb{1}\{\mathrm{RSOA}_\varepsilon(x_t \mid L^\star_{<t}) \notin S_t\}\, \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) < \varepsilon\}\right] = \hat{\mu}^\star_t(S^c_t)\, \mathbb{1}\{\hat{\mu}^\star_t(S^c_t) < \varepsilon\} \leq \varepsilon,
$$

we obtain

$$
\mathbb{E}\left[\inf_{E\in\mathcal{E}} \sum_{t\in R^\star} \mathbb{1}\{E(x_t) \notin S_t\}\right] \leq \varepsilon |R^\star| + d_\varepsilon \leq \varepsilon T + d_\varepsilon.
$$

Finally, plugging this bound in Equation (B.3) yields

$$
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{P}(x_t) \notin S_t\}\right] \leq \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \notin S_t\} + d_\varepsilon + \varepsilon T + \sqrt{2 d_\varepsilon T \ln(T)}.
$$

Since $\varepsilon > 0$ is arbitrary, this completes our proof. ■

B.4.2 Agnostic Lowerbound

Proof. (of (ii) in Theorem 18) Let $d = \mathrm{SL}_2(\mathcal{H})$ and $d_\gamma = \mathrm{MS}_\gamma(\mathcal{H})$ for $\gamma \in [0, 1]$. The lowerbound of $\sup_{\gamma>0} \gamma\, d_\gamma$ on the expected regret in the agnostic setting follows trivially from the lowerbound on the expected cumulative loss in the realizable setting (see (ii) in Theorem 17). Moreover, when $\sup_{\gamma>0} d_\gamma = 0$, there is no non-negative lowerbound on the expected regret. Indeed, consider the case where $\mathcal{Y} = [5]$, $\mathcal{S}(\mathcal{Y}) = \{\{3, 4\}, \{4, 5\}\})$, and

$\mathcal{H} = \{h_1, h_2\}$, where h_i is a constant hypothesis that always outputs i. Then, $\sup_{\gamma>0} d_\gamma = 0$ trivially. However, the expected regret of the algorithm that always outputs 4 is $-T$.

Next, we will focus on showing how the lowerbound of $\sqrt{\frac{dT}{8}}$ can be obtained. When $d = 0$, the claimed lowerbound is $\max\left\{\sqrt{dT/8}\,, \sup_{\gamma>0}\, d_\gamma\right\} = \sup_{\gamma>0}\, \gamma\, d_\gamma$, which we have already established. Let $d > 0$ and $\mathcal{T}$ be a SL_2 tree of depth d shattered by $\mathcal{H}$. With a binary tree $\mathcal{T}$, we now use the technique from Ben-David et al. [2009] to obtain the aforementioned lowerbound.

Consider $T = k\,d$ for some odd $k \in \mathbb{N}$. For $\sigma \in \{\pm 1\}^T$, define $\tilde{\sigma}_i = \operatorname{sign}\left(\sum_{t=(i-1)k+1}^{ik} \sigma_t\right)$ for all $i \in \{1, 2 \ldots, d\}$. Note that the sequence $(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_d)$ gives a path down the tree $\mathcal{T}$. The game proceeds as follows. The adversary samples a string $\sigma \in \{\pm 1\}^T$ uniformly at random and generates a sequence of labeled instances $(x_1, S_1), \ldots (x_T, S_T)$ such that for all $i \in \{1, 2, \ldots, d\}$ and all $t \in \{(i-1)k+1, \ldots, ik\}$, we have $x_t = \mathcal{T}_i(\tilde{\sigma}_{<i})$ and $S_t = f_i((\tilde{\sigma}_{<i}, \sigma_t))$. That is, on round $t \in \{(i-1)k+1, \ldots, ik\}$, the adversary always reveals the instance $\mathcal{T}_i(\tilde{\sigma}_{<i})$ but alternates between revealing the sets labeling the left and right outgoing edges from $\mathcal{T}_i(\tilde{\sigma}_{<i})$ depending on σ_t.

Let $\mathcal{A}$ be any randomized online learner. Then, for each block $i \in [d]$, we have

$$\mathbb{E}\left[\sum_{t=(i-1)k+1}^{ik} \mathbb{1}\left\{\mathcal{A}(x_t) \notin S_t\right\}\right] \geq \sum_{t=(i-1)k+1}^{ik} \frac{1}{2} = \frac{k}{2}.$$

The inequality above holds because S_t is chosen uniformly at random from two disjoint sets $f_i((\tilde{\sigma}_{<i}, -1))$ and $f_i((\tilde{\sigma}_{<i}, +1))$, so the expected loss of any randomized algorithm is at least $1/2$.

Let $h_{\tilde{\sigma}}$ be the hypothesis at the end of the path $(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_d)$ in $\mathcal{T}$. For each block $i \in [d]$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=(i-1)k+1}^{ik} \mathbb{1}\left\{h_{\tilde{\sigma}}(x_t) \notin S_t\right\}\right] = \mathbb{E}\left[\sum_{t=(i-1)k+1}^{ik} \mathbb{1}\{\tilde{\sigma}_i \neq \sigma_t\}\right] &= \frac{k}{2} - \frac{1}{2}\,\mathbb{E}\left[\sum_{t=(i-1)k+1}^{ik} \tilde{\sigma}_i\,\sigma_t\right] \\
&= \frac{k}{2} - \frac{1}{2}\,\mathbb{E}\left[\left|\sum_{t=(i-1)k+1}^{ik} \sigma_j\right|\right] \\
&\leq \frac{k}{2} - \sqrt{\frac{k}{8}},
\end{aligned}$$

where the final step follows upon using Khinchine's inequality [Cesa-Bianchi and Lugosi,

2006, Page 364]. Combining these two bounds above, we obtain

$$\mathbb{E}\left[\sum_{t=(i-1)k+1}^{ik} \mathbb{1}\left\{\mathcal{A}(x_t) \notin S_t\right\} - \sum_{t=(i-1)k+1}^{ik} \mathbb{1}\left\{h_{\tilde{\sigma}}(x_t) \notin S_t\right\}\right] \geq \sqrt{\frac{k}{8}}.$$

Summing this inequality over d blocks, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(x_t) \notin S_t\right\} - \inf_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left\{h(x_t) \notin S_t\right\}\right] \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{\mathcal{A}(x_t) \notin S_t\right\} - \sum_{t=1}^{T} \mathbb{1}\left\{h_{\tilde{\sigma}}(x_t) \notin S_t\right\}\right] \\
&\geq d\sqrt{\frac{k}{8}} = \sqrt{\frac{dT}{8}}.
\end{aligned}$$

which completes our proof. ■

B.5 Applications

B.5.1 Online Multilabel Ranking

In this section, we prove Lemma 7, establishing lower and upperbounds on Helly numbers of permutation sets. Before we prove Lemma 7, we define some new notation. For any bit string $r \in \mathcal{R}$, let $P(r) := \{i : r^i = 1\}$ and let $|r| := |P(r)|$ denote the number of 1's. Given two bit strings r_1, r_2 where $|r_1| \geq |r_2|$, we say that $r_2 \subseteq r_1$ iff $P(r_2) \subseteq P(r_1)$. The following property will also be useful. Let $r_1, r_2 \in \mathcal{R}$ and without loss of generality suppose $|r_1| \geq |r_2|$. If $\mathcal{Y}(r_1) \cap \mathcal{Y}(r_2) \neq \emptyset$ then $r_2 \subseteq r_1$. To prove the contraposition, suppose that $r_2 \nsubseteq r_1$. Then, there exist an index $j \in [K]$ such that $r_2^j = 1$ but $r_1^j = 0$. Thus, every permutation in $\mathcal{Y}(r_2)$ ranks label j in the top $|r_2|$, but every permutation in $\mathcal{Y}(r_1)$ ranks label j outside the top $|r_1|$. That is, for all $\pi_2 \in \mathcal{Y}(r_2)$ we have $\pi_2^j \leq |r_2|$ but for all $\pi_1 \in \mathcal{Y}(r_1)$, we have $\pi_1^j > |r_1|$. Since $|r_2| \leq |r_1|$, we have $\mathcal{Y}(r_1) \cap \mathcal{Y}(r_2) = \emptyset$. We are now ready to prove the main claim. At a high-level, our proof exploits the fact that if we have a sequence of bit strings such that $r_Q \subseteq r_{Q-1} \subseteq ... \subseteq r_1$, then we can iteratively construct a permutation that lies in all $\mathcal{Y}(r_i)$.

Proof. (of Lemma 7) Let $Q \geq 2$ and let $\{r_i\}_{i=1}^{Q} \subseteq \mathcal{R}$ be a sequence of bit strings. It suffices to show that if for all $i, j \in [Q]$ we have $\mathcal{Y}(r_i) \cap \mathcal{Y}(r_j) \neq \emptyset$, then we have $\bigcap_{i\in[Q]} \mathcal{Y}(r_i) \neq \emptyset$. Without loss of generality, suppose $\{r_i\}_{i=1}^{Q}$ is sorted in increasing order of size. That is, for all $i, j \in [Q]$ such that $i > j$, we have $|r_i| \geq |r_j|$. Then, by the property above, for all

$i, j \in [Q]$ where $i > j$ we have $r_j \subseteq r_i$. We now construct a permutation $\pi : [K] \to [K]$ such that for all $i \in [Q]$, we have $\pi \in \mathcal{Y}(r_i)$.

For every $i \in \{2, ..., Q\}$, let $\phi_i : P(r_i) \setminus P(r_{i-1}) \to [|r_i|] \setminus [|r_{i-1}|]$ denote an arbitrary bijective mapping from $P(r_i) \setminus P(r_{i-1})$ to $[|r_i|] \setminus [|r_{i-1}|]$. For $i = 1$, let $\phi_1 : P(r_1) \to [|r_1|]$ be a bijective mapping from $P(r_1)$ to $[|r_1|]$. Finally, let $\phi_{Q+1} : [K] \setminus P(r_Q) \to [K] \setminus [|r_Q|]$ denote an arbitrary bijective mapping from $[K] \setminus P(r_Q)$ to $[K] \setminus [|r_Q|]$. Note that by definition, for all $i, j \in \{1, ..., Q+1\}$, the image space of ϕ_i and ϕ_j are disjoint. Moreover, the union of the image space across all bijective mappings ϕ_i's is $[K]$. Accordingly, we now use these bijective mappings to construct a permutation $\pi \in \mathcal{Y}$. In particular, let π be the permutation such that for all $j \in P(r_1)$, we have $\pi^j = \phi_1(j)$, for all $i \in \{2, ..., Q\}$ and $j \in P(r_i) \setminus P(r_{i-1})$, we have $\pi^j = \phi_i(j)$, and for all $j \in [K] \setminus P(r_Q)$ we have $\pi^j = \phi_{Q+1}(j)$. We now need to show that for all $i \in [Q]$, $\pi \in \mathcal{Y}(r_i)$.

Fix an $i \in \mathcal{Q}$ and consider r_i. It suffices to show that for all $j \in P(r_i)$, we have $\pi^j \leq |r_i|$. That is, π ranks the labels in $P(r_i)$ in the top $|r_i|$. By the subset property, we have

$$P(r_i) = P(r_1) \cup \bigcup_{j=2}^{i} P(r_j) \setminus P(r_{j-1}).$$

Consider some $p \in P(r_i)$. Then, by the equality above, either $p \in P(r_1)$ or $p \in \bigcup_{j=2}^{i} P(r_j) \setminus P(r_{j-1})$. Suppose $p \in P(r_1)$, then by definition $\pi^p = \phi_1(p) \in [|r_1|]$ and therefore $\pi^p \leq |r_i|$. Suppose $p \in \bigcup_{j=2}^{i} P(r_j) \setminus P(r_{j-1})$. In particular, suppose $p \in P(r_j) \setminus P(r_{j-1})$ for some $Q \geq j > 1$. Then by definition, $\pi^p = \phi_j(p) \in [|r_j|] \setminus [|r_{j-1}|]$ and therefore $\pi^p \leq |r_i|$ since $|r_j| \leq |r_i|$. This shows that for every $j \in P(r_i)$, π ranks j in the top $|r_i|$ and therefore $\ell_{0\text{-}1}(\pi, r_i) = 0$. Since $i \in [Q]$ is arbitrary, this completes the proof as we have shown that $\bigcap_{i=1}^{Q} \mathcal{Y}(r_i) \neq \emptyset$. ■

B.5.2 Ranking Littlestone dimension

We end this section by defining an equivalent, arguably more natural, dimension that provides a tight quantitative characterization of online multilabel ranking learnability under binary relevance score feedback. The key insight is that we can actually label the edges in the SL_2 tree with bit strings instead of sets from $\mathcal{S}(\mathcal{Y})$. This intuition leads to the following dimension for online multilabel ranking.

Definition 39 (Ranking Littlestone dimension)**.** *Let $\mathcal{T}$ be a complete $\mathcal{X}$-valued binary tree of depth d. The tree $\mathcal{T}$ is shattered by $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ if there exists a sequence $(f_1, ..., f_d)$ of edge-labeling functions $f_t : \{\pm 1\}^t \to \mathcal{R}$ such that for every path $\sigma = (\sigma_1, ..., \sigma_d) \in \{\pm 1\}^d$, there exists a hypothesis $h_\sigma \in \mathcal{H}$ such that for all $t \in [d]$, $\ell_{0\text{-}1}(h_\sigma(\mathcal{T}_t(\sigma_{<t})), f_t(\sigma_{\leq t})) = 0$,*

but $f_t((\sigma_{<t}, +1)) \not\subseteq f_t((\sigma_{<t}, -1))$ *and* $f_t((\sigma_{<t}, -1)) \not\subseteq f_t((\sigma_{<t}, +1))$*. The Ranking Littlestone dimension of* $\mathcal{H}$*, denoted* $\mathrm{RL}(\mathcal{H}, \mathcal{S}(\mathcal{Y}))$*, is the maximal depth of a tree* $\mathcal{T}$ *that is shattered by* $\mathcal{H}$*. If there exists shattered trees of arbitrarily large depth, we say* $\mathrm{RL}(\mathcal{H}, \mathcal{S}(\mathcal{Y})) = \infty$*.*

Since bit strings map one-to-one with sets in $\mathcal{S}(\mathcal{Y})$, $r_1 \not\subseteq r_2, r_2 \not\subseteq r_1$ iff $\mathcal{Y}(r_1) \cap \mathcal{Y}(r_2) = \emptyset$, and $\ell_{0\text{-}1}(\pi, r) = 0$ iff $\pi \in \mathcal{Y}(r)$, it follows that $\mathrm{SL}_2(\mathcal{H}) = \mathrm{RL}(\mathcal{H})$. Corollary 2 immediately shows that $\mathrm{RL}(\mathcal{H})$ provides a tight quantitative characterization of online multilabel ranking learnability in both the realizable and agnostic settings.

B.5.3 Online Multilabel Classification

Lemma 14 (Helly Number of Hamming Balls)**.** *Let* $\mathcal{Y} = \{0,1\}^K$ *and* $\mathcal{S}_q(\mathcal{Y}) = \{\mathcal{B}(y, q) : y \in \mathcal{Y}\}$*. Then, for all* $q \in [K-1]$*, we have*

$$2^{q+1} \leq \mathrm{H}(\mathcal{S}_q(\mathcal{Y}) \leq \sum_{r=0}^{q} \binom{K}{r} + 1.$$

Proof. (of Lemma 14) Fix $q \in [K-1]$ and let $\mathcal{S}_q(\mathcal{Y}) = \{\mathcal{B}(y, q) : y \in \mathcal{Y}\}$. To see the upperbound, observe that for any bit string $b_1 \in \{0,1\}^K$, there are $\sum_{r=0}^{q} \binom{K}{r}$ sets in $\mathcal{S}_q(\mathcal{Y})$ which contain b_1. This follows from the fact that $b_1 \in \mathcal{B}(b_2, q)$ if and only if $b_2 \in \mathcal{B}(b_1, q)$. Therefore, $|\{A \in \mathcal{S}_q(Y) : b_1 \in A\}| = |\mathcal{B}(b_1, q)| = \sum_{r=0}^{q} \binom{K}{r}$. The upperbound on $\mathrm{H}(\mathcal{S}_q(\mathcal{Y}))$ then follows from the fact that every sequence of sets of size at least $\sum_{r=0}^{q} \binom{K}{r} + 1$ must have an empty intersection.

To establish the lowerbound, it suffices to construct a family of 2^{q+1} Hamming balls that have an empty intersection, but every subfamily of size $2^{q+1} - 1$ has a common element. Let $S = \{y_1, \ldots, y_{2^{q+1}}\} \subset \{0,1\}^K$ be a family of bitstrings that embeds a hypercube of size $q+1$ and is 0 everywhere else. That is, there exists a set of indices $I \subset [K]$ of size $|I| = q+1$ such that $S_{|I} = \{0,1\}^{q+1}$ and $S_{|[K]\setminus I} = 00\ldots00$, where $S_{|I}$ denotes the restriction of bitstrings in S to indices in I. We will first show that

$$\bigcap_{i=1}^{2^{q+1}} B(y_i, q) = \emptyset.$$

To see why this is true, pick a $y \in \{0,1\}^K$. Since S embeds a boolean cube in I, there exists $i, j \in [2^{q+1}]$ such that $y_{|I} = y_{i\,|I}$ and $\neg y_{|I} = y_{j\,|I}$, where $\neg y$ is obtained by flipping every bit in y. Given that $|I| = q+1$, we have $\ell_H(y, y_j) \geq q+1$ and thus $y \notin B(y_j, q)$. Since $y \in \{0,1\}^K$ is arbitrary, $\bigcap_{i=1}^{2^{q+1}} B(y_i, q) = \emptyset$.

Next, we will show that for every $j \in [2^{q+1}]$, we have

$$\bigcap_{i \neq j} B(y_i, q) \neq \emptyset.$$

For each $y_j \in S$, define $\tilde{y}_j \in \{0,1\}^K$ such that $\tilde{y}_{j\,|I} = \neg y_{j\,|I}$ and $\tilde{y}_{j\,|[K]\setminus I} = 00\ldots00 = y_{j\,|[K]\setminus I}$. Recall that a ball of radius q centered at a vertex v of a $q+1$ dimensional boolean cube contains all vertices except $\neg v$. Thus, $y_i \in B(\tilde{y}_j, q)$ for all $i \neq j$. Therefore, $\tilde{y}_j \in \bigcap_{i \neq j} B(y_i, q)$, completing our proof. ■

As a result of Lemma 14, we do not generally have that $\mathrm{SL}(\mathcal{H}) = \mathrm{SL}_2(\mathcal{H})$. Accordingly, unlike multilabel ranking, the quantitative lowerbound implied by Theorem 18 does not immediately follow from the structural properties in Theorem 14. Instead, Lemma 15 shows that when K is sufficiently large, we are guaranteed that $\mathrm{SL}_2(\mathcal{H}) > 0$ for any non-trivial hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and thus the lowerbound of Theorem 18 still gives us a meaningful lowerbound scaling with T.

Lemma 15 (Lowerbound on $\mathrm{SL}_2(\mathcal{H})$)**.** *Fix $q \in \mathbb{N}$ and $K \geq 2q+1$. Let $\mathcal{Y} = \{0,1\}^K$, $\mathcal{S}_q(\mathcal{Y}) = \{\mathcal{B}(y,q) : y \in \mathcal{Y}\}$, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class such that $|\mathcal{H}| \geq 2$. Then,* $\mathrm{SL}_2(\mathcal{H}) \geq 1$.

Proof. (of Lemma 15) Suppose $K \geq 2q+1$ and $|\mathcal{H}| \geq 2$. Then, there exists a $x \in \mathcal{X}$ and a pair of hypothesis $h_1, h_2 \in \mathcal{H}$ such that $h_1(x) \neq h_2(x)$. Our goal will be to construct a shattered SL_2 tree of depth one according to Definition 10 with the root node being labeled by x. To do so, it suffices to find two disjoint balls $S_1, S_2 \in \mathcal{S}_q(\mathcal{Y})$ such that $h_1(x) \in S_1$ and $h_2(x) \in S_2$. We can then label the left and right outgoing edge from x by S_1 and S_2 respectively.

Let p denote the number of indices where $h_1(x)$ and $h_2(x)$ disagree. Note that since $h_1(x) \neq h_2(x)$, we have $p \geq 1$. Let $J \subset [K]$, $|J| = 2q+1-p$ denote an arbitrary subset of the indices where $h_1(x)$ and $h_2(x)$ *agree*. If $2q+1-p$ is even, partition J into two equally sized parts J_1 and J_2. If $2q+1-p$ is odd, partition J into J_1 and J_2 such that $|J_1| - |J_2| = 1$. For every index in J_1 flip the bit in the corresponding position in $h_1(x)$. Let $y_1 \in \mathcal{Y}$ be the bit string resulting from this operation. Likewise, for every index in J_2, flip the bit in the corresponding position in $h_2(x)$. Let $y_2 \in \mathcal{Y}$ denote the resulting bitstring. We now claim that the balls $B(y_1, q), B(y_2, q) \in \mathcal{S}_q(\mathcal{Y})$ satisfy the aforementioned properties.

First, we show that $B(y_1, q) \cap B(y_2, q) = \emptyset$. By construction, y_1 and y_2 differ in the locations where $h_1(x)$ and $h_2(x)$ differ plus all the indices in J. Thus, $\ell_H(y_1, y_2) \geq 2q+1$. Finally, we show that $h_1(x) \in B(y_1, q)$ and $h_2(x) \in B(y_2, q)$. By construction of y_1 and y_2 and the fact that $p \geq 1$, we get that $\ell_H(h_1(x), y_1) \leq \left\lceil \frac{2q+1-p}{2} \right\rceil \leq q$ and $\ell_H(h_2(x), y_2) \leq$

$\left\lceil \frac{2q+1-p}{2} \right\rceil \leq q$. Accordingly, we have that $h_1(x) \in B(y_1, q)$ and $h_2(x) \in B(y_2, q)$ as needed. This completes the proof as we have given two disjoint balls, $B(y_1, q)$ and $B(y_2, q)$, such that $h_1(x) \in B(y_1, q)$ and $h_2(x) \in B(y_2, q)$. ■

Combining Lemma 15 and Theorems 16, 17, and 18 gives a quantitative characterization of online multilabel classification in both the realizable and agnostic settings.

Corollary 5 (Quantitative Online Learnability of Multilabel Classification)**.** *Fix $q \in \mathbb{N}$ and let $K \geq 2q+1$. Let $\mathcal{Y} = \{0,1\}^K$, $\mathcal{S}_q(\mathcal{Y}) = \{\mathcal{B}(y,q) : y \in \mathcal{Y}\}$, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class. Then, in the realizable setting,*

$$\frac{\mathrm{SL}_2(\mathcal{H})}{2} \leq \inf_{\mathcal{A}} \mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) \leq \mathrm{SL}(\mathcal{H}).$$

In the agnostic setting,

$$\sqrt{\frac{\mathrm{SL}_2(\mathcal{H})\, T}{8}} \leq \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}) \leq \mathrm{SL}(\mathcal{H}) + \sqrt{2\ \mathrm{SL}(\mathcal{H})\, T \ln(T)}.$$

We leave it as an interesting future direction to get matching upper and lowerbounds for online multilabel classification.

B.5.4 Online Interval Learning

In this section, we expand on Section 3.6 by providing one more application of set learning to a real-valued setting that we term online interval learning. Consider an arbitrary instance space $\mathcal{X}$, a range space $\mathcal{Y} = [-B, B]$ for some $B > 0$, and a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. We study an online supervised model where, in each round $t \in [T]$, the adversary reveals an example x_t, and the learner makes a prediction $\hat{y}_t \in [-B, B]$. The adversary then reveals an interval $[a_t, b_t]$, and the learner suffers the loss $\mathbb{1}\left\{\hat{y}_t \notin [a_t, b_t]\right\}$. This framework models natural scenarios where the ground truth is a range of values instead of a single value. For instance, consider a model that predicts appropriate clothing size using some structural features of a customer. Instead of one fixed size, there is usually a range of sizes that fits the customer. Since any size outside a particular range is not useful to the customer, the notion of 0-1 mistake is more natural than a regression loss. In fact, interval-valued feedback is ubiquitous in experimental fields such as natural science and medicine because of the inherent uncertainty in measurement.

By defining $\mathcal{S}(\mathcal{Y}) = \left\{[a,b] : -B \leq a < b \leq B\right\}$, a qualitative characterization of online interval learnability in terms of $\mathrm{SL}(\mathcal{H})$ and $\mathrm{MS}_\gamma(\mathcal{H})$ follows immediately from Theorems 16 and 18. Thus, in this section, we instead focus on establishing a quantitative characterization of online interval learnability. As in ranking, we start by computing $\mathrm{H}(\mathcal{S}(\mathcal{Y}))$.

Lemma 16 (Helly Number of Intervals)**.** *Let* $\mathcal{S}(\mathcal{Y}) = \Big\{[a, b] \,:\, -B \le a < b \le B\Big\}$*. Then,* $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$*.*

Lemma 16 is a special case of the celebrated Helly's Theorem (see Radon [1921], Eckhoff [1993]). Since $\mathrm{H}(\mathcal{S}(\mathcal{Y})) = 2$, by Theorem 14, we know that for all $\gamma \in [0, \frac{1}{2}]$, $\mathrm{MS}_\gamma(\mathcal{H}) = \mathrm{SL}(\mathcal{H}) = \mathrm{SL}_2(\mathcal{H})$. Therefore the $\mathrm{SL}_2(\mathcal{H})$ characterizes both deterministic and randomized online interval learnability in the realizable setting. Moreover, we can use Theorems 16, 17, and 18 to give Corollary 6, a sharp quantitative characterization of online interval learning in both the realizable and agnostic settings.

Corollary 6 (Online Interval Learnability)**.** *Let* $\mathcal{Y} = [-B, B]$*,* $\mathcal{S}(\mathcal{Y}) = \Big\{[a, b] \,:\, -B \le a < b \le B\Big\}$*, and* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *be a scalar-valued hypothesis class. Then, in the realizable setting,*

$$\frac{\mathrm{SL}_2(\mathcal{H})}{2} \le \inf_{\mathcal{A}} \mathrm{M}_{\mathcal{A}}(T, \mathcal{H}) \le \mathrm{SL}(\mathcal{H}).$$

In the agnostic setting,

$$\sqrt{\frac{\mathrm{SL}_2(\mathcal{H})\, T}{8}} \le \inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{H}) \le \mathrm{SL}(\mathcal{H}) + \sqrt{2\ \mathrm{SL}(\mathcal{H})\, T \ln(T)}.$$

APPENDIX C

A Unified Theory of Supervised Online Learnability

C.1 A More General Lower Bound

In Appendix C.4, we derive a lower bound on the expected regret for online learning algorithms that satisfy Definition 14. Here, we show that the same lower bound in Theorem 20 applies to a much larger family of algorithms which can also use the realizations of past plays to make future predictions. The proof is identical except now the adversary computes and uses the "expected" measure that the learner will play on round to traverse down the SM tree. We expand on this below.

In full generality, a randomized learner is a sequence of maps $f_1, f_2, \ldots, f_T$ where $f_1 : \mathcal{X} \to \Pi(\mathcal{Z})$ and $f_t : (\mathcal{X} \times \mathcal{Y})^{t-1} \times \mathcal{Z}^{t-1} \times \mathcal{X} \to \Pi(\mathcal{Z})$. On round t, if the learner's past predictions are $z_1, \ldots, z_{t-1}$, then its prediction on round t is $z_t \sim f_t(x_{1:t-1}, y_{1:t-1}, z_{1:t-1}, x_t)$. Now, we can define the "expected" measure on round t as:

$$
\begin{aligned}
&g_t(x_{1:t-1}, y_{1:t-1}, x_t) \\
&:= \mathop{\mathbb{E}}_{z_1 \sim f_1(x_1)} \left[\mathop{\mathbb{E}}_{z_2 \sim f_2(x_1, y_1, z_1, x_2)} \left[\cdots \mathop{\mathbb{E}}_{z_{t-1} \sim f_{t-1}(x_{1:t-2}, y_{1:t-2}, z_{1:t-2}, x_{t-1})} \left[f_t(x_{1:t-1}, y_{1:t-1}, z_{1:t-1}, x_t) \right] \right] \right].
\end{aligned}
$$

Note that $g_t(x_{1:t-1}, y_{1:t-1}, x_t)$ is only a function of the data stream $(x_1, y_1), \ldots, (x_T, y_T)$ and so it can be computed by the adversary before the game begins. Moreover, the expected regret can be written in terms of the "expected" measures, and so our lower bounds applies to this learning algorithm if the adversary uses g_t's to traverse down the SM tree as in the proof in Appendix C.4.

C.2 Proof of Theorem 19

In this section, we show that SMdim reduces to existing combinatorial dimensions. We start with Lemma 17, which shows that SMdim $\equiv$ Ldim.

Lemma 17 (SMdim $\equiv$ Ldim)**.** *Let $\mathcal{Y} = \mathcal{Z}$, $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, and $\ell(y, z) = \mathbb{1}\{y \neq z\}$. Then, $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{L}(\mathcal{H})$ for all $\gamma \in [0, \frac{1}{2}]$.*

Proof. Fix $\gamma \in (0, \frac{1}{2}]$. We first show that $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{L}(\mathcal{H})$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $\Pi(\mathcal{Z})$-ary tree of depth $d = \mathrm{SM}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$. Let v be the root node of $\mathcal{T}$ and x denote the instance labeling the node. Recall that v has an outgoing edge for each measure $\mu \in \Pi(\mathcal{Z})$. Let $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$ be the set of elements in $\mathcal{Y}$ that label the outgoing edges from v. We first claim that there at least two distinct elements in the set $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$. For the sake of contradiction, suppose this is not the case. That is, there is only one distinct element that labels the outgoing edges from v. Let y denote the element that labels the outgoing edges from v. That is, $y_\mu = y$ for all $\mu \in \Pi(\mathcal{Z})$. Consider the Dirac measure δ_y that puts all mass on y. Note that $\delta_y \in \Pi(\mathcal{Z})$ and therefore there exists an outgoing edge from v indexed by δ_y and labeled by y. However, it must be the case that $\mathbb{P}_{z \sim \delta_y}[y \neq z] = 0$. Since $\gamma > 0$, the shattering condition required by Definition 20 cannot be met, which is a contradiction. Accordingly, there is at least two distinct elements in the set $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$.

Let y_{-1}, y_{+1} be the distinct elements of the set $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$, and μ_{-1}, μ_{+1} be the indices of the edges labeled by y_{-1} and y_{+1} respectively. Let $\mathcal{H}_{-1} = \{h_\mu : \mu \in \Pi(\mathcal{Z})^d, \mu_1 = \mu_{-1}\}$ denote the set of shattering hypothesis that corresponds to following a path down $\mathcal{T}$ that takes the outgoing edge indexed μ_{-1} from the root node. Likewise define $\mathcal{H}_{+1}$. Keep the edges indexed by μ_{-1} and μ_{+1} and remove all other outgoing edges along with their corresponding subtree. Reindex the two edges using $\{\pm 1\}$. The root node v should now have two outgoing edges, indexed by $\{\pm 1\}$ and labeled by distinct elements of $\mathcal{Y}$, matching the first constraint of a Littlestone tree. As for the second constraint, observe that for all $h_{-1} \in \mathcal{H}_{-1}$ and $h_{+1} \in \mathcal{H}_{+1}$ the shattering condition from Definition 20 implies that $\mathbb{P}_{z \sim \mu_{-1}}[y_{-1} \neq z] \geq \mathbb{1}\{y_{-1} \neq h_{-1}(x)\} + \gamma$ and $\mathbb{P}_{z \sim \mu_{+1}}[y_{+1} \neq z] \geq \mathbb{1}\{y_{+1} \neq h_{+1}(x)\} + \gamma$. However, this can only be true if both $\mathbb{1}\{y_{-1} \neq h_{-1}(x)\} = 0 \implies y_{-1} = h_{-1}(x)$ and $\mathbb{1}\{y_{+1} \neq h_{+1}(x)\} = 0 \implies y_{+1} = h_{+1}(x)$. Accordingly, the hypotheses that shatters the edges indexed by μ_{-1} and μ_{+1} in the original tree according to Definition 20 also shatters the newly re-indexed edges according to Definition 16. Recursively repeating the above procedure on the subtrees following the two reindexed edges results in a Littlestone tree shattered by $\mathcal{H}$ of depth d. Thus, $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{L}(\mathcal{H})$ for $\gamma \in (0, \frac{1}{2}]$. The case when $\gamma = 0$ follows similarly and uses the fact that when $\gamma = 0$, we define the shattering condition in SMdim with a strict inequality (see last sentence in Definition 20).

We now prove the inequality that $\mathrm{SM}_\gamma(\mathcal{H}) \geq \mathrm{L}(\mathcal{H})$. Fix $\gamma \in [0, \frac{1}{2}]$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $\{\pm 1\}$-ary tree of depth $d = \mathrm{L}(\mathcal{H})$ shattered by $\mathcal{H}$ according to Definition 16. Our goal will be to expand $\mathcal{T}$ into a $\Pi(\mathcal{Z})$-ary tree that is γ-shattered by $\mathcal{H}$ according to Definition 20. Let v be the root node of $\mathcal{T}$, x be the instance that labels the root node, and y_{-1}, y_{+1} denote

the distinct elements of $\mathcal{Y}$ that label the left and right outgoing edges from v respectively. Let $\mathcal{H}_{-1} = \{h_\sigma : \sigma \in \{\pm 1\}^d, \sigma_1 = -1\} \subset \mathcal{H}$ denote the set of shattering hypothesis that correspond to following a path down $\mathcal{T}$ that takes the edge indexed by -1 in the first level. Define $\mathcal{H}_{+1}$ analogously. Then, for all $h_{-1} \in \mathcal{H}_{-1}$ and $h_{+1} \in \mathcal{H}_{+1}$, the shattering condition implies that $h_{-1}(x) = y_{-1}$ and $h_{+1}(x) = y_{+1}$.

For every measure $\mu \in \Pi(\mathcal{Z})$, we claim that there exists a $\sigma_\mu \in \{\pm 1\}$ such that $\mathbb{P}_{z\sim\mu}\left[y_{\sigma_\mu} \neq z\right] = \mu(\{y_{\sigma_\mu}\}^c) \geq \gamma$. Suppose for the sake of contradiction that this is not true. Then, there exists a measure $\mu \in \Pi(\mathcal{Z})$ such that for both $\sigma \in \{\pm 1\}$, we have $\mu(\{y_\sigma\}^c) < \gamma$. Then, $1 = \mu(\mathcal{Z}) = \mu(\{y_{-1}\}^c \cup \{y_{+1}\}^c) < 2\gamma < 1$, a contradiction. Thus, for every measure $\mu \in \Pi(\mathcal{Z})$ there exists a $\sigma_\mu \in \{\pm 1\}$ such that $\mathbb{P}_{z\sim\mu}\left[y_{\sigma_\mu} \neq z\right] \geq \gamma$. Combining this with the fact that for any $h_{-1} \in \mathcal{H}_{-1}$ and $h_{+1} \in \mathcal{H}_{+1}$, we have $y_{-1} = h_{-1}(x)$ and $y_{+1} = h_{+1}(x)$, gives that, for every measure $\mu \in \Pi(\mathcal{Z})$, there exists a $\sigma_\mu \in \{\pm 1\}$ such that for all $h_{\sigma_\mu} \in \mathcal{H}_{\sigma_\mu}$,we have $\mathbb{P}_{z\sim\mu}\left[y_{\sigma_\mu} \neq z\right] \geq \mathbb{1}\{y_{\sigma_\mu} \neq h_{\sigma_\mu}(x)\} + \gamma$. Note that if we take y_{σ_μ} to be the label on an edge indexed by μ, then the inequality above matches the shattering condition required by Definition 20.

To that end, for every measure $\mu \in \Pi(\mathcal{Z})$, add an outgoing edge from v indexed by μ and labeled by the y_{σ_μ}, where σ_μ is the index as promised by the analysis above. Take the sub-tree in $\mathcal{T}$ following the original outgoing edge from v indexed by σ_μ, and append it to the newly constructed outgoing edge from v indexed by μ. Remove the original outgoing edges from v indexed by $\{\pm 1\}$ and their corresponding subtrees. Recursively repeat the above procedure on the subtrees following the newly created edges indexed by measures. Upon repeated this process for every internal node in $\mathcal{T}$, we obtain a $\Pi(\mathcal{Z})$-ary tree that is γ-shattered by $\mathcal{H}$ of depth d. Thus, we have that $\mathrm{L}(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H})$ for $\gamma \in [0, \frac{1}{2}]$. ■

Next, we show an equivalence between SMdim and seq-fat.

Lemma 18 (SMdim ≡ seq-fat)**.** *Let $\mathcal{Y} = \mathcal{Z} = [-1, 1]$, $\mathcal{H} \subseteq \mathcal{Z}^\mathcal{X}$, and $\ell(y, z) = |y - z|$. Then for every $\gamma \in (0, 1]$ and $\gamma' < \gamma$,*

$$\mathrm{sfat}_\gamma(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{sfat}_{\gamma'}(\mathcal{H}).$$

Proof. We first prove the upper bound. Let $\gamma \in (0, 1]$ and $\gamma' < \gamma$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $\Pi(\mathcal{Z})$-ary tree of depth $d = \mathrm{SM}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$. Let v be the root node of $\mathcal{T}$ and x denote the instance labeling the node. Recall that v has an outgoing edge for each measure $\mu \in \Pi(\mathcal{Z})$. In particular, this means that v has outgoing edges corresponding to the Dirac measures on $\mathcal{Z}$, which we denote by $\{\delta_z\}_{z\in\mathcal{Z}}$. Fix a $z \in \mathcal{Z}$ and consider the outgoing edge from v indexed by δ_z. Let $y_z \in \mathcal{Y}$ be the element that labels the outgoing edge indexed by δ_z. Let $\mathcal{H}_z = \{h_\mu : \mu \in \Pi(\mathcal{Z})^d, \mu_1 = \delta_z\} \subset \mathcal{H}$ denote the set of shattering hypothesis that

corresponds to following a path down $\mathcal{T}$ that takes the edge δ_z in the root node. Then, for all $h \in \mathcal{H}_z$ the shattering condition from Definition 20 implies that

$$|z - y_z| \geq |h(x) - y_z| + \gamma > |h(x) - y_z| + \gamma'.$$

Taking the supremum on both sides, gives that:

$$|z - y_z| > \sup_{h \in \mathcal{H}_z} |h(x) - y_z| + \gamma' = r_z + \gamma'. \tag{C.1}$$

where we let $r_z = \sup_{h \in \mathcal{H}_z} |h(x) - y_z|$. Let $I_z := [y_z - (r_z + \gamma'), y_z + (r_z + \gamma')] \subset [-3, 3]$ denote an interval corresponding to z. Inequality (C.1) above implies that $z \notin I_z$ (note that I_z changes depending on z). Since $z \in \mathcal{Z}$ was arbitrary, it must be the case that $z \notin I_z$ for all $z \in \mathcal{Z}$. This means that $\bigcap_{z \in \mathcal{Z}} I_z = \emptyset$. Since $[-3, 3]$ is compact and $\{I_z\}_{z \in \mathcal{Z}}$ is a family of closed intervals whose intersection is empty, the celebrated Helly's theorem states that there exists two intervals in $\{I_z\}_{z \in \mathcal{Z}}$ that are disjoint [Eckhoff, 1993, Radon, 1921]. Accordingly, let z_1, z_2 be such that $I_{z_1} \cap I_{z_2} = \emptyset$. As before, let y_{z_1} and y_{z_2} be the labels on the outgoing edges from v indexed by the Dirac measures δ_{z_1} and δ_{z_2} respectively. Without loss of generality, let $y_{z_1} < y_{z_2}$ (we have strict inequality because we are guaranteed that I_{z_1} and I_{z_2} are disjoint). By inequality C.1, for all $h_{z_1} \in \mathcal{H}_{z_1}$ and $h_{z_2} \in \mathcal{H}_{z_2}$ we have that

$$h_{z_1}(x) \in [y_{z_1} - r_{z_1}, y_{z_1} + r_{z_1}] \quad \text{and} \quad h_{z_2}(x) \in [y_{z_2} - r_{z_2}, y_{z_2} + r_{z_2}]\,.$$

Let $s = \frac{y_{z_1} + r_{z_1} + y_{z_2} - r_{z_2}}{2} \in [-1, 1]$ be a witness. Then, for all $h_{z_1} \in \mathcal{H}_{z_1}$ and $h_{z_2} \in \mathcal{H}_{z_2}$, we have that $s - h_{z_1}(x) \geq \gamma'$ and $h_{z_2}(x) - s \geq \gamma'$. Relabel the two edges indexed by δ_{z_1} and δ_{z_2} with the same witness s. Reindex the two edges indexed by δ_{z_1} and δ_{z_2} with -1 and $+1$ respectively. Remove all other edges indexed by measures and their corresponding subtrees. There should now only be two outgoing edges from v, each labeled by the same witness. Next, recall that for all $h_{z_1} \in \mathcal{H}_{z_1}$ and $h_{z_2} \in \mathcal{H}_{z_2}$ we have that $s - h_{z_1} \geq \gamma'$ and $h_{z_2} - s \geq \gamma'$. Accordingly, the hypotheses that shatter the edges indexed by δ_{z_1} and δ_{z_2} in the original tree according to Definition 20 also shatter the newly re-indexed and relabeled edges according to Definition 17. Recursively repeating the above procedure on the subtrees following the two newly reindexed and relabeled edges results in a seq-fat tree γ'-shattered by $\mathcal{H}$ of depth d. Thus, $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{sfat}_{\gamma'}(\mathcal{H})$ for $\gamma' < \gamma$.

We now move on to prove the lower bound. Let $\gamma \in (0, 1]$ and $\mathcal{T}$ be a $\mathcal{X}$-valued, $\{\pm 1\}$-ary tree of depth $d = \mathrm{sfat}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$ according to Definition 17. Our goal will be expand $\mathcal{T}$ into a $\Pi(\mathcal{Z})$-ary tree that is γ-shattered by $\mathcal{H}$ according to Definition 20. Let v be the root node, x the instance that labels the root node, and s be the witness that labels

the two outgoing edges of v. Let $\mathcal{H}_{-1} = \{h_\sigma : \sigma \in \{\pm 1\}^d, \sigma_1 = -1\} \subset \mathcal{H}$ denote the set of shattering hypothesis that corresponds to following a path down $\mathcal{T}$ that takes the outgoing edge indexed by -1 from the root node. Likewise define $\mathcal{H}_{+1}$. Then, for all $h_{-1} \in \mathcal{H}_{-1}$ and $h_{+1} \in \mathcal{H}_{+1}$, the shattering condition implies that $s - h_{-1}(x) \geq \gamma$ and $h_{+1}(x) - s \geq \gamma$ respectively.

For every measure $\mu \in \Pi(\mathcal{Z})$, we claim that there exist a $\sigma_\mu \in \{-1, 1\}$ such that $\mathbb{E}_{z\sim\mu}[|\sigma_\mu - z|] \geq |s - \sigma_\mu|$. Suppose for the sake of contradiction that this is not true. That is, there exists $\mu \in \Pi(\mathcal{Z})$ such that for all $\tau \in \{-1, 1\}$ we have that $\mathbb{E}_{z\sim\mu}[|\tau - z|] < |s - \tau|$. Then, when $\tau = -1$, we have that $\mathbb{E}_{z\sim\mu}[z] < |s + 1| - 1$ and when $\tau = 1$, we have $1 - |s - 1| < \mathbb{E}_{z\sim\mu}[z]$, using the fact that $|\tau - z| = 1 - \tau z$. Combining the two inequalities together and using the fact that $s \in [-1, 1]$ gives that $2 < |s + 1| + |s - 1| = 2$, which is a contradiction. Accordingly, for every measure $\mu \in \Pi(\mathcal{Z})$, there exists a $\sigma_\mu \in \{-1, 1\}$ such that $\mathbb{E}_{z\sim\mu}[|\sigma_\mu - z|] \geq |s - \sigma_\mu|$. Next, crucially note that for any $\tau \in \{\pm 1\}$ and any $h_\tau \in \mathcal{H}_\tau$, we have $|h_\tau(x) - \tau| = |s - \tau| - |h_\tau(x) - s| \leq |s - \tau| - \gamma$ by the seq-fat shattering condition from Definition 17. Therefore, for every measure $\mu \in \Pi(\mathcal{Z})$, there exists $\sigma_\mu \in \{\pm 1\}$ such that for all $h_{\sigma_\mu} \in \mathcal{H}_{\sigma_\mu}$, we have that $\mathbb{E}_{z\sim\mu}[|\sigma_\mu - z|] \geq |\sigma_\mu - h_{\sigma_\mu}(x)| + \gamma$. Note that if we take σ_μ to be the label on a edge indexed by μ, then $\mathbb{E}_{z\sim\mu}[|\sigma_\mu - z|] \geq |\sigma_\mu - h_{\sigma_\mu}(x)| + \gamma$ exactly matches the shattering condition required by Definition 20.

To that end, for every measure $\mu \in \Pi(\mathcal{Z})$, add an outgoing edge from v indexed by μ and labeled by the $\sigma_\mu \in \{\pm 1\}$ promised in the analysis above. Take the sub-tree in $\mathcal{T}$ following the original outgoing edge from v indexed by σ_μ, and append it to the newly constructed outgoing edge from v indexed by μ. Remove the original outgoing edges from v indexed by -1 and $+1$ and their corresponding subtrees. Recursively repeat the above procedure on the subtrees following the newly created edges indexed by measures. Upon repeating this process for every internal node in $\mathcal{T}$, we obtain a $\Pi(\mathcal{Z})$-ary tree that is γ-shattered by $\mathcal{H}$ of depth d. Thus, we have that $\mathrm{sfat}_\gamma(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H})$. ■

Next, we show that SMdim reduces to $(k+1)$-Ldim from Moran et al. [2023].

Lemma 19 (SMdim $\equiv$ $(k+1)$-Ldim)**.** *Let $\mathcal{Z} = \{S : S \subset \mathcal{Y}, |S| \leq k\}$, $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, and $\ell(y, z) = \mathbb{1}\{y \notin z\}$. Then for every $\gamma \in [0, \frac{1}{k+1}]$, we have $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{L}_{k+1}(\mathcal{H})$.*

Proof. Fix $\gamma \in (0, 1]$. We first show that $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{L}_{k+1}(\mathcal{H})$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $\Pi(\mathcal{Z})$-ary tree of depth $d = \mathrm{SM}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$. Let v be the root node of $\mathcal{T}$ and x denote the instance labeling the node. Recall that v has an outgoing edge for each measure $\mu \in \Pi(\mathcal{Z})$. Let $\{y_\mu\}_{\mu\in\Pi(\mathcal{Z})}$ be the set of elements in $\mathcal{Y}$ that label the outgoing edges from v. We first claim that there at least $k + 1$ distinct elements in the set $\{y_\mu\}_{\mu\in\Pi(\mathcal{Z})}$. For the sake of contradiction, suppose this was not the case. That is, there are only k distinct elements

that label the outgoing edges from v. Let $y_1, \ldots, y_k$ denote the k distinct elements that label the outgoing edges from v, Consider the measure $\tilde{\mu}$ that puts all mass on $\{y_1, \ldots, y_k\}$. Note that $\tilde{\mu} \in \Pi(\mathcal{Z})$ and let $\tilde{y} \in \{y_1, \ldots, y_k\}$ be the label on the outgoing edge from v indexed by $\tilde{\mu}$. By definition of $\tilde{\mu}$ and $\tilde{y}$, it must be the case that $\mathbb{P}_{z \sim \tilde{\mu}}[\tilde{y} \notin z] = 0$. Since $\gamma > 0$, the shattering condition required by Definition 20 cannot be met, which is a contradiction. Accordingly, there exists at least $k+1$ distinct elements in the set $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$.

Let $y_1, \ldots, y_{k+1}$ be the distinct elements of the set $\{y_\mu\}_{\mu \in \Pi(\mathcal{Z})}$, and $\mu_1, \ldots, \mu_{k+1}$ be the indices of the edges labeled by $y_1, \ldots, y_{k+1}$ respectively, breaking ties arbitrarily. For $\mu_i \in \{\mu_1, \ldots, \mu_{k+1}\}$, let $\mathcal{H}_{\mu_i}$ denote the set of shattering hypothesis that corresponds to following a path down $\mathcal{T}$ that takes the outgoing edge μ_i from the root node. Keep the edges indexed by $\mu_1, \ldots, \mu_{k+1}$, and remove all other outgoing edges along with their corresponding subtree. Reindex the $k+1$ edges using distinct numbers in $[k+1]$. The root node v should now have $k+1$ outgoing edges, each indexed by a different natural number in $[k+1]$ and labeled by a distinct element of $\mathcal{Y}$, matching the first constraint of a $(k+1)$-Littlestone tree. As for the second constraint, observe that for all $h \in \mathcal{H}_{\mu_i}$ the shattering condition implies that $\mathbb{P}_{z \sim \mu_i}[y_i \notin z] \geq \mathbb{1}\{y_i \notin h(x)\} + \gamma$. However, this can only be true if $\mathbb{1}\{y_i \notin h(x)\} = 0 \implies y_i \in h(x)$. Accordingly, the hypotheses that shatter the edges indexed by $\mu_1, \ldots, \mu_{k+1}$ in the original tree according to Definition 20 also shatter the newly re-indexed edges according to Definition 18. Recursively repeating the above procedure on the subtrees following the $k+1$ reindexed edges results in a $(k+1)$-Littlestone tree shattered by $\mathcal{H}$ of depth d. Thus, $\mathrm{SM}_\gamma(\mathcal{H}) \leq \mathrm{L}_{k+1}(\mathcal{H})$ for $\gamma \in (0, 1]$. The case when $\gamma = 0$ follows similarly and uses the fact that when $\gamma = 0$, we define the shattering condition in SMdim with a strict inequality (see last sentence in Definition 20).

We now prove the inequality that $\mathrm{SM}_\gamma(\mathcal{H}) \geq \mathrm{L}_{k+1}(\mathcal{H})$. Fix $\gamma \in [0, \frac{1}{k+1}]$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $[k+1]$-ary tree of depth $d = \mathrm{L}_{k+1}(\mathcal{H})$ shattered by $\mathcal{H}$ according to Definition 18. Our goal will be to expand $\mathcal{T}$ into a $\Pi(\mathcal{Z})$-ary tree that is γ-shattered by $\mathcal{H}$ according to Definition 20. Let v be the root node of $\mathcal{T}$, x be the instance that labels the root node, and $\{y_i\}_{i=1}^{k+1}$ denote the distinct elements of $\mathcal{Y}$ that label the $k+1$ outgoing edges from v. For each $i \in [k+1]$, let $\mathcal{H}_i = \{h_p : p \in [k+1]^d, p_1 = i\} \subset \mathcal{H}$ denote the set of shattering hypothesis that corresponds to following a path down $\mathcal{T}$ that takes the outgoing edge indexed by i from v. Then, for all $i \in [k+1]$ and $h_i \in \mathcal{H}_i$, the shattering condition implies that $y_i \in h_i(x) \implies \mathbb{1}\{y_i \notin h_i(x)\} = 0$.

For every measure $\mu \in \Pi(\mathcal{Z})$, we claim that there exists a $i_\mu \in [k+1]$ such that $\mathbb{P}_{z \sim \mu}\left[y_{i_\mu} \notin z\right] \geq \gamma$. Suppose for the sake of contradiction that this is not true. Then, there exists a measure $\mu \in \Pi(\mathcal{Z})$ such that for all $i \in [k+1]$, we have $\mathbb{P}_{z \sim \mu}[y_i \notin z] < \gamma$. This

implies that

$$\mathbb{P}_{z\sim\mu}\left[\exists i \in [k+1] \text{ such that } y_i \notin z\right] < (k+1)\gamma < 1.$$

However, since μ is supported over subsets of $\mathcal{Y}$ of size $\leq k$, we have $\mathbb{P}_{z\sim\mu}\left[\exists i \in [k+1] \text{ such that } y_i \notin z\right] = 1$, a contradiction. Thus, for every measure $\mu \in \Pi(\mathcal{Z})$ there exists a $i_\mu \in [k+1]$ such that $\mathbb{P}_{z\sim\mu}\left[y_{i_\mu} \notin z\right] \geq \gamma$. Combining this with the fact that for every $i \in [k+1]$ and $h_i \in \mathcal{H}_i$ we have that $y_i \in h_i(x)$ gives that, for every measure $\mu \in \Pi(\mathcal{Z})$, there exists a $i_\mu \in [k+1]$ such that for all $h_{i_\mu} \in \mathcal{H}_{i_\mu}$, we have $\mathbb{P}_{z\sim\mu}\left[y_{i_\mu} \notin z\right] \geq \mathbb{1}\{y_{i_\mu} \notin h_{i_\mu}(x)\} + \gamma$. Note that if we take y_{i_μ} to be the label on an edge indexed by μ, then the inequality above matches the shattering condition required by Definition 20.

To that end, for every measure $\mu \in \Pi(\mathcal{Z})$, add an outgoing edge from v indexed by μ and labeled by the y_{i_μ}, where i_μ is the index as promised by the analysis above. Take the sub-tree in $\mathcal{T}$ following the original outgoing edge from v indexed by i_μ, and append it to the newly constructed outgoing edge from v indexed by μ. Remove the original outgoing edges from v indexed by numbers in $[k+1]$ and their corresponding subtrees. Recursively repeat the above procedure on the subtrees following the newly created edges indexed by measures. Upon repeating this process for every internal node in $\mathcal{T}$, we obtain a $\Pi(\mathcal{Z})$-ary tree of depth d that is γ-shattered by $\mathcal{H}$. Thus, we have that $\mathrm{L}_{k+1}(\mathcal{H}) \leq \mathrm{SM}_\gamma(\mathcal{H})$ for $\gamma \in [0, \frac{1}{k+1}]$. ■

Finally, we show that the SMdim ≡ MSdim.

Lemma 20 (SMdim ≡ MSdim)**.** *Let $\mathcal{Y} \subset \sigma(\mathcal{Z})$, $\mathcal{H} \subseteq \mathcal{Z}^{\mathcal{X}}$, and $\ell(y,z) = \mathbb{1}\{z \notin y\}$. Then for every $\gamma \in [0,1]$, we have $\mathrm{SM}_\gamma(\mathcal{H}) = \mathrm{MS}_\gamma(\mathcal{H})$.*

Proof. The equality follows directly from the fact that $\mathbb{E}_{z\sim\mu}\left[\ell(y,z)\right] = \mu(y^c)$ and the fact that $\mathbb{E}_{z\sim\mu_t}\left[\ell(z, f_t(\mu_{\leq t}))\right] \geq \ell(h_\mu(\mathcal{T}_t(\mu_{<t})), f_t(\mu_{\leq t})) + \gamma \iff h_\mu(\mathcal{T}_t(\mu_{<t})) \in f_t(\mu_{\leq t})$ and $\mu_t(f_t(\mu_{\leq t})) \leq 1-\gamma$. ■

C.3 Proof of Lemma 8

We now prove that given any target accuracy $\gamma > 0$ and any ε_t-realizable sequence $\{(x_t, (y_t, \varepsilon_t))\}_{t=1}^T$, Algorithm 8 computes distributions $\mu_t \in \Pi(\mathcal{Z})$ such that

$$\sum_{t=1}^{T} \mathbb{1}\{\mathbb{E}_{z\sim\mu_t}\left[\ell(y_t,z)\right] \geq \gamma + \varepsilon_t\} \leq \mathrm{SM}_\gamma(\mathcal{H}).$$

To prove this guarantee, it suffices to show that (i) on any round where $\mathbb{E}_{z_t\sim\mu_t}\left[\ell(y_t,z_t)\right] \geq \gamma + \varepsilon_t$ and $\mathrm{SM}_\gamma(V_{t-1}) > 0$, we have $\mathrm{SM}_\gamma(V_t) \leq \mathrm{SM}_\gamma(V_{t-1}) - 1$, and (ii) if $\mathrm{SM}_\gamma(V_{t-1}) = 0$ there

always exists a distribution $\mu_t \in \Pi(\mathcal{Z})$ such that $\mathbb{E}_{z_t \sim \mu_t}[\ell(y_t, z_t)] < \gamma + \varepsilon_t$.

Let $t \in [T]$ be a round where $\mathbb{E}_{z_t \sim \mu_t}[\ell(y_t, z_t)] \geq \gamma + \varepsilon_t$ and $\mathrm{SM}_\gamma(V_{t-1}) > 0$. For the sake contradiction, suppose that $\mathrm{SM}_\gamma(V_t) = \mathrm{SM}_\gamma(V_{t-1}) = d$. Then, by the min-max computation in Algorithm 8, for every measure $\mu \in \Pi(\mathcal{Z})$, there exists a pair $(y_\mu, \varepsilon_\mu) \in \mathcal{Y} \times [0, c]$ such that $\mathbb{E}_{z \sim \mu}[\ell(y_\mu, z)] \geq \varepsilon_\mu + \gamma$ and $\mathrm{SM}_\gamma(V_{t-1}(y_\mu, \varepsilon_\mu)) = d$. Now construct a tree $\mathcal{T}$ with x_t labeling the root node. For each measure $\mu \in \Pi(\mathcal{Z})$, construct an outgoing edge from x_t indexed by μ and labeled by y_μ. Append the tree of depth d associated with the version space $V_{t-1}(y_\mu, \varepsilon_\mu)$ to the edge indexed by μ. Note that the depth of $\mathcal{T}$ must be $d+1$. Furthermore, observe that for every hypothesis $h \in V_{t-1}(y_\mu, \varepsilon_\mu)$, we have that $\mathbb{E}_{z \sim \mu}[\ell(y_\mu, z)] \geq \ell(y_\mu, h(x_t)) + \gamma$, matching the shattering condition in Definition 20. Therefore, by definition of SMdim, we have that $\mathrm{SM}_\gamma(V_{t-1}) \geq d+1$, a contradiction. Thus, it must be the case that $\mathrm{SM}_\gamma(V_t) \leq \mathrm{SM}_\gamma(V_{t-1}) - 1$.

Now, suppose $t \in [T]$ is a round such that $\mathrm{SM}_\gamma(V_{t-1}) = 0$. We show that there always exist a distribution $\mu_t \in \Pi(\mathcal{Z})$ such that for all $(y, \varepsilon) \in \mathcal{C}_t$, we have $\mathbb{E}_{z_t \sim \mu_t}[\ell(y, z_t)] < \gamma + \varepsilon$. Since we are in the ε_t-realizable setting, it must be the case that $(y_t, \varepsilon_t) \in \mathcal{C}_t$. To see why such a μ_t must exist, suppose for the sake of contradiction that it does not exist. Then, for all $\mu \in \Pi(\mathcal{Z})$, there exists a pair $(y_\mu, \varepsilon_\mu) \in \mathcal{C}_t$ such that $\mathbb{E}_{z \sim \mu}[\ell(y_\mu, z)] \geq \gamma + \varepsilon_\mu$. As before, consider a tree with root node labeled by x_t. For each measure $\mu \in \Pi(\mathcal{Z})$, construct an outgoing edge from x_t indexed by μ and labeled by y_μ. Since $(y_\mu, \varepsilon_\mu) \in \mathcal{C}_t$, there exists a hypothesis $h_\mu \in V_{t-1}$ such that $\ell(y_\mu, h_\mu(x_t)) \leq \varepsilon_\mu$. Therefore, we have $\mathbb{E}_{z \sim \mu}[\ell(y_\mu, z)] \geq \ell(y_\mu, h_\mu(x_t)) + \gamma$. By definition of SMdim, this implies that $\mathrm{SM}_\gamma(V_{t-1}) \geq 1$, which contradicts the fact that $\mathrm{SM}_\gamma(V_{t-1}) = 0$. Thus, there must be a distribution $\mu_t \in \Pi(\mathcal{Z})$ such that for for all $(y, \varepsilon) \in \mathcal{C}_t$, we have $\mathbb{E}_{z \sim \mu_t}[\ell(y, z)] < \gamma + \varepsilon$. Since this is precisely the distribution that Algorithm 8 plays whenever $\mathrm{SM}_\gamma(V_{t-1}) = 0$ and since $\mathrm{SM}_\gamma(V_{t'}) \leq \mathrm{SM}_\gamma(V_{t-1})$ for all $t' \geq t$, the algorithm no longer suffers expected loss more than $\gamma + \varepsilon_{t'}$ for all $t' \geq t$. This completes the proof.

C.4 Proof of lower bound in Theorem 20

We now prove the lower bound in Theorem 20. Fix $\gamma > 0$ and $d_\gamma := \mathrm{SM}_\gamma(\mathcal{H})$. By definition of SMdim, there exists a $\mathcal{X}$-valued, $\Pi(\mathcal{Z})$-ary tree $\mathcal{T}$ of depth d_γ shattered by $\mathcal{H}$. Let $(f_1, \ldots, f_d)$ be the sequence of edge-labeling functions $f_t : \Pi(\mathcal{Z})^t \to \mathcal{Y}$ associated with $\mathcal{T}$. Let $\mathcal{A}$ be any randomized learner for $\mathcal{H}$. Our goal will be to use $\mathcal{T}$ and its edge-labeling functions $(f_1, \ldots, f_d)$ to construct a difficult stream for $\mathcal{A}$ such that on every round, the expected loss of $\mathcal{A}$ is at least γ more than the loss of the optimal hypothesis in hindsight. This stream is obtained by traversing $\mathcal{T}$ adapting to the sequence of distributions output by $\mathcal{A}$.

To that end, for every round $t \in [d_\gamma]$, let μ_t denote the distribution that $\mathcal{A}$ computes

before making its prediction $z_t \sim \mu_t$. Consider the stream $\{(\mathcal{T}_t(\mu_{<t}), f_t(\mu_{\leq t}))\}_{t=1}^{d_\gamma}$, where $\mu = (\mu_1, \ldots, \mu_{d_\gamma})$ denotes the sequence of distributions output by $\mathcal{A}$. This stream is obtained by starting at the root of $\mathcal{T}$, passing $\mathcal{T}_1$ to $\mathcal{A}$, observing the distribution μ_1 computed by $\mathcal{A}$, passing the label $f_t(\mu_{\leq 1})$ to $\mathcal{A}$, and then finally moving along the edge indexed by μ_1. This process then repeats $d_\gamma - 1$ times until the end of the tree $\mathcal{T}$ is reached. Note that we can observe and use the distribution computed by $\mathcal{A}$ on round t to generate the label because $\mathcal{A}$ *deterministically* maps a sequence of labeled instances to a distribution.

Recall that the shattering condition implies that $\exists h_\mu \in \mathcal{H}$ such that $\mathbb{E}_{z_t \sim \mu_t}[\ell(f_t(\mu_{\leq t}), z_t)] \geq \ell(f_t(\mu_{\leq t}), h_\mu(\mathcal{T}_t(\mu_{<t})) + \gamma$ for all $t \in [d_\gamma]$. Therefore, the regret of $\mathcal{A}$ on the stream described above is at least

$$\mathrm{R}_{\mathcal{A}}(T, \mathcal{H}, \ell) \geq \sum_{t=1}^{d_\gamma} \mathop{\mathbb{E}}_{z_t \sim \mu_t}[\ell(f_t(\mu_{\leq t}), z_t)] - \sum_{t=1}^{d_\gamma} \ell(f_t(\mu_{\leq t}), h_\mu(\mathcal{T}_t(\mu_{<t})) \geq \sum_{t=1}^{d_\gamma} \gamma = \gamma d_\gamma.$$

Since our choice of γ and the randomized algorithm $\mathcal{A}$ is arbitrary, this holds true for any $\gamma > 0$ and randomized online learner. This completes our proof.

C.5 Proof of Lemma 9

Let $p = \mathrm{H}(\mathcal{Y}, \mathcal{Z}, \ell)$. Fix $\gamma \in (0, 1]$ and $\gamma' < \gamma$. Let $\mathcal{T}$ be a $\mathcal{X}$-valued, $\Pi(\mathcal{Z})$-ary tree of depth $d = \mathrm{SM}_\gamma(\mathcal{H})$ shattered by $\mathcal{H}$. Let v be the root node of $\mathcal{T}$ and x denote the instance labeling v. Recall that v has an outgoing edge for each measure $\mu \in \Pi(\mathcal{Z})$. In particular, this means that v has outgoing edges corresponding to the Dirac measures on $\mathcal{Z}$, which we denote by $\{\delta_z\}_{z \in \mathcal{Z}}$. Fix a $z \in \mathcal{Z}$ and consider the outgoing edge from v indexed by δ_z. Let $y_z \in \mathcal{Y}$ be the element that labels the outgoing edge indexed by δ_z. Let $\mathcal{H}_{\delta_z} = \{h_\mu : \mu \in \Pi(\mathcal{Z})^d, \mu_1 = \delta_z\} \subset \mathcal{H}$ denote the set of shattering hypothesis that corresponds to following a path down $\mathcal{T}$ that takes the edge δ_z in the first level. Then, for all $h \in \mathcal{H}_\delta$, the shattering condition implies that

$$\ell(y_z, z) \geq \ell(y_z, h(x)) + \gamma > \ell(y_z, h(x)) + \gamma'.$$

Taking the supremum on both sides further gives that

$$\ell(y_z, z) > \sup_{h \in \mathcal{H}_{\delta_z}} \ell(y_z, h(x)) + \gamma'.$$

Let $B_z := \mathrm{B}_\ell(y_z, r_z + \gamma') \in \mathrm{B}_\ell(\mathcal{Y})$ be the ball centered around y_z of radius $r_z + \gamma'$ where $r_z := \sup_{h \in \mathcal{H}_{\delta_z}} \ell(y_z, h(x))$. The inequality above implies that $z \notin B_z$ (note that B_z changes depending on z). Since $z \in \mathcal{Z}$ was arbitrary, it must be the case that $z \notin B_z$ for all $z \in \mathcal{Z}$.

This means that $\bigcap_{z\in\mathcal{Z}} B_z = \emptyset$. Then, using the fact that $(\mathcal{Y}, \mathcal{Z}, \ell)$ is a Helly space with Helly number p, there exists p balls in $\{B_z\}_{z\in\mathcal{Z}}$ such that their collection-wise intersection is also empty. Accordingly, let $z_1, \ldots, z_p$ be such that $\bigcap_{i=1}^{p} B_{z_i} = \emptyset$. As before, for every $i \in [p]$, let y_{z_i} denote the label on the outgoing edge from v indexed by the Dirac measure δ_{z_i}. By definition, for all $i \in [p]$ and $h_{\delta_{z_i}} \in \mathcal{H}_{\delta_{z_i}}$ we have that

$$h_{\delta_{z_i}}(x) \in \mathrm{B}_\ell(y_{z_i}, r_{z_i}) := \tilde{B}_{z_i}$$

Note that B_{z_i} is the γ' expansion of $\tilde{B}_{z_i}$. For each $i \in [p]$, relabel the outgoing edge from v indexed by δ_{z_i} with the tuple (y_{z_i}, r_{z_i}). For each $i \in [p]$, reindex the outgoing edge from v indexed by δ_{z_i} with i. Remove all other edges indexed by measures and their corresponding subtrees. There should now only be p outgoing edges from v, each indexed by a number $i \in [p]$ and labeled by a tuple in $\mathcal{Y} \times [0, c]$. Note that $\bigcap_{i=1}^{p} \mathrm{B}_\ell(y_{z_i}, r_{z_i} + \gamma') = \bigcap_{i=1}^{p} B_{z_i} = \emptyset$, which matches the second constraint imposed by Definition 24. As for the first constraint on shattering, note that for all $i \in [p]$ and all $h_{\delta_{z_i}} \in \mathcal{H}$, we have that $h_{\delta_{z_i}}(x) \in \tilde{B}_{z_i}$. Thus, the hypothesis that shatters the edges indexed by δ_{z_i} in the original tree according to Definition 20 also shatters the newly re-indexed and relabeled edges according to Definition 24. Thus, for the root node v, both constraints imposed by Definition 24 are met. Recursively repeating the above procedure on the subtrees following the p newly reindexed and relabeled edges results in a p-dim tree γ'-shattered by $\mathcal{H}$ of depth d. Thus, $\mathrm{SM}_\gamma(\mathcal{H}) \leq p\text{-dim}_{\gamma'}(\mathcal{H})$ for $\gamma' < \gamma$.

APPENDIX D

Online Infinite-Dimensional Regression: Learning Linear Operators

D.1 Upperbound Proofs for Online Setting

Our proof of Theorem 21 also relies on the following technical Lemma.

Lemma 21. *Let $v \in \mathcal{V}$, $w \in \mathcal{W}$, and $f \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. Then, we have $\langle f(v), w\rangle = \mathrm{tr}(f \circ (v \otimes w))$.*

Proof. (of Lemma 21) Let $\{\psi_n\}_{n=1}^{\infty}$ be an orthonormal basis of $\mathcal{W}$ and $w = \sum_{n=1}^{\infty} \alpha_n \psi_n$ for an ℓ_2 summable sequence $\{\alpha_n\}_{n\in\mathbb{N}}$. Then, by definition of the trace operator, we have

$$\begin{aligned}\mathrm{tr}(f \circ (v \otimes w)) &= \sum_{n=1}^{\infty} \langle f \circ (v \otimes w)(\psi_n), \psi_n\rangle \\ &= \sum_{n=1}^{\infty} \langle \alpha_n\, f(v), \psi_n\rangle = \left\langle f(v), \sum_{n=1}^{\infty} \alpha_n \psi_n \right\rangle = \langle f(v), w\rangle\,,\end{aligned}$$

which completes our proof. ■

D.1.1 Proof of Lemma 10

Let $F = \sum_{t=1}^{T} \sigma_t\, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})$. Since

$$\mathrm{rank}\,(F) \leq \sum_{t=1}^{T} \mathrm{rank}\,(\sigma_t\, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})) \leq T,$$

F can have at most T non-zero singular values. Let $\{s_t\}_{t=1}^{T}$ be the singular values of the operator F, possibly with multiplicities. Then, for $q \in [1, 2)$, we have

$$\|F\|_q = \left(\sum_{t=1}^{T} s_t^q\right)^{\frac{1}{q}} \leq \left(\left(\sum_{t=1}^{T} (s_t^q)^{\frac{2}{q}}\right)^{\frac{q}{2}} \left(\sum_{t=1}^{T} 1^{\frac{2}{2-q}}\right)^{\frac{2-q}{2}}\right)^{\frac{1}{q}} = \left(\sum_{t=1}^{T} s_t^2\right)^{\frac{1}{2}} T^{\frac{1}{q}-\frac{1}{2}} = \|F\|_2\, T^{\frac{1}{q}-\frac{1}{2}},$$

where the inequality is due to Hölder. As for $q \geq 2$, we trivially have $\|F\|_q \leq \|F\|_2$. In either case, we obtain

$$\|F\|_q \leq \max\left\{T^{\frac{1}{q}-\frac{1}{2}}, 1\right\} \|F\|_2 .$$

Hence, to prove Lemma 10, it suffices to show that

$$\mathbb{E}[\,\|F\|_2] \leq c_1\, c_2\, T^{\frac{1}{2}}.$$

Recall that by definition of the 2-Schatten norm, we have $\|F\|_2 = \sqrt{\mathrm{tr}\,(F^\star F)}$. Using linearity of trace and Jensen's inequality gives $\mathbb{E}\left[\sqrt{\mathrm{tr}\,(F^\star F)}\right] \leq \sqrt{\mathrm{tr}\,(\mathbb{E}\,[F^\star F])}$. Then,

$$\begin{aligned}
&\mathbb{E}\,[F^\star F] \\
&= \mathbb{E}\left[\left(\sum_{t=1}^{T} \sigma_t\, w_t(\sigma_{<t}) \otimes v_t(\sigma_{<t})\right)\left(\sum_{t=1}^{T} \sigma_t\, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right)\right] \\
&= \mathbb{E}\left[\sum_{t,r} \sigma_t\, \sigma_r\, \langle v_t(\sigma_{<t}), v_r(\sigma_{<r})\rangle\, w_t(\sigma_{<t}) \otimes w_r(\sigma_{<r})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \|v_t(\sigma_{<t})\|^2\, w_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right] + \mathbb{E}\left[\sum_{t\neq r} \sigma_t\sigma_r\, \langle v_t(\sigma_{<t}), v_r(\sigma_{<r})\rangle\, w_t(\sigma_{<t}) \otimes w_r(\sigma_{<r})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \|v_t(\sigma_{<t})\|^2\, w_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right].
\end{aligned}$$

To see why the second term above is 0, consider the case $t < r$. We have

$$\begin{aligned}
&\mathbb{E}\,[\sigma_t\sigma_r\, \langle v_t(\sigma_{<t}), v_r(\sigma_{<r})\rangle\, w_t(\sigma_{<t}) \otimes w_r(\sigma_{<r})] \\
&= \mathbb{E}\,[\mathbb{E}\,[\sigma_t\sigma_r\, \langle v_t(\sigma_{<t}), v_r(\sigma_{<r})\rangle\, w_t(\sigma_{<t}) \otimes w_r(\sigma_{<r}) \mid \sigma_{<r}]] \\
&= \mathbb{E}\,[\sigma_t\, \langle v_t(\sigma_{<t}), v_r(\sigma_{<r})\rangle\, w_t(\sigma_{<t}) \otimes w_r(\sigma_{<r})\; \mathbb{E}\,[\sigma_r \mid \sigma_{<r}]] \\
&= 0.
\end{aligned}$$

The last equality follows because σ_r is independent of $\sigma_{<r}$ and thus $\mathbb{E}\left[\sigma_r \mid \sigma_{<r}\right] = \mathbb{E}[\sigma_r] = 0$. The case where $t > r$ is symmetric. Putting everything together, we have

$$\begin{aligned}
\operatorname{tr}\left(\mathbb{E}[F^\star F]\right) &= \operatorname{tr}\left(\mathbb{E}\left[\sum_{t=1}^{T} \left\|v_t(\sigma_{<t})\right\|^2 w_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right]\right) \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\|v_t(\sigma_{<t})\right\|^2 \operatorname{tr}\left(w_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left\|v_t(\sigma_{<t})\right\|^2 \left\|w_t(\sigma_{<t})\right\|^2\right] \\
&\leq \sum_{t=1}^{T} c_1^2 c_2^2 = (c_1 c_2)^2\, T,
\end{aligned}$$

which implies that $\mathbb{E}[\,\|F\|_2] \leq \sqrt{\operatorname{tr}\left(\mathbb{E}\left[F^\star F\right]\right)} \leq \sqrt{(c_1 c_2)^2 T} = c_1\, c_2\, T^{\frac{1}{2}}$. This completes our proof.

D.1.2 Proof of Theorem 21

Define the normalized loss class $\{(u, v) \mapsto \frac{1}{4c^2} \left\|f(u) - v\right\|^2 : f \in \mathcal{F}_p\}$ such that every function in this class maps to $[0, 1]$. Applying [Rakhlin et al., 2015b, Theorem 2] to this normalized loss class, we obtain that the expected regret of $\mathcal{A}$ is $\leq 8c^2 \operatorname{Rad}_T(\overline{\mathcal{F}}_p)$, where $\overline{\mathcal{F}}_p = \{\frac{1}{4c^2} f \mid f \in \mathcal{F}_p\}$ is the normalized operator class. Since $\operatorname{Rad}_T(\overline{\mathcal{F}}_p) = \frac{1}{4c^2} \operatorname{Rad}_T(\mathcal{F}_p)$, the expected regret of $\mathcal{A}$ is $\leq 2 \operatorname{Rad}_T(\mathcal{F}_p)$. This completes the proof of the first inequality. We now focus on proving the second inequality here. By definition, we have

$$\begin{aligned}
&\operatorname{Rad}_T(\mathcal{F}_p) = \sup_{x,y} \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t})) - y_t(\sigma_{<t})\right\|^2\right] \\
&\leq \sup_{x,y} \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t}))\right\|^2\right] + 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} -\sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle\right] \right. \\
&\quad \left. + \mathbb{E}\left[\sum_{t=1}^{T} \sigma_t \left\|y_t(\sigma_{<t}))\right\|^2\right] \right) \\
&= \sup_{x,y} \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t}))\right\|^2\right] + 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle\right] \right).
\end{aligned}$$

To handle the second term above, recall that Lemma 21 implies $\left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle =$

$\mathrm{tr}(f \circ (x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t})))$. Using the linearity of the trace operator, we obtain

$$\begin{aligned}\sum_{t=1}^{T} \sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t}) \right\rangle &= \mathrm{tr}\left(f \circ \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t}) \right) \\ &\leq \|f\|_p \, \left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t}) \right\|_q ,\end{aligned}$$

where $q := 1 - \frac{1}{p}$ is the Hölder conjugate of p [Reed and Simon, 1975, Page 41]. This implies the bound

$$\begin{aligned}\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t}) \right\rangle \right] &\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \|f\|_p \, \left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t}) \right\|_q \right] \\ &\leq c \, \mathbb{E}\left[\left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t}) \right\|_q \right],\end{aligned}$$

where the last inequality follows from the definition of $\mathcal{F}_p$.

To handle the first term in the bound of $\mathrm{Rad}_T(\mathcal{F}_p)$ above, note that

$$\begin{aligned}\|f(x_t(\sigma_{<t}))\|^2 &= \langle f(x_t(\sigma_{<t})), f(x_t(\sigma_{<t})) \rangle \\ &= \langle f^\star f(x_t(\sigma_{<t})), x_t(\sigma_{<t}) \rangle \\ &= \mathrm{tr}(f^\star f \circ (x_t(\sigma_{<t}) \otimes x_t(\sigma_{<t}))),\end{aligned}$$

where the final equality follows from Lemma 21. Using linearity of trace, and the generalized Hölder's inequality for Schatten norms [Reed and Simon, 1975, Page 41], we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \, \|f(x_t(\sigma_{<t}))\|^2 \right] &\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \|f^\star f\|_p \left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes x_t(\sigma_{<t}) \right\|_q \right] \\ &\leq c^2 \, \mathbb{E}\left[\left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes x_t(\sigma_{<t}) \right\|_q \right],\end{aligned}$$

where the last inequality uses the fact that $\|f^\star f\|_p \leq \|f\|_p^2$. Combining everything, we obtain

$$\begin{aligned}\mathrm{Rad}_T(\mathcal{F}_p) &\leq c^2 \, \mathbb{E}\left[\left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes x_t(\sigma_{<t}) \right\|_q \right] + 2c \, \mathbb{E}\left[\left\| \sum_{t=1}^{T} \sigma_t \, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t}) \right\|_q \right] \\ &\leq 3c^2 \, T^{\max\left\{\frac{1}{2}, \frac{1}{q}\right\}},\end{aligned}$$

where the final inequality follows from using Lemma 10 twice. Recalling that $\frac{1}{q} = 1 - \frac{1}{p}$ completes our proof of second inequality.

D.2 Proof of Theorem 23

D.2.1 Proof of lowerbound of $\frac{c^2}{12}\, n^{-\frac{1}{p-1}}$.

Proof. Fix $n, m \in \mathbb{N}$. Let $\mathcal{D}$ be an arbitrary joint distribution on $\mathcal{X} \times \mathcal{Y}$, and U denote the uniform distribution on $\{e_1, \ldots, e_{mn}\}$. For each $\sigma \in \{-1, 1\}^{mn}$, define $h_\sigma = \sum_{i=1}^{mn} c\, \sigma_i\, \psi_i \otimes e_i$. Note that $h_\sigma \notin \mathcal{F}_p$ for large n. The minimax expected excess risk of $\mathcal{F}$ is

$$\begin{aligned}
\mathcal{E}_n(\mathcal{F}) &= \inf_{\hat{f}_n} \sup_{\mathcal{D}} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^n} \left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\left\| \hat{f}_n(x) - y \right\|^2 \right] - \inf_{f\in\mathcal{F}_p} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\|f(x) - y\|^2 \right] \right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\mathop{\mathbb{E}}_{S\sim(U\times h_\sigma)^n} \left[\mathop{\mathbb{E}}_{x\sim U} \left[\left\| \hat{f}_n(x) - h_\sigma(x) \right\|^2 \right] - \inf_{f\in\mathcal{F}_p} \mathop{\mathbb{E}}_{x\sim U} \left[\|f(x) - h_\sigma(x)\|^2 \right] \right] \right],
\end{aligned}$$

where the first inequality follows upon replacing supremum over $\mathcal{D}, \sigma$ with U and expectation over σ respectively. Let $S_x \in \mathcal{X}^n$ denote the instances from labeled samples $S \in (\mathcal{X} \times \mathcal{Y})^n$. We first lower bound the expected risk of the learner, and then upper bound the expected risk of the optimal function in $\mathcal{F}_p$. Exchanging the order of the first two expectations, the lower bound of the expected risk of the learner is

$$\begin{aligned}
&\inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x\sim U^n} \left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\mathop{\mathbb{E}}_{x\sim U} \left[\left\| \hat{f}_n(x) - h_\sigma(x) \right\|^2 \right] \right] \right] \\
&= \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x\sim U^n} \left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\frac{1}{mn} \sum_{i=1}^{mn} \left\| \hat{f}_n(e_i) - h_\sigma(e_i) \right\|^2 \right] \right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N\sim\mathrm{Unif}(\{1,,\ldots,mn\})^n} \left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\frac{1}{mn} \sum_{i\notin N} \left\| \hat{f}_n(e_i) - c\,\sigma_i\psi_i \right\|^2 \right] \right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N\sim\mathrm{Unif}(\{1,,\ldots,mn\})^n} \left[\frac{1}{mn} \sum_{i\notin N} \left(\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\left\| \hat{f}_n(e_i) - c\,\sigma_i\psi_i \right\| \right] \right)^2 \right].
\end{aligned}$$

In order to get the second to the last inequality, we reinterpret sampling x uniformly from $\{e_1, \ldots, e_{mn}\}$ as sampling index i uniformly from $\{1, \ldots, mn\}$ and drawing e_i. The final inequality follows upon exchanging the sum and expectation and applying Jensen's. Note that, whenever $i \notin N$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}} \left[\left\| \hat{f}_n(e_i) - c\,\sigma_i\psi_i \right\| \right] = \mathbb{E} \left[\mathop{\mathbb{E}}_{\sigma_i} \left[\left\| \hat{f}_n(e_i) - c\,\sigma_i\psi_i \right\| \right] \mid \hat{f}_n \right] \\
&= \mathbb{E} \left[\frac{1}{2} \left(\left\| \hat{f}_n(e_i) - c\,\psi_i \right\| + \left\| \hat{f}_n(e_i) + c\,\psi_i \right\| \right) \mid \hat{f}_n \right] \\
&\geq \frac{1}{2} \|c\psi_i + c\psi_i\| \\
&= c,
\end{aligned}$$

where we use the fact $\hat{f}_n$ is independent of σ_i for all $i \notin N$ and triangle inequality. Thus, combining everything, our lower bound is

$$\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N \sim \text{Unif}(\{1,,\ldots,mn\})^n} \left[\frac{1}{mn} \sum_{i \notin N} c^2 \right] = \frac{c^2}{mn} \sum_{i=1}^{mn} \mathbb{P}(i \notin N) = c^2 \left(1 - \frac{1}{mn}\right)^n.$$

For the last equality, we use the fact that the probability of i not appearing in the set N obtained by n random uniform draw from $\{1, 2, \ldots, mn\}$ with replacement is $\left(1 - \frac{1}{mn}\right)^n$.

Next, we upperbound optimal expected risk amongst functions in $\mathcal{F}_p$. Consider

$$f_{\sigma,p} = \sum_{j=1}^{mn} \frac{c\,\sigma_j}{(mn)^{1/p}}\, \psi_j \otimes e_j.$$

Clearly, $\|f_{\sigma,p}\|_p \leq c$ for all $p \in [1, \infty]$ and thus $f_{\sigma,p} \in \mathcal{F}_p$. Therefore, we can write

$$\begin{aligned}
\inf_{f \in \mathcal{F}_p} \mathop{\mathbb{E}}_{x \sim U} \left[\|f(x) - h_\sigma(x)\|^2 \right] &\leq \mathop{\mathbb{E}}_{x \sim U} \left[\|f_{\sigma,p}(x) - h_\sigma(x)\|^2 \right] \\
&= \frac{1}{mn} \sum_{i=1}^{mn} \|f_{\sigma,p}(e_i) - h_\sigma(e_i)\|^2 \\
&= \frac{1}{mn} \sum_{i=1}^{mn} \left\| \frac{c\,\sigma_i}{(mn)^{1/p}}\, \psi_i - c\,\sigma_i \psi_i \right\|^2 \\
&= \frac{1}{mn} \sum_{i=1}^{mn} c^2 \left(1 - \frac{1}{(mn)^{1/p}}\right)^2 \\
&= c^2 \left(1 - \frac{1}{(mn)^{1/p}}\right)^2 \leq c^2 \left(1 - \frac{1}{(mn)^{1/p}}\right).
\end{aligned}$$

Thus, putting everything together, the minimax expected excess risk is

$$\begin{aligned}
&\geq c^2 \left(1 - \frac{1}{mn}\right)^n - c^2 \left(1 - \frac{1}{(mn)^{1/p}}\right) \\
&\geq c^2 \left(1 - \frac{1}{2m}\right)^2 - c^2 \left(1 - \frac{1}{(mn)^{1/p}}\right) \qquad (\text{ for } n \geq 2) \\
&\geq c^2 \left(\frac{1}{(mn)^{\frac{1}{p}}} - \frac{1}{2m} \right).
\end{aligned}$$

Next, pick $m = \left\lceil 2n^{\frac{1}{p-1}} \right\rceil$. Then, we have that $2n^{\frac{1}{p-1}} \leq m \leq 3n^{\frac{1}{p-1}}$. So, the expression above

is further lower bounded by

$$c^2\left(\frac{1}{(3n^{\frac{1}{p-1}}\,n)^{\frac{1}{p}}}-\frac{1}{2\,2n^{\frac{1}{p-1}}}\right)\geq c^2\left(\frac{1}{3n^{\frac{1}{p-1}}}-\frac{1}{4n^{\frac{1}{p-1}}}\right)=\frac{c^2}{12n^{\frac{1}{p-1}}}.$$

This completes our proof. ■

D.2.2 Proof of lowerbound of $\frac{c^2}{8}\,n^{-\frac{2}{p}}$.

Our proof here follows similar arguments as the proof in D.2.1. However, the lowerbound in this section is derived in the realizable setting.

Proof. Fix $n, m \in \mathbb{N}$. Let $\mathcal{D}$ be an arbitrary joint distribution on $\mathcal{X}\times\mathcal{Y}$, and let U denote the uniform distribution on $\{e_1,\ldots,e_{mn}\}$. For each $\sigma\in\{-1,1\}^{mn}$, define $f_{\sigma,p}=\sum_{i=1}^{mn}\frac{c}{(mn)^{1/p}}\,\sigma_i\,\psi_i\otimes e_i$. Note that $f_{\sigma,p}\in\mathcal{F}_p$ for all $p\geq 1$. The minimax expected excess risk of $\mathcal{F}$ is

$$\begin{aligned}
\mathcal{E}_n(\mathcal{F}) &= \inf_{\hat{f}_n}\sup_{\mathcal{D}}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left\|\hat{f}_n(x)-y\right\|^2\right]-\inf_{f\in\mathcal{F}_p}\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\|f(x)-y\|^2\right]\right]\\
&\geq \inf_{\hat{f}_n}\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\mathop{\mathbb{E}}_{S\sim(U\times f_{\sigma,p})^n}\left[\mathop{\mathbb{E}}_{x\sim U}\left[\left\|\hat{f}_n(x)-f_{\sigma,p}(x)\right\|^2\right]-\inf_{f\in\mathcal{F}_p}\mathop{\mathbb{E}}_{x\sim U}\left[\|f(x)-f_{\sigma,p}(x)\|^2\right]\right]\right]\\
&\geq \inf_{\hat{f}_n}\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\mathop{\mathbb{E}}_{S\sim(U\times f_{\sigma,p})^n}\left[\mathop{\mathbb{E}}_{x\sim U}\left[\left\|\hat{f}_n(x)-f_{\sigma,p}(x)\right\|^2\right]\right]\right]
\end{aligned}$$

where the first inequality follows upon replacing supremum over $\mathcal{D},\sigma$ with U and expectation over σ. The second inequality follows because $\inf_{f\in\mathcal{F}_p}\mathbb{E}_{x\sim U}\left[\|f(x)-f_{\sigma,p}(x)\|^2\right]\leq\mathbb{E}_{x\sim U}\left[\|f_{\sigma,p}(x)-f_{\sigma,p}(x)\|^2\right]=0$ as $f_{\sigma,p}\in\mathcal{F}_p$.

Let S_x denote the instances from labeled samples S. We first lower bound the expected risk of the learner $\hat{f}_n$. Following the same calculation as in the first part of the proof, the lower bound of the expected risk of the learner is

$$\begin{aligned}
&\inf_{\hat{f}_n}\mathop{\mathbb{E}}_{S_x\sim U^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\mathop{\mathbb{E}}_{x\sim U}\left[\left\|\hat{f}_n(x)-f_{\sigma,p}(x)\right\|^2\right]\right]\right]\\
&=\inf_{\hat{f}_n}\mathop{\mathbb{E}}_{S_x\sim U^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\frac{1}{mn}\sum_{i=1}^{mn}\left\|\hat{f}_n(e_i)-f_{\sigma,p}(e_i)\right\|^2\right]\right]\\
&\geq\inf_{\hat{f}_n}\mathop{\mathbb{E}}_{N\sim\mathrm{Unif}(\{1,,\ldots,mn\})^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\frac{1}{mn}\sum_{i\notin N}\left\|\hat{f}_n(e_i)-\frac{c\,\sigma_i}{(mn)^{1/p}}\psi_i\right\|^2\right]\right]\\
&\geq\inf_{\hat{f}_n}\mathop{\mathbb{E}}_{N\sim\mathrm{Unif}(\{1,,\ldots,mn\})^n}\left[\frac{1}{mn}\sum_{i\notin N}\left(\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{mn}}\left[\left\|\hat{f}_n(e_i)-\frac{c\,\sigma_i}{(mn)^{1/p}}\psi_i\right\|\right]\right)^2\right].
\end{aligned}$$

To get the second to the last inequality, we reinterpret sampling x uniformly from $\{e_1, \ldots, e_{mn}\}$ as sampling index i uniformly from $\{1, \ldots, mn\}$ and drawing e_i. The final inequality follows upon exchanging the sum and expectation and applying Jensen's. Note that, whenever $i \notin N$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{mn}} \left[\left\| \hat{f}_n(e_i) - c\,\sigma_i \psi_i \right\| \right] = \mathbb{E}\left[\mathop{\mathbb{E}}_{\sigma_i} \left[\left\| \hat{f}_n(e_i) - \frac{c\,\sigma_i}{(mn)^{1/p}} \psi_i \right\| \right] \mid \hat{f}_n \right] \\
&= \mathbb{E}\left[\frac{1}{2} \left(\left\| \hat{f}_n(e_i) - \frac{c}{(mn)^{1/p}}\,\psi_i \right\| + \left\| \hat{f}_n(e_i) + \frac{c}{(mn)^{1/p}} \right\| \right) \mid \hat{f}_n \right] \\
&\geq \frac{c}{(mn)^{1/p}}
\end{aligned}$$

where we use the fact $\hat{f}_n$ is independent of σ_i as $i \notin N$ and triangle inequality. Thus, combining everything, our lower bound is

$$\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N \sim \mathrm{Unif}(\{1,,\ldots,mn\})^n} \left[\frac{1}{mn} \sum_{i \notin N} \frac{c^2}{(mn)^{2/p}} \right] = \frac{c^2}{(mn)^{2/p}} \left(1 - \frac{1}{mn} \right)^n.$$

For the last equality, we use the fact that the probability of i not appearing in the set N obtained by n random uniform draw from $\{1, 2, \ldots, mn\}$ with replacement is $\left(1 - \frac{1}{mn}\right)^n$. Picking $m = 2$ and using the fact that $\left(1 - \frac{1}{2n}\right)^n \geq 1 - 1/2 = 1/2$, we obtain the lowerbound of $\frac{c^2}{8} n^{-\frac{2}{p}}$.

■

APPENDIX E

Controlling Statistical, Discretization, and Truncation Errors in Learning Fourier Linear Operators

E.1 Fourier Linear Operators

In this section, we provide a formal treatment of Fourier linear operators and the corresponding parametrization in FNOs. Recall that, in the Fourier Neural operator, one assumes that $\mathcal{X} = \mathcal{Y} = \mathbb{T}^d$ and the kernel is translation invariant. This implies that $\mathcal{K}_\theta$ defined in Section 6.1 is a convolution operator. That is,

$$\mathcal{K}_\theta\, v = k_\theta \star v, \qquad \text{where} \qquad (k_\theta \star v)(y) = \int_{\mathbb{T}^d} k_\theta(y - x)\, v(x)\, dx.$$

The convolution is done elementwise, $(\mathcal{K}_\theta v)_i(y) = \sum_{j=1}^{p} \Big([k_\theta]_{ij} \star v_j\Big)(y)$, where $[k_\theta]_{ij} : \mathbb{T}^d \to \mathbb{R}$ is the scalar-valued kernel defined by the $(i,j)^{th}$ component of k_θ and $(\mathcal{K}_\theta v)_i$ is the i^{th} component of a $\mathbb{R}^q$-valued function. Similarly, $v_j : \mathbb{T}^d \to \mathbb{R}$ is the j^{th} component function of $\mathbb{R}^p$-valued function v. Next, using the linearity of the Fourier transform and the Convolution Theorem, we can write

$$\left(\mathcal{K}_\theta v\right)_i = \mathcal{F}^{-1}\left(\sum_{j=1}^{p} \mathcal{F}\Big([k_\theta]_{ij}\Big)\, \mathcal{F}(v_j)\right).$$

where $\mathcal{F}$ is Fourier transform operator, and $\mathcal{F}^{-1}$ is the inverse Fourier transform. Here, $\mathcal{F}([k_\theta]_{ij}) : \mathbb{Z}^d \to \mathbb{C}$ and $\mathcal{F}(v_j) : \mathbb{Z}^d \to \mathbb{C}$ are Fourier transforms of $[k_\theta]_{ij}$ and v_j respectively. Note that only discrete Fourier modes are defined because all the functions are defined on a periodic domain $\mathbb{T}^d$.

The key insight in FNO is that instead of parametrizing the kernel k_θ, we parametrize its Fourier transform $\mathcal{F}(k_\theta)$ directly. That is, we parametrize the kernel transform operator as $\left(\mathcal{K}_\beta v\right)_i = \mathcal{F}^{-1}\left(\sum_{j=1}^{p} [\Lambda_\beta]_{ij}\, \mathcal{F}(v_j)\right)$ for some $\Lambda_\beta : \mathbb{Z}^d \to \mathbb{C}^{q\times p}$ that maps Fourier modes to a complex-valued matrix. Using the linearity of the inverse Fourier transform, we can write

this more succinctly in a matrix form as $\mathcal{K}_\beta\, v = \mathcal{F}^{-1}\Big(\,\Lambda_\beta\;\mathcal{F}(v)\Big)$.

Since $\mathcal{F}^{-1}\Big(\,\Lambda_\beta\;\mathcal{F}(v)\Big)$ is a function defined on periodic domain $\mathbb{T}^d$, it has a Fourier series representation. So, we can write

$$\mathcal{F}^{-1}\Big(\,\Lambda_\beta\;\mathcal{F}(v)\Big)(\cdot) = \sum_{m\in\mathbb{Z}^d} \varphi_m(\cdot)\,\Lambda_\beta(m)\,(\mathcal{F}v)(m),$$

as $\varphi_m(\cdot) := e^{2\pi\,\mathrm{i}\langle m,\cdot\rangle}$ and the m^{th} Fourier coefficient of $\mathcal{F}^{-1}\Big(\,\Lambda_\beta\;\mathcal{F}(v)\Big)$ is $\Lambda_\beta(m)\;(\mathcal{F}v)\,(m)$.

We have not specified in what metric the sum on the right-hand side converges. However, the convergence is not really an issue from a practical standpoint. In practice, Λ_β is a trainable parameter, and it has been observed in Li et al. [2021] that parametrizing Λ_β as a function from $\mathbb{Z}^d$ to $\mathbb{C}^{q\times p}$ yields sub-optimal results, possibly due to discrete structure of the lattice $\mathbb{Z}^d$. So, one picks a large $K > 0$ and parametrize Λ_β as a collection of matrices $\{\Lambda_\beta(m)\;:\; m \in \mathbb{Z}^d \text{ such that } |m|_\infty \le K\}$. In this case, the sum contains $\le K^d$ terms and thus always converges. If one still wants to deal with the infinite sum, a standard assumption would be $[\Lambda_\beta]_{ij} \in \ell^1(\mathbb{Z}^d)$ for all (i,j) pairs. That is, $\sum_{m\in\mathbb{Z}^d} |[\Lambda_\beta(m)]_{ij}| < \infty$ for all (i,j) pairs. Then, the Weirstrass M-test implies that the sum above converges uniformly over all $y \in \mathbb{T}^d$.

Reparametrizing $\mathcal{K}_\theta$ as $\mathcal{F}^{-1}\Big(\Lambda_\beta\;\mathcal{F}(v)\Big)$ was proposed by Li et al. [2021] from the perspective of the convolution theorem, as discussed earlier. However, a more natural way to derive $\mathcal{F}^{-1}\Big(\Lambda_\beta\;\mathcal{F}(v)\Big)$ from $\mathcal{K}_\theta$ is to assume that k_θ has a Mercer-type decomposition.

Proposition 6. *Let $k_\theta : \mathbb{Z}^d \to \mathbb{C}^{q\times p}$ be a kernel with decomposition*

$$[k_\theta(y,x)]_{ij} = \sum_{m\in\mathbb{Z}^d} [\Lambda_\beta(m)]_{ij}\;\varphi_m(y)\;\varphi_{-m}(x) \quad \forall (i,j)$$

for some $\Lambda_\beta : \mathbb{Z}^d \to \mathbb{C}^{q\times p}$ such that $\Lambda_\beta \in \ell^1(\mathbb{Z}^d)$. Then, $\mathcal{K}_\theta v = \mathcal{F}^{-1}\Big(\,\Lambda_\beta\;\mathcal{F}(v)\Big)$ for all $v \in \mathcal{V}$.

Given such decomposition, a simple algebra shows that

$$\int_{\mathbb{T}^d} [k_\theta(y,x)]_{ij}\,\varphi_k(x)dx = [\Lambda_\beta(k)]_{ij}\;\varphi_k(y).$$

In other words, $[\Lambda_\beta(k)]_{ij}$ are the eigenvalues of the integral operator defined by the kernel $[k_\theta]_{ij}$. This suggests that the Fourier layer of FNOs is parametrizing the eigenvalues of an operator while fixing the eigenfunctions to be φ_k's. So, setting $\Lambda_\beta(m) = 0$ for $m \in \mathbb{Z}^d_{>K}$ amounts to parametrizing the low-rank version of such operator. This viewpoint shows that FNO is just a special case of a Low-rank Neural Operator defined in [Kovachki et al., 2023, Section 4.2].

More importantly, Proposition 6 (see Appendix E.1.1 for the proof) provides a natu-

ral way to generalize Fourier Neural Operators. That is, we can consider $[k_\theta(y,x)]_{ij} = \sum_{m\in\mathcal{J}} [\Lambda_\beta(m)]_{ij}\ \psi_m(y)\,\phi_m(x)$, where $\mathcal{J}$ is some countable index-set and $\{\psi_m\}_{m\in\mathcal{J}}$, $\{\phi_m\}_{m\in\mathcal{J}}$ are some orthonormal sequences. Some common orthonormal sequences that allow efficient computation like FFT include the Chebyshev polynomial and wavelet basis. Some works have already explored the practical advantage of replacing Fourier basis with wavelet basis in certain problem settings Gupta et al. [2021], Tripura and Chakraborty [2023].

E.1.1 Proof of Proposition 6

We now end this section by proving Proposition 6.

Proof. Let $\lambda_{ij}(m) := [\Lambda_\beta(m)]_{ij}$ and assume that

$$[k_\theta(y,x)]_{ij} = \sum_{m\in\mathbb{Z}^d} \lambda_{ij}(m)\ \varphi_m(y)\ \varphi_{-m}(x).$$

Using this decomposition, we obtain

$$\begin{aligned}
(\mathcal{K}_\theta v)_i(y) &= \int_{\mathbb{T}^d} \sum_{j=1}^{p} [k_\theta(y,x)]_{ij}\ v_j(x)\,dx \\
&= \int_{\mathbb{T}^d} \sum_{j=1}^{p} \sum_{m\in\mathbb{Z}^d} \lambda_{ij}(m)\ \varphi_m(y)\ \varphi_{-m}(x)\,v_j(x)\,dx \\
&= \sum_{m\in\mathbb{Z}^d} \varphi_m(y)\ \sum_{j=1}^{p} \lambda_{ij}(m)\ \int_{\mathbb{T}^d} \varphi_{-m}(x)\,v_j(x)\,dx.
\end{aligned}$$

Note that swapping the integral and the summation is justified through Fubini's because the sum over $\mathbb{Z}^d$ converges absolutely (as $\Lambda_\beta \in \ell^1$) and $\mathbb{T}^d$ is a bounded set. Since

$$\int_{\mathbb{T}^d} \varphi_{-m}(x)\,v_j(x)\,dx = \int_{\mathbb{T}^d} e^{-2\pi\,\mathrm{i}\langle m,x\rangle} v_j(x)\,dx = \mathcal{F}(v_j)(m),$$

we can write

$$(\mathcal{K}_\theta v)_i(y) = \sum_{m\in\mathbb{Z}^d} \varphi_m(y) \sum_{j=1}^{p} \lambda_{ij}(m)\ \mathcal{F}(v_j)(m).$$

Next, consider the function $w := \mathcal{F}^{-1}\left(\sum_{j=1}^{p} \lambda_{ij}\,\mathcal{F}(v_j)\right)$. Our proof will be complete upon showing that $w(y) = (\mathcal{K}_\theta v)_i(y)$ for every $y \in \mathbb{T}^d$. Since the function $w : \mathbb{T}^d \to \mathbb{C}$ is defined on a periodic domain, it has a Fourier series representation. That is,

$$w(y) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,y\rangle}\,\mathcal{F}(w)(m) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,y\rangle} \sum_{j=1}^{p} \lambda_{ij}(m)\,\mathcal{F}(v_j)(m),$$

where the final equality follows because

$$\mathcal{F}\left(\mathcal{F}^{-1}\left(\sum_{j=1}^{p}\lambda_{ij}\,\mathcal{F}(v_j)\right)\right)(m)=\sum_{j=1}^{p}\lambda_{ij}(m)\,\mathcal{F}(v_j)(m).$$

As usual, $\Lambda_\beta \in \ell^1$ implies that the sum above converges uniformly over $y \in \mathbb{T}^d$. Recalling that $\varphi_m(y) = e^{2\pi\,\mathrm{i}\langle m,y\rangle}$, we have shown that $(\mathcal{K}_\theta v)_i(y) = w(y)$ for all $y \in \mathbb{T}^d$. This subsequently implies that

$$(\mathcal{K}_\theta v)_i = w = \mathcal{F}^{-1}\left(\sum_{j=1}^{p}\lambda_{ij}\,\mathcal{F}(v_j)\right).$$

Finally, using the linearity of the inverse Fourier transform and writing this in the matrix form establishes that $\mathcal{K}_\theta v = \mathcal{F}^{-1}(\Lambda_\beta\,\mathcal{F}(v))$ for any $v \in \mathcal{V}$. ■

E.2 Proof of Proposition 1

Proof. Fix $v \in \mathcal{V}$ and define $w := \mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(v)\Big)$. By definition of the operator $\mathcal{F}^{-1}\Big(\lambda\ \mathcal{F}(\cdot)\Big)$, we have

$$w = \mathcal{F}^{-1}\left(\lambda\,\mathcal{F}(v)\,\right).$$

Using the Fourier series representation of w, we have

$$w(\cdot) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,\cdot\rangle}\,(\mathcal{F}w)(m) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,\cdot\rangle}\,\lambda_m\ \mathcal{F}(v)(m).$$

This step is rigorously justified because $\lambda \in \ell^1$. Noting that

$$(\mathcal{F}v)(m) = \int_{\mathbb{T}^d} e^{-2\pi\,\mathrm{i}\langle m,x\rangle}\,v(x)\,dx = \langle\varphi_{-m}, v\rangle_{L^2}\,,$$

we can write

$$w(\cdot) = \sum_{m\in\mathbb{Z}^d} e^{2\pi\,\mathrm{i}\langle m,\cdot\rangle}\,\lambda_m\ \langle\varphi_{-m}, v\rangle_{L^2}\,.$$

Thus, $w = \sum_{m\in\mathbb{Z}^d}\lambda_m\ \langle\varphi_{-m}, v\rangle_{L^2}\ \varphi_m$, where the convergence is uniform over $\mathbb{T}^d$. This implies that

$$\mathcal{F}^{-1}\left(\lambda\,\mathcal{F}(v)\,\right) = \sum_{m\in\mathbb{Z}^d}\lambda_m\ \langle\varphi_{-m}, v\rangle_{L^2}\ \varphi_m.$$

Since this equality holds for every $v \in \mathcal{V}$, we have

$$\mathcal{F}^{-1}\left(\lambda\,\mathcal{F}(\cdot)\,\right) = \sum_{m\in\mathbb{Z}^d}\lambda_m\ \varphi_m\otimes\varphi_{-m}.$$

■

E.3 Technical Lemmas

In this section, we state and derive some technical Lemmas that we use to prove Theorems 25 and 26.

Lemma 22. *For any $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, we have*

$$| \langle \varphi_{-m}, u \rangle_{L^2} | \leq \frac{\|u\|_{\mathcal{H}^s}}{(2\pi)^s \, |m|_\infty^s} \qquad \forall m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}.$$

Proof. Fix $m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ and let $|m_j| = |m|_\infty = \max_{1 \leq i \leq d} |m_i|$. Clearly, $m_j \neq 0$. Integrating by parts s times with respect to variable x_j in $x = (x_1, \ldots, x_d)$, we obtain

$$\begin{aligned} \langle \varphi_{-m}, u \rangle = \int_{\mathbb{T}^d} u(x) e^{-2\pi \, \mathrm{i} \langle m, x \rangle} \, dx &= (-1)^s \int_{\mathbb{T}^d} (\partial_j^s u)(x) \frac{e^{-2\pi \, \mathrm{i} \langle m, x \rangle}}{(-2\pi \, \mathrm{i} \; m_j)^s} \, dx \\ &= \left(\frac{1}{2\pi \, \mathrm{i} \, m_j} \right)^s \left\langle \varphi_{-m}, \partial_j^s \, u \right\rangle. \end{aligned}$$

Here, all boundary terms vanish because $\mathbb{T}^d$ does not have a boundary ([Grafakos, 2008, Proof of Theorem 3.3.9]). Taking absolute value on both sides, we obtain that

$$|m_j|^s \, | \langle \varphi_{-m}, u \rangle | = (2\pi)^{-s} \, | \left\langle \varphi_{-m}, \partial_j^s u \right\rangle |$$

Finally, using the fact that $\left| \left\langle \varphi_{-m}, \partial_j^s u \right\rangle \right| \leq \|u\|_{\mathcal{H}^s}$ completes our proof. ■

Lemma 23. *For any $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$, we have*

$$\sum_{m \in \mathbb{Z}^d} (1 + |m|_\infty^{2s}) \; | \langle \varphi_{-m}, u \rangle |^2 \leq \|u\|_{\mathcal{H}^s}^2 .$$

Proof. Fix $m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ and let $|m_j| = |m|_\infty = \max_{1 \leq i \leq d} |m_i|$. Clearly, $m_j \neq 0$. Integrating by parts s times with respect to variable x_j in $x = (x_1, \ldots, x_d)$, we obtain

$$\begin{aligned} \langle \varphi_{-m}, u \rangle = \int_{\mathbb{T}^d} u(x) e^{-2\pi \, \mathrm{i} \langle m, x \rangle} \, dx &= (-1)^s \int_{\mathbb{T}^d} (\partial_j^s u)(x) \frac{e^{-2\pi \, \mathrm{i} \langle m, x \rangle}}{(-2\pi \, \mathrm{i} \; m_j)^s} \, dx \\ &= \left(\frac{1}{2\pi \, \mathrm{i} \, m_j} \right)^s \left\langle \varphi_{-m}, \partial_j^s \, u \right\rangle. \end{aligned}$$

Here, all boundary terms vanish because $\mathbb{T}^d$ does not have a boundary ([Grafakos, 2008, Proof of Theorem 3.3.9]). Taking absolute value on both sides, we obtain that

$$|m_j|^s \, |\langle \varphi_{-m}, u \rangle| = (2\pi)^{-s} \, \left| \left\langle \varphi_{-m}, \partial_j^s u \right\rangle \right|$$

Noting that $|m_j| = |m|_\infty$, squaring and summing over all $m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}$ to get

$$\sum_{m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}} |m|_\infty^{2s} \, |\langle \varphi_{-m}, u \rangle|^2 = (2\pi)^{-2s} \sum_{m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}} \left| \left\langle \varphi_{-m}, \partial_j^s u \right\rangle \right|^2 \leq (2\pi)^{-2s} \left\| \partial_j^s u \right\|_{L^2}^2,$$

where the final inequality uses Parseval's identity and the fact that $\partial_j^s u \in L^2(\mathbb{T}^d, \mathbb{R})$. Thus, we obtain

$$\begin{aligned}
\sum_{m \in \mathbb{Z}^d} (1 + |m|_\infty^{2s}) \, |\langle \varphi_{-m}, u \rangle|^2 &= \sum_{m \in \mathbb{Z}^d} |\langle \varphi_{-m}, u \rangle|^2 + \sum_{m \in \mathbb{Z}^d \backslash \{\mathbf{0}\}} |m|_\infty^{2s} \, |\langle \varphi_{-m}, u \rangle|^2 \\
&\leq \|u\|_{L^2}^2 + (2\pi)^{-2s} \left\| \partial_j^s u \right\|_{L^2}^2 \\
&\leq \|u\|_{L^2}^2 + \left\| \partial_j^s u \right\|_{L^2}^2 \\
&\leq \|u\|_{\mathcal{H}^s}^2,
\end{aligned}$$

completing our proof. ■

Lemma 24. *For any $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$ such that $s \geq 0$ and $K \in \mathbb{Z}_{>0}$, we have*

$$\sum_{m \in \mathbb{Z}^d_{>K}} |\langle \varphi_{-m}, u \rangle|^2 \leq \frac{\|u\|_{\mathcal{H}^s}^2}{K^{2s}}$$

Proof. Observe that

$$\begin{aligned}
\sum_{m \in \mathbb{Z}^d_{>K}} |\langle \varphi_{-m}, u \rangle|^2 &= \sum_{m \in \mathbb{Z}^d_{>K}} (1 + |m|_\infty^{2s}) \, |\langle \varphi_{-m}, u \rangle|^2 \, \frac{1}{(1 + |m|_\infty^{2s})} \\
&\leq \frac{1}{1 + K^{2s}} \sum_{m \in \mathbb{Z}^d_{>K}} (1 + |m|_\infty^{2s}) \, |\langle \varphi_{-m}, u \rangle|^2 \\
&\leq \frac{\|u\|_{\mathcal{H}^s}^2}{K^{2s}},
\end{aligned}$$

using Lemma 23. ■

Lemma 25. *For any $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$ such that $s > d/2$, we have*

$$\sum_{m\in\mathbb{Z}^d_{>K}} |\langle\varphi_{-m},u\rangle| \le \|u\|_{\mathcal{H}^s}\sqrt{\frac{3^d}{2s-d}}\frac{1}{\sqrt{K^{2s-d}}},.$$

Proof. First, we use Cauchy-Schwarz to get

$$\sum_{m\in\mathbb{Z}^d_{>K}} |\langle\varphi_{-m},u\rangle| = \sqrt{\sum_{m\in\mathbb{Z}^d_{>K}}(1+|m|_\infty^{2s})|\langle\varphi_{-m},u\rangle|^2}\sqrt{\sum_{m\in\mathbb{Z}^d_{>K}}\frac{1}{(1+|m|_\infty^{2s})}}$$

Lemma 24 implies that the first term is $\le \|u\|_{\mathcal{H}^s}$. To bound the second term, note that for any $j\in\mathbb{N}$, we have $|\{m\in\mathbb{Z}^d : |m|_\infty = j\}| = 2(2j+1)^{d-1}$. This is because one of the entry of m has to be $\pm j$ and other $d-1$ entries could be anything in $\{-j\ldots,-1,0,1,\ldots,j\}$. So,

$$\begin{aligned}\sum_{m\in\mathbb{Z}^d_{>K}}\frac{1}{(1+|m|_\infty^{2s})} = \sum_{j>K}\frac{2\,(2j+1)^{d-1}}{(1+j^{2s})} \le 3^d\sum_{j>K}\frac{1}{j^{2s-d+1}} &\le 3^d\int_K^\infty t^{-2s+d-1}\,dt\\ &= \frac{3^d}{2s-d}\frac{1}{K^{2s-d}},\end{aligned}$$

for all $s > d/2$. Thus, overall, we obtain

$$\sum_{m\in\mathbb{Z}^d_{>K}} |\langle\varphi_{-m},u\rangle| \le \|u\|_{\mathcal{H}^s}\sqrt{\frac{3^d}{2s-d}}\frac{1}{\sqrt{K^{2s-d}}},$$

completing our proof. ■

Lemma 26. *Let* $\mathrm{G} := \left\{j/N : j\in\{0,\ldots,N-1\}^d\right\}$ *be the* N*-uniform grid of* $[0,1]^d$*. Then, for any* $m\in\mathbb{Z}^d_{<N}$*, we have*

$$\frac{1}{N^d}\sum_{x\in\mathrm{G}} e^{2\pi\,\mathrm{i}\langle k-m,x\rangle} = \mathbb{1}[k\equiv m\,(mod\ N)].$$

Here, we say $k\equiv m(\mathrm{mod}\ N)$ if $\exists\ell\in\mathbb{Z}^d$ such that $k = N\ell+m$.

Proof. We first prove it for $d=1$. For this case, we need to show that

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{2\pi\,\mathrm{i}(k-m)\frac{j}{N}} = \mathbb{1}[k\equiv m(\mathrm{mod}\ N)].$$

First, consider the case where $k = \tau N+m$ for some $\tau\in\mathbb{Z}$. Then, $e^{2\pi\,\mathrm{i}(k-m)\frac{j}{N}} = e^{2\pi\,\mathrm{i}\,\tau\,j} = 1$ by Euler's identity. Thus, the overall sum must be 1. Next, assume that $k\not\equiv m\ (\mathrm{mod}\ N)$.

Then, the geometric series formula implies that

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{2\pi\,\mathrm{i}(k-m)\frac{j}{N}} = \frac{1}{N}\,\frac{1-e^{2\pi\,\mathrm{i}(k-m)j}}{1-e^{2\pi\,\mathrm{i}(k-m)\frac{j}{N}}} = 0.$$

Here, the final equality holds because $e^{2\pi\,\mathrm{i}(k-m)j} = 1$ by Euler's identity, whereas $e^{2\pi\,\mathrm{i}(k-m)\frac{j}{N}} \neq 1$ for every $j \in \{0, 1\ldots, N-1\}$. This completes our proof for the case $d = 1$.

Next, to prove it for general d, we write the sum as d-fold summation

$$\begin{aligned}\frac{1}{N^d}\sum_{x\in\mathrm{G}} e^{2\pi\,\mathrm{i}\langle k-m,x\rangle} &= \frac{1}{N^d}\sum_{j_1=0}^{N-1}\cdots\sum_{j_d=0}^{N-1} e^{2\pi\,\mathrm{i}(k_1-m_1)\frac{j_1}{N}}\cdots e^{2\pi\,\mathrm{i}(k_d-m_d)\frac{j_d}{N}}\\ &= \prod_{t=1}^{d}\frac{1}{N}\sum_{j_t=0}^{N-1} e^{2\pi\,\mathrm{i}(k_t-m_t)\frac{j_t}{N}}.\end{aligned}$$

Using the result of $d = 1$ case for each term in the product, we have

$$\frac{1}{N^d}\sum_{x\in\mathrm{G}} e^{2\pi\,\mathrm{i}\langle k-m,x\rangle} = \prod_{t=1}^{d}\mathbb{1}[k_t \equiv m_t\,(\mathrm{mod}\ N)] = \mathbb{1}[k \equiv m\ (\mathrm{mod}\, N)].$$

■

Lemma 27. *Let $u \in \mathcal{H}^s(\mathbb{T}^d, \mathbb{R})$ such that $\|u\|_{\mathcal{H}^s} \leq B$ and $u^N := \{u(x) : x \in \mathrm{G}\}$ be its values on the uniform grid G. Then, for all $|m|_\infty < N$, we have*

$$|\,\mathrm{DFT}(u^N)(-m) - \langle\varphi_{-m}, u\rangle\,| \;\leq\; \left|\sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}}\Big\langle\varphi_{-(\ell N+m)}, u\Big\rangle\right|.$$

Proof. Recall that

$$\mathrm{DFT}(u^N)(-m) = \frac{1}{N^d}\sum_{x\in\mathrm{G}} u(x)\, e^{-2\pi\,\mathrm{i}\langle m,x\rangle}.$$

Pick some $M > N$ and write

$$u(x) = \sum_{k\in\mathbb{Z}^d_{\leq M}}\langle\varphi_{-k}, u\rangle\; e^{2\pi\,\mathrm{i}\langle k,x\rangle} + \left(u(x) - \sum_{k\in\mathbb{Z}^d_{\leq M}}\langle\varphi_{-k}, u\rangle\; e^{2\pi\,\mathrm{i}\langle k,x\rangle}\right).$$

We can then write

$$
\begin{aligned}
&\mathrm{DFT}(u^N)(-m) \\
&= \frac{1}{N^d} \sum_{x \in \mathrm{G}} \left(\sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} + \left(u(x) - \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right) \right) e^{-2\pi \, \mathrm{i} \langle m, x \rangle} \\
&= \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \left(\frac{1}{N^d} \sum_{x \in \mathrm{G}} e^{2\pi \, \mathrm{i} \langle k - m, x \rangle} \right) \\
&+ \frac{1}{N^d} \sum_{x \in \mathrm{G}} \left(u(x) - \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right) e^{-2\pi \, \mathrm{i} \langle m, x \rangle} \\
&= \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\mathrm{mod}\ N)] + \frac{1}{N^d} \sum_{x \in \mathrm{G}} \left(u(x) - \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right) e^{-2\pi \, \mathrm{i} \langle m, x \rangle},
\end{aligned}
$$

where the final equality follows from Lemma 26 as $|m|_\infty < N$. Note that we can swap sums over G and $\mathbb{Z}^d$ in the first term because the sums converge absolutely when $s > d/2$ (see Lemma 25). Thus, we obtain

$$
\begin{aligned}
&|\, \mathrm{DFT}(u^N)(-m) - \langle \varphi_{-m}, u \rangle \,| \\
&\leq \left| \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\mathrm{mod}\ N)] - \langle \varphi_{-m}, u \rangle \right| \\
&+ \left| \frac{1}{N^d} \sum_{x \in \mathrm{G}} \left(u(x) - \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right) e^{-2\pi \, \mathrm{i} \langle m, x \rangle} \right|
\end{aligned}
$$

Using the uniform bound over $x \in \mathrm{G}$ for the second term and the following identity for the first term

$$
\sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\mathrm{mod}\ N)] - \langle \varphi_{-m}, u \rangle = \sum_{k \in \mathbb{Z}^d_{\leq M} \setminus \{m\}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\mathrm{mod}\ N)],
$$

we obtain

$$
\begin{aligned}
&|\, \mathrm{DFT}(u^N)(-m) - \langle \varphi_{-m}, u \rangle \,| \\
&\leq \left| \sum_{k \in \mathbb{Z}^d_{\leq M} \setminus \{m\}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\mathrm{mod}\ N)] \right| + \sup_{x \in \mathrm{G}} \left| u(x) - \sum_{k \in \mathbb{Z}^d_{\leq M}} \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right|
\end{aligned}
$$

Recall that we have (i) $|\, \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle}| \leq B$ and $\sum_{k \in \mathbb{Z}^d} \left| \langle \varphi_{-k}, u \rangle \; e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right| < \infty$ for

$s > d/2$ using Lemma 25. The Weierstrass M-test implies that the second term converges to 0 uniformly over $x \in \mathbb{T}^d$ as $M \to \infty$. Thus, we obtain

$$\sum_{k \in \mathbb{Z}^d_{\leq M} \setminus \{m\}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\text{mod } N)] \xrightarrow[M \to \infty]{} \sum_{k \in \mathbb{Z}^d \setminus \{m\}} \langle \varphi_{-k}, u \rangle \; \mathbb{1}[k \equiv m (\text{mod } N)]$$
$$= \sum_{\ell \in \mathbb{Z}^d \setminus \{\mathbf{0}\}} \left\langle \varphi_{-(\ell N + m)}, u \right\rangle,$$

which completes our proof.

Lemma 28. *For any $s \in \mathbb{N}$ such that $s > d/2$, we have*

$$\sum_{k \in \mathbb{Z}^d \setminus \{\mathbf{0}\}} \frac{1}{|k|_\infty^{2s}} \leq \pi^2 \, 3^{d-2}.$$

Proof. Recall that $|\{m \in \mathbb{Z}^d \, : \, |m|_\infty = j\}| = 2(2j+1)^{d-1}$. This is because one of the entry of m has to be $\pm j$ and other $d-1$ entries could be anything in $\{-j \ldots, -1, 0, 1, \ldots, j\}$. Thus,

$$\sum_{\ell \in \mathbb{Z}^d \setminus \{\mathbf{0}\}} \frac{1}{|\ell|_\infty^{2s}} \leq \sum_{j=1}^{\infty} \frac{2(2j+1)^{d-1}}{j^{2s}} \leq 2 \cdot 3^{d-1} \sum_{j=1}^{\infty} \frac{1}{j^{2s-d+1}} \leq 2 \cdot 3^{d-1} \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{2 \cdot 3^{d-1} \pi^2}{6} = \pi^2 \, 3^{d-2}.$$

The third inequality uses $2s - d \leq 1$ as $s > d/2$ and $s \in \mathbb{N}$. ■

■

E.4 Proof of Upper Bound (Theorem 25)

Before we prove Theorem 25, we need some notation. For any $T \in \mathcal{T}$ such that $T = \sum_{m \in \mathbb{Z}^d} \lambda_m \, \varphi_m \otimes \varphi_{-m}$, we define

$$r(T) := \mathop{\mathbb{E}}_{(v,w) \sim \mu} \left[\|Tv - w\|_{L^2}^2 \right] = \mathop{\mathbb{E}}_{(v,w) \sim \mu} \left[\sum_{m \in \mathbb{Z}^d} |\lambda_m \langle \varphi_{-m}, v \rangle - \langle \varphi_{-m}, w \rangle|^2 \right]$$
$$\widehat{r}(T) := \frac{1}{n} \sum_{i=1}^{n} \|Tv_i - w_i\|_{L^2}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{m \in \mathbb{Z}^d} |\lambda_m \langle \varphi_{-m}, v_i \rangle - \langle \varphi_{-m}, w_i \rangle|^2$$

where $\{(v_i, w_i)\}_{i=1}^n$ is the sample accessible to the learner on a uniform grid of $[0,1]^d$. Then, using these definitions, we can write

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) = \mathbb{E}\left[r(\widehat{T}_K^N) - \inf_{T \in \mathcal{T}} r(T) \right] = \mathbb{E}\left[r(\widehat{T}_K^N) - \inf_{T \in \mathcal{T}_K} r(T) \right] + \inf_{T \in \mathcal{T}_K} r(T) - \inf_{T \in \mathcal{T}} r(T),$$

where $\mathcal{T}_K$ is the truncated class defined as

$$\mathcal{T}_K := \left\{ \sum_{m \in \mathbb{Z}^d_{\leq K}} \lambda_m \, \varphi_m \otimes \varphi_{-m} \,\middle|\, \sup_{m \in \mathbb{Z}^d_{\leq K}} |\lambda_m| \leq C \right\}.$$

Furthermore, defining

$$\widehat{T}_K \in \operatorname*{arg\,min}_{T \in \mathcal{T}_K} \widehat{r}(T),$$

we can decompose

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) = \underbrace{\mathbb{E}\left[r(\widehat{T}_K^N) - r(\widehat{T}_K)\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[r(\widehat{T}_K) - \inf_{T \in \mathcal{T}_K} r(T)\right]}_{\text{(II)}} + \underbrace{\inf_{T \in \mathcal{T}_K} r(T) - \inf_{T \in \mathcal{T}} r(T)}_{\text{(III)}}.$$

First, it is easy to see that

$$\text{(III)} \leq \sup_{T \in \mathcal{T}} \inf_{T_K \in \mathcal{T}_K} |r(T) - r(T_K)|.$$

To upper bound (II), let $T_K^\star \in \mathcal{T}_K$ such that $r(T_K^\star) = \inf_{T \in \mathcal{T}_K} r(T)$. Formally, for every $\varepsilon > 0$, we may only be guaranteed the existence of $T_K^\star$ such that $r(T_K^\star) \leq \inf_{T \in \mathcal{T}_K} r(T) + \varepsilon$. However, as ε can be made arbitrarily small, we can just choose it to be smaller than any error bound we obtain at the end. So, the arguments below are rigorously justified.

Given such $T_K^\star$, we can write

$$\text{(II)} = \mathbb{E}[r(\widehat{T}_K) - r(T_K^\star)] = \mathbb{E}[r(\widehat{T}_K) - \widehat{r}(\widehat{T}_K)] + \mathbb{E}[\widehat{r}(\widehat{T}_K) - \widehat{r}(T_K^\star)] + \mathbb{E}[\widehat{r}(T_K^\star) - r(T_K^\star)].$$

The last term of the sum vanishes because $\mathbb{E}[\widehat{r}(T_K^\star)] = r(T_K^\star)$. As for the second term, $\widehat{T}_K$ minimizes empirical loss over the samples, implying $\widehat{r}(\widehat{T}_K) \leq \widehat{r}(T_K^\star)$. For the first term, we use the trivial bound $r(\widehat{T}_K) - \widehat{r}(\widehat{T}_K) \leq \sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}(T)|$. Overall, we obtain

$$\text{(II)} \leq \mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}(T)|\right].$$

Finally, we upper bound the term (I). Given K and N, for any $T \in \mathcal{T}_K$ such that $T = \sum_{m \in \mathbb{Z}^d_{\leq K}} \lambda_m \, \varphi_m \otimes \varphi_{-m}$, define

$$\widehat{r}_N(T) := \frac{1}{n} \sum_{i=1}^n \sum_{m \in \mathbb{Z}^d_{\leq K}} \left| \lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m) \right|^2 + \frac{1}{n} \sum_{i=1}^n \sum_{m \in \mathbb{Z}^d_{> K}} |\, \langle \varphi_{-m}, w_i \rangle \,|^2.$$

Technically, the term $\widehat{r}_N(T)$ also depends on K, but we drop K to avoid cluttered notation. Here, the first term above is the empirical DFT-based least squares loss of T define in 6.4.2. The second term is introduced purely for technical reasons to make our calculations work (see Section E.4.2). Since the second term does not depend on T, our estimator $\widehat{T}_K^N$ is still the operator obtained by minimizing $\widehat{r}_N$. Then, note that

$$\text{(I)} = \mathbb{E}[r(\widehat{T}_K^N) - \widehat{r}_N(\widehat{T}_K^N)] + \mathbb{E}[\widehat{r}_N(\widehat{T}_K^N) - \widehat{r}_N(\widehat{T}_K)] + \mathbb{E}[\widehat{r}_N(\widehat{T}_K) - r(\widehat{T}_K)]$$

Note that the second term above satisfies $\widehat{r}_N(\widehat{T}_K^N) - \widehat{r}_N(\widehat{T}_K) \leq 0$ almost surely because $\widehat{T}_K^N$ minimizes $\widehat{r}_N(T)$ over all $T \in \mathcal{T}_K$. For the first and the third term, we use the bound

$$\mathbb{E}[r(\widehat{T}_K^N) - \widehat{r}_N(\widehat{T}_K^N)] \leq \mathbb{E}[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}_N(T)|]$$

and

$$\mathbb{E}[\widehat{r}_N(\widehat{T}_K) - r(\widehat{T}_K)] \leq \mathbb{E}[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}_N(T)|].$$

Thus, we have

$$\text{(I)} \leq 2\,\mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}_N(T)|\right] \leq 2\,\mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}(T)|\right] + 2\,\mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |\widehat{r}(T) - \widehat{r}_N(T)|\right],$$

where the final step uses the triangle inequality. Combining everything, we have established that

$$\begin{aligned}
&\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu) \\
&\leq 3\,\mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |r(T) - \widehat{r}(T)|\right] + 2\,\mathbb{E}\left[\sup_{T \in \mathcal{T}_K} |\widehat{r}(T) - \widehat{r}_N(T)|\right] + \sup_{T \in \mathcal{T}} \inf_{T_K \in \mathcal{T}_K} |r(T) - r(T_K)|
\end{aligned}$$

The first term is the statistical error, the second is the discretization error, and the final is the truncation error. Next, we bound each of these terms individually.

E.4.1 Upper bound on the truncation error $\sup_{T \in \mathcal{T}} \inf_{T_K \in \mathcal{T}_K} |r(T) - r(T_K)|$

Pick any $T \in \mathcal{T}$. Then, there exists a sequence $\{\lambda_m\}_{m \in \mathbb{Z}^d}$ such that $T = \sum_{m \in \mathbb{Z}^d} \lambda_m \, \varphi_m \otimes \varphi_{-m}$. Define

$$T_K := \sum_{m \in \mathbb{Z}^d_{\leq K}} \lambda_m \, \varphi_m \otimes \varphi_{-m}.$$

Clearly, $T_K \in \mathcal{T}_K$. Then, we have

$$\begin{aligned} r(T) - r(T_K) &= \mathop{\mathbb{E}}_{(v,w)\sim\mu} [\|Tv - w\|_{L^2}^2 - \|T_K v - w\|_{L^2}^2] \\ &= \mathop{\mathbb{E}}_{(v,w)\sim\mu} [\|Tv\|_{L^2}^2 - \|T_K v\|_{L^2}^2 + 2\langle (T_K - T)v, w\rangle] \\ &\leq \mathop{\mathbb{E}}_{(v,w)\sim\mu} \left[\sum_{m\in\mathbb{Z}^d_{>K}} |\lambda_m|^2 |\langle \varphi_{-m}, v\rangle|^2 + 2 \sum_{m\in\mathbb{Z}^d_{>K}} \left| \lambda_m \langle \varphi_{-m}, v\rangle \langle \varphi_m, w\rangle \right| \right] \end{aligned}$$

The final equality uses the following facts. First, we have

$$\|Tv\|_{L^2}^2 = \left\| \sum_{m\in\mathbb{Z}^d} \lambda_m \langle \varphi_{-m}, v\rangle \varphi_m \right\|_{L^2}^2 = \sum_{m\in\mathbb{Z}^d} |\lambda_m|^2 |\langle \varphi_{-m}, v\rangle|^2.$$

Analogously,

$$\|T_K v\|_{L^2}^2 = \sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m|^2 |\langle \varphi_{-m}, v\rangle|^2.$$

As for the second term, we use

$$\langle (T_K - T)v, w\rangle = \left\langle \sum_{m\in\mathbb{Z}^d_{>K}} \lambda_m \langle \varphi_{-m}, v\rangle \varphi_m, w \right\rangle = \sum_{m\in\mathbb{Z}^d_{>K}} \lambda_m \langle \varphi_{-m}, v\rangle \langle \varphi_m, w\rangle .$$

Next, using the fact that $|\lambda_m| \leq C$ followed by Lemma 24, the first term is

$$\sum_{m\in\mathbb{Z}^d_{>K}} |\lambda_m|^2 |\langle \varphi_{-m}, v\rangle|^2 \leq \frac{B^2 C^2}{K^{2s}}.$$

As for the second term, using $|\lambda_m| \leq C$ followed by Cauchy-Schwarz implies

$$2 \sum_{m\in\mathbb{Z}^d_{>K}} |\lambda_m \langle \varphi_{-m}, v\rangle \langle \varphi_m, w\rangle| \leq 2C \sqrt{\sum_{m\in\mathbb{Z}^d_{>K}} |\langle \varphi_{-m}, v\rangle|^2} \sqrt{\sum_{m\in\mathbb{Z}^d_{>K}} |\langle \varphi_m, w\rangle|^2} \leq \frac{2CB^2}{K^{2s}},$$

where the final inequality holds because of Lemma 24. Since $T \in \mathcal{T}$ is arbitrary, we have shown that

$$\sup_{T\in\mathcal{T}} \inf_{T_K\in\mathcal{T}_K} |r(T) - r(T_K)| \leq \frac{B^2 C(C+2)}{K^{2s}} \leq \frac{B^2 (C+1)^2}{K^{2s}}.$$

E.4.2 Upper bound on the discretization error

Fix $T \in \mathcal{T}_K$. Then, there exists $\{\lambda_m\}_{m\in\mathbb{Z}^d_{\leq K}}$ with $|\lambda_m| \leq C$ such that $T = \sum_{\mathbb{Z}^d_{\leq K}} \lambda_m \, \varphi_m \otimes \varphi_{-m}$. Then, recall that

$$\widehat{r}_N(T) := \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{\leq K}} \left|\lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m)\right|^2 + \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{>K}} |\,\langle \varphi_{-m}, w_i\rangle\,|^2.$$

Moreover, we also have

$$\widehat{r}(T) = \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle|^2 + \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{>K}} |\,\langle \varphi_{-m}, w_i\rangle\,|^2,$$

which yields

$$\begin{aligned}
&\widehat{r}_N(T) - \widehat{r}(T) \\
&= \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{\leq K}} \left(\left|\lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m)\right|^2 - |\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle|^2\right).
\end{aligned}$$

Next, we define

$$\alpha_{im} = \mathrm{DFT}(v_i^N)(-m) - \langle \varphi_{-m}, v_i\rangle \quad \text{and} \quad \beta_{im} = \mathrm{DFT}(w_i^N)(-m) - \langle \varphi_{-m}, w_i\rangle\,.$$

We can then write

$$\begin{aligned}
&\left|\lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m)\right|^2 \\
&= |\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle + \lambda_m \, \alpha_{im} - \beta_{im}|^2 \\
&\leq |\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle\,|^2 \\
&+ 2|\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle\,| \; |\lambda_m \alpha_{im} - \beta_{im}| \; + |\lambda_m \alpha_{im} - \beta_{im}|^2.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
&|\widehat{r}_N(T) - \widehat{r}(T)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathbb{Z}^d_{\leq K}} \left(2|\lambda_m \langle \varphi_{-m}, v_i\rangle - \langle \varphi_{-m}, w_i\rangle\,| \, |\lambda_m \alpha_{im} - \beta_{im}| \; + |\lambda_m \alpha_{im} - \beta_{im}|^2\right) \\
&\leq \max_{i\in[n]} \sum_{m\in\mathbb{Z}^d_{\leq K}} 2 \; (|\lambda_m \langle \varphi_{-m}, v_i\rangle\,| + |\,\langle \varphi_{-m}, w_i\rangle\,|) \; |\lambda_m \alpha_{im} - \beta_{im}|\,| + |\lambda_m \alpha_{im} - \beta_{im}|^2.
\end{aligned}$$

Next, using Cauchy-Schwarz inequality, the first term of the summand can be bounded

as

$$\begin{aligned}
&\sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m \langle \varphi_{-m}, v_i\rangle| \, |\lambda_m\alpha_{im} - \beta_{im}| \\
&\qquad \leq \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m|^2(1+|m|_\infty^{2s}) \, |\langle \varphi_{-m}, v_i\rangle|^2} \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} \frac{|\lambda_m\alpha_{im} - \beta_{im}|^2}{1+|m|_\infty^{2s}}} \\
&\qquad \leq B\,C \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} \frac{|\lambda_m\alpha_{im} - \beta_{im}|^2}{1+|m|_\infty^{2s}}},
\end{aligned}$$

where the final inequality uses Lemma 24 and the fact that $|\lambda_m| \leq C$. Similar arguments show that

$$\sum_{m\in\mathbb{Z}^d_{\leq K}} |\langle \varphi_{-m}, w_i\rangle| \, |(\lambda_m\alpha_{im} - \beta_{im})| \leq B \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} \frac{|\lambda_m\alpha_{im} - \beta_{im}|^2}{1+|m|_\infty^{2s}}}.$$

Overall, we have shown that

$$|\widehat{r}_N(T) - \widehat{r}(T)| \leq \max_{i\in[n]} \left(2B(C+1) \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} \frac{|\lambda_m\alpha_{im} - \beta_{im}|^2}{1+|m|_\infty^{2s}}} + \sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m\alpha_{im} - \beta_{im}|^2 \right).$$

Now, recall that Lemma 27 implies

$$\max\{|\alpha_{im}|, |\beta_{im}|\} \leq \left| \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \left\langle \varphi_{-(\ell N+m)}, u \right\rangle \right|,$$

which subsequently yields

$$|\lambda_m\alpha_{im} - \beta_{im}|^2 \leq (C+1)^2 \left| \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \left\langle \varphi_{-(\ell N+m)}, u \right\rangle \right|^2.$$

Thus, we have

$$\begin{aligned}
&\sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m\alpha_{im}-\beta_{im}|^2 \\
&\leq (C+1)^2 \sum_{m\in\mathbb{Z}^d_{\leq K}} \left|\sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \left\langle \varphi_{-(\ell N+m)}, u\right\rangle\right|^2 \\
&\leq (C+1)^2 \left(\sum_{m\in\mathbb{Z}^d_{\leq K}} \left(\sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \frac{1}{1+|m+\ell N|_\infty^{2s}}\right) \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} (1+|m+\ell N|_\infty^{2s})\,|\left\langle \varphi_{-(\ell N+m)}, u\right\rangle|^2\right).
\end{aligned}$$

where the final step follows from Cauchy-Schwarz inequality.

To upper bound the first sum within inner parenthesis, note that $|m+\ell N|_\infty \geq |\ell N|_\infty - |m|_\infty \geq N|\ell|_\infty - N/2 \geq N/2\,|\ell|_\infty$. Here, we use the fact that $|m|_\infty \leq K \leq N/2$. So, we have

$$\sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \frac{1}{1+|m+\ell N|_\infty^{2s}} \leq \left(\frac{2}{N}\right)^{2s} \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \frac{1}{|\ell|_\infty^{2s}} \leq \frac{2^{2s}\pi^2\,3^{d-2}}{N^{2s}},$$

where the final inequality uses Lemma 28. Next, note that

$$\sum_{m\in\mathbb{Z}^d_{\leq K}} \sum_{\ell\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} (1+|m+\ell N|_\infty^{2s})\,|\left\langle \varphi_{-(\ell N+m)}, u\right\rangle|^2 \leq \sum_{k\in\mathbb{Z}^d} (1+|k|_\infty^{2s})\,|\langle \varphi_{-k}, u\rangle|^2 \leq B^2,$$

where the second inequality follows from Lemma 24. The first inequality holds because for each $k\in\mathbb{Z}^d$, we have $|\{(m,\ell) : m+\ell N = k,\, m\in\mathbb{Z}^d_{\leq K} \text{ and } \ell\in\mathbb{Z}^d\setminus\{0\}\}| \leq 1$. That is, for each $k\in\mathbb{Z}^d$, there is only one possible pair (m,ℓ) such that $k = m+\ell N$. Suppose, for the sake of contradiction, there exists $k\in\mathbb{Z}^d$ such that two distinct pairs exist in the set, namely (m_1,ℓ_1) and (m_2,ℓ_2). Note that $m_1+\ell_1 N-(m_2+\ell_2 N) = k-k = 0$, which implies $(m_1-m_2) = (\ell_2-\ell_1)N$. Clearly, we cannot have $\ell_2=\ell_1$, otherwise, we will have $m_2=m_1$, contradicting the fact that there are two distinct pairs. So, we must have $\ell_2\neq\ell_1$. That is, $|\ell_2-\ell_1|_\infty \geq 1$, and thus $|m_1-m_2|_\infty \geq N$. Moreover, $|m_1-m_2|_\infty \leq |m_1|_\infty + |m_2|_\infty \leq 2K$, which implies that $2K\geq N$. This contradicts the fact that $K < N/2$. Therefore, overall, we have shown that

$$\sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m\alpha_{im}-\beta_{im}|^2 \leq \frac{2^{2s}\pi^2\,3^{d-2}\,B^2(C+1)^2}{N^{2s}}.$$

Next, we have

$$\sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} \frac{|\lambda_m\alpha_{im}-\beta_{im}|^2}{1+|m|_\infty^{2s}}} \leq \sqrt{\sum_{m\in\mathbb{Z}^d_{\leq K}} |\lambda_m\alpha_{im}-\beta_{im}|^2} \leq \frac{2^s\pi\sqrt{3^{d-2}}\,B(C+1)}{N^s}.$$

Therefore, by combining everything, we have shown that

$$|\widehat{r}_N(T)-\widehat{r}(T)| \leq \frac{2^{s+1}B^2(C+1)^2}{N^s}\pi\sqrt{3^{d-2}}+\frac{B^2(C+1)^2 4^s}{N^{2s}}\,\pi^2 3^{d-2} \leq 2\,\frac{2^{s+1}B^2(C+1)^2}{N^s}\pi\sqrt{3^{d-2}}.$$

The final inequality holds when $N^s \geq 2^{s-1}\,\pi\sqrt{3^{d-2}}$, which is satisfied as long as $N \geq 6$. As $T \in \mathcal{T}_K$ is arbitrary, we have shown that the discretization error

$$2\,\mathbb{E}\left[\sup_{T\in\mathcal{T}_K} |\widehat{r}(T)-\widehat{r}_N(T)|\right] \leq \frac{2^{s+3}\pi\sqrt{3^{d-2}}B^2(C+1)^2}{N^s} \leq \frac{2^{s+3}\sqrt{\pi^d}B^2(C+1)^2}{N^s}.$$

E.4.3 Upper Bound on the Statistical Error

In fact, we will bound $\mathbb{E}\left[\sup_{T\in\mathcal{T}} |r(T)-\widehat{r}(T)|\right]$. This can be viewed as the limit of the statistical error as $K \to \infty$. To that end, let $\sigma_1,\ldots,\sigma_n$ denote iid random variables such that $\sigma_i \sim \text{Uniform}(\{-1,1\})$. Standard symmetrization arguments show that

$$\begin{aligned}\mathbb{E}\left[\sup_{T\in\mathcal{T}} |r(T)-\widehat{r}(T)|\right] &\leq 2\,\mathbb{E}\left[\sup_{T\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \left\|Tv_i-w_i\right\|_{L^2}^2\right|\right] \\ &= 2\,\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\leq C}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i \sum_{m\in\mathbb{Z}^d} |\lambda_m\langle\varphi_{-m},v_i\rangle - \langle\varphi_{-m},w_i\rangle|^2\right|\right]\end{aligned}$$

Note that

$$\begin{aligned}&|\lambda_m\langle\varphi_{-m},v_i\rangle - \langle\varphi_{-m},w_i\rangle|^2 \\ &= (\lambda_m\langle\varphi_{-m},v_i\rangle - \langle\varphi_{-m},w_i\rangle)\,\overline{(\lambda_m\langle\varphi_{-m},v_i\rangle - \langle\varphi_{-m},w_i\rangle)} \\ &= |\lambda_m|^2\,|\langle\varphi_{-m},v_i\rangle|^2 - \left(\lambda_m\langle\varphi_{-m},v_i\rangle\,\overline{\langle\varphi_{-m},w_i\rangle} + \overline{\lambda_m}\langle\varphi_{-m},w_i\rangle\,\overline{\langle\varphi_{-m},v_i\rangle}\right) + |\langle\varphi_{-m},w_i\rangle|^2.\end{aligned}$$

The first and the last term above are real numbers, so the term in the parenthesis must also be a real number. Using triangle inequality, the term Rademacher sum above can be

upper-bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}|\lambda_m\langle\varphi_{-m},v_i\rangle-\langle\varphi_{-m},w_i\rangle|^2\right|\right]\\
&\quad\le\underbrace{\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}|\lambda_m|^2\,|\langle\varphi_{-m},v_i\rangle|^2\right|\right]}_{\text{(i)}}\\
&\quad+\underbrace{\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}\lambda_m\langle\varphi_{-m},v_i\rangle\overline{\langle\varphi_{-m},w_i\rangle}\right|\right]}_{\text{(ii)}}\\
&\quad+\underbrace{\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}\overline{\lambda_m}\langle\varphi_{-m},w_i\rangle\overline{\langle\varphi_{-m},v_i\rangle}\right|\right]}_{\text{(iii)}}+\underbrace{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}|\langle\varphi_{-m},w_i\rangle|^2\right|\right]}_{\text{(iv)}}.
\end{aligned}$$

Let us start with the term (iv) first. Swapping the sum over m and i and using triangle inequality yields

$$\begin{aligned}
\text{(iv)}=\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}|\langle\varphi_{-m},w_i\rangle|^2\right|\right]&\le\sum_{m\in\mathbb{Z}^d}\frac{1}{n}\mathbb{E}\left[\left|\sum_{i=1}^n\sigma_i|\langle\varphi_{-m},w_i\rangle|^2\right|\right]\\
&\le\sum_{m\in\mathbb{Z}^d}\frac{1}{n}\left(\sum_{i=1}^n|\langle\varphi_{-m},w_i\rangle|^4\right)^{1/2},
\end{aligned}$$

where the final step follows from Khintchine's inequality. Note that swapping the sums is justified because both sums converge absolutely.

For the term (iii), swapping the sum over m and i and using the fact that $|\lambda_m|\le C$ yields

$$\begin{aligned}
\text{(iii)}&=\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\sum_{m\in\mathbb{Z}^d}\overline{\lambda_m}\langle\varphi_{-m},w_i\rangle\overline{\langle\varphi_{-m},v_i\rangle}\right|\right]\\
&=\mathbb{E}\left[\sup_{|\lambda|_{\ell^\infty}\le C}\left|\frac{1}{n}\sum_{m\in\mathbb{Z}^d}\overline{\lambda_m}\sum_{i=1}^n\sigma_i\langle\varphi_{-m},w_i\rangle\overline{\langle\varphi_{-m},v_i\rangle}\right|\right]\\
&\le C\,\mathbb{E}\left[\sum_{m\in\mathbb{Z}^d}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\langle\varphi_{-m},w_i\rangle\overline{\langle\varphi_{-m},v_i\rangle}\right|\right]\\
&\le C\sum_{m\in\mathbb{Z}^d}\frac{1}{n}\left(\sum_{i=1}^n|\langle\varphi_{-m},w_i\rangle\overline{\langle\varphi_{-m},v_i\rangle}|^2\right)^{1/2},
\end{aligned}$$

where the final step uses Khintchine's inequality. Since $|\lambda_m|\le C$, we can use the same

arguments to show that

$$\text{(ii)} \leq C \sum_{m \in \mathbb{Z}^d} \frac{1}{n} \left(\sum_{i=1}^{n} | \langle \varphi_{-m}, v_i \rangle \overline{\langle \varphi_{-m}, w_i \rangle} |^2 \right)^{1/2},$$

and

$$\text{(i)} \leq C^2 \sum_{m \in \mathbb{Z}^d} \frac{1}{n} \left(\sum_{i=1}^{n} | \langle \varphi_{-m}, v_i \rangle |^4 \right)^{1/2}.$$

Next, note that we can bound $| \langle \varphi_0, u \rangle | \leq B$ for all $\|u\|_{\mathcal{H}^s} \leq B$. Moreover, Lemma 22 implies that $| \langle \varphi_{-m}, u \rangle | \leq \frac{B}{(2\pi)^s |m|_\infty^s}$ for all $m \neq \mathbf{0}$. Thus, we obtain the bound

$$\begin{aligned}
\text{(i)} &\leq \frac{B^2 C^2}{\sqrt{n}} + C^2 \sum_{m \in \mathbb{Z}^d \setminus \{\mathbf{0}\}} \frac{1}{n} \left(\sum_{i=1}^{n} \frac{B^4}{(2\pi)^{4s}} \frac{1}{|m|_\infty^{4s}} \right)^{1/2} \\
&\leq B^2 C^2 \frac{1}{\sqrt{n}} + \frac{B^2 C^2}{(2\pi)^{2s}} \frac{1}{\sqrt{n}} \sum_{m \in \mathbb{Z}^d \setminus \{\mathbf{0}\}} \frac{1}{|m|_\infty^{2s}} \\
&\leq B^2 C^2 \frac{1}{\sqrt{n}} + \frac{B^2 C^2 \pi^2 3^{d-2}}{(2\pi)^{2s}} \frac{1}{\sqrt{n}},
\end{aligned}$$

where the final inequality uses Lemma 28. Similar calculations can be done to show that

$$\text{(ii)}, \text{(iii)} \leq B^2 C \frac{1}{\sqrt{n}} + \frac{B^2 C \pi^2 \, 3^{d-2}}{(2\pi)^{2s}} \frac{1}{\sqrt{n}} \quad \text{and} \quad \text{(iv)} \leq B^2 \frac{1}{\sqrt{n}} + \frac{B^2 \pi^2 3^{d-2}}{(2\pi)^{2s}} \frac{1}{\sqrt{n}}.$$

Thus, we have overall shown that

$$\begin{aligned}
\mathbb{E}\left[\sup_{T \in \mathcal{T}} |r(T) - \widehat{r}(T)| \right] &\leq 2 \left(\text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)} \right) \\
&\leq 2(B^2 C^2 + 2B^2 C + B^2) \left(1 + \frac{\pi^2 3^{d-2}}{(2\pi)^{2s}} \right) \frac{1}{\sqrt{n}} \\
&= \frac{2B^2 (C+1)^2}{\sqrt{n}} \left(1 + \frac{\pi^2 3^{d-2}}{(2\pi)^{2s}} \right) \\
&\leq \frac{5}{2} \frac{B^2 (C+1)^2}{\sqrt{n}}
\end{aligned}$$

where we use the fact that

$$\frac{\pi^2 3^{d-2}}{(2\pi)^{2s}} \leq \frac{1}{2^{2s}} \frac{\pi^d}{\pi^{2s}} \leq \frac{1}{2^{2s}} \leq \frac{1}{4}$$

as $2s > d$ and $s \geq 1$. Therefore, the overall statistical error is

$$3\,\mathbb{E}\left[\sup_{T\in\mathcal{T}} |r(T) - \widehat{r}(T)|\right] \leq \frac{8B^2(C+1)^2}{\sqrt{n}}.$$

E.5 Proof of Lower Bound (Theorem 26)

Proof. To define a difficult distribution for the learner, we need some notations. Let

$$\psi_0 = \varphi_0 \quad \text{and } \psi_m = \frac{1}{\sqrt{2}}\,(\varphi_{-m} + \varphi_m) \quad \text{for } m \in \mathbb{Z}^d\backslash\{\mathbf{0}\}.$$

Note that $\psi_m : \mathbb{T}^d \to \mathbb{R}$ is a *real-valued* function such that $\|\psi_m\|_{L^2} = 1$. We work with ψ_m's to ensure that the distribution is only supported over real-valued functions. For any $\{\lambda_k\}_{k\in\mathbb{Z}^d}$ such that $\lambda_k = \lambda_{-k} \in \mathbb{R}$, the operator $T = \sum_{m\in\mathbb{Z}^d} \lambda_m \varphi_m \otimes \varphi_{-m}$ satisfies

$$T\psi_m = \frac{1}{\sqrt{2}}(\lambda_m\,\varphi_m + \lambda_{-m}\varphi_{-m}) = \frac{\lambda_m}{\sqrt{2}}\,(\varphi_{-m} + \varphi_m) = \lambda_m\psi_m \quad \forall m \in \mathbb{Z}^d\backslash\{\mathbf{0}\}.$$

Clearly, $T\psi_0 = \lambda_0\psi_0$. Next, let us define a sequence $\{\gamma_m\}_{m\in\mathbb{Z}^d}$ such that

$$\gamma_0 = \frac{B}{\sqrt{s+1}} \quad \text{and} \quad \gamma_m = \frac{B}{\sqrt{s+1}\,|m|_\infty^s} \quad \forall m \in \mathbb{Z}^d\backslash\{\mathbf{0}\}.$$

Finally, define a set

$$\mathcal{J} = \{m \in \mathbb{Z}^d : \ m_1 \in \mathbb{N} \text{ and } m_j = 0 \quad \forall j \neq 1\}.$$

For any $M, N \in \mathbb{N}$, define $\mathcal{J}_M^N = \{m \in \mathcal{J} : m_1 \not\equiv 0 \pmod N \text{ and } m_1 \leq M\}$. Let $r \in \mathbb{Z}^d$ such that $r \in \mathcal{J}$ and $r_1 = 1$. That is, $r = (1, 0, 0, \ldots, 0)$. For any $q \in \mathbb{Z}$, we write $qr = (q, 0, 0, \ldots, 0)$.

We now describe a difficult distribution for the learner. To that end, first draw a $\xi := \{\xi_m\}_{m\in\mathbb{Z}^d}$ such that $\xi_m = \xi_{-m}$ is drawn from Uniform($\{-1, 1\}$). Then, given such ξ, let μ_ξ be any joint distribution on $\mathcal{V} \times \mathcal{W}$ such that its marginal on $\mathcal{V}$ assigns $1/3$ mass uniformly on $\left\{\gamma_m\psi_m : m \in \mathcal{J}_M^N\right\}$, $1/3$ mass on $\gamma_0\psi_0$, and the remaining $1/3$ mass on $\gamma_{(K+j)r}\,\psi_{(K+j)r}$ for either $j = 1$ or $j = 2$ ensuring that $K + j \not\equiv 0 \pmod N$. Moreover, given a $v = \gamma_k\psi_k$ drawn from the marginal of μ_ξ, assign $w \mid v$ to be $\xi_k\gamma_k\psi_k$ if $k \neq 0$. On the other hand, if $k = 0$, then $w \mid v$ is $\xi_{Nr}\,\gamma_{Nr}\,\psi_{Nr}$.

This is a valid distribution as

$$\begin{aligned}\|v\|_{\mathcal{H}^s}^2 = \sum_{k\in\mathbb{N}_0^d\,:\,|k|_\infty\le s} \left\|\partial^k v\right\|_{L^2}^2 &= \sum_{k\in\mathbb{N}_0^d\,:\,|k|_\infty\le s} (m_1^{k_1}\gamma_m)^2 \mathbb{1}[k_j = 0 \text{ for all } j\neq 1] \\ &= \gamma_m^2 \sum_{k_1=0}^{s} |m|_\infty^{2k_1} \\ &\le (s+1)\gamma_m^2 |m|_\infty^{2s} \\ &\le B^2\end{aligned}$$

Similar arguments show that $\|w\|_{\mathcal{H}^s}^2 \le B^2$.

Next, we establish that

$$\mathop{\mathbb{E}}_{\xi}\left[\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu_\xi)\right] \ge \frac{B^2}{3(s+1)}\left(\frac{1}{8n} + \frac{2}{(K+2)^{2s}} + \frac{1}{N^{2s}}\right).$$

Since the lower bound above holds in expectation, we can use the probabilistic method to argue that there must exist a sequence $\xi^\star$ such that

$$\mathcal{E}_n(\widehat{T}_K^N, \mathcal{T}, \mu_{\xi^\star}) \ge \frac{B^2}{3(s+1)}\left(\frac{1}{8n} + \frac{2}{(K+2)^{2s}} + \frac{1}{N^{2s}}\right).$$

We now proceed with the proof of the claimed lowerbound. Let $\widehat{T}_K^N$ denote the estimator produced by the algorithm. Then, there exists $\{\widehat{\lambda}_m\}_{m\in\mathbb{Z}^d_{\le K}}$ such that

$$\widehat{T}_K^N = \sum_{m\in\mathbb{Z}^d_{\le K}} \widehat{\lambda}_m\ \varphi_m \otimes \varphi_{-m}.$$

For convenience, we will extend the sum to the entire $\mathbb{Z}^d$ and write $\widehat{T}_K^N = \sum_{m\in\mathbb{Z}^d} \widehat{\lambda}_m\ \varphi_m \otimes \varphi_{-m}$, where $\widehat{\lambda}_m = 0$ for all $m \in \mathbb{Z}^d_{>K}$.

Given a ξ, we now lowerbound the expected loss of $\widehat{T}_K^N$ on μ_ξ. Using the definition of the

distribution μ_ξ, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\|\widehat{T}_K^N v - w\|_{L^2}^2\right] \\
&= \frac{1}{3}\frac{1}{|\mathcal{J}_M^N|}\sum_{m\in\mathcal{J}_M^N}\left(\widehat{\lambda}_m - \xi_m\right)^2\gamma_m^2 + \frac{1}{3}\left\|\widehat{\lambda}_0\gamma_0\psi_0 - \xi_{Nr}\,\gamma_{Nr}\,\psi_{Nr}\right\|_{L^2}^2 + \frac{1}{3}\left\|\mathbf{0} - \gamma_{(K+j)r}\,\psi_{(K+j)r}\right\|_{L^2}^2 \\
&\geq \frac{1}{3|\mathcal{J}_M^N|}\sum_{m\in\mathcal{J}_M^N}\gamma_m^2\,\mathbb{1}[\widehat{\lambda}_m\xi_m \leq 0] + \frac{\widehat{\lambda}_0^2\gamma_0^2 + \gamma_{Nr}^2}{3} + \frac{\gamma_{(K+j)r}^2}{3} \\
&\geq \frac{\gamma_r^2}{3|\mathcal{J}_M^N|}\,\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0] + \frac{\widehat{\lambda}_0^2\gamma_0^2 + \gamma_{Nr}^2}{3} + \frac{\gamma_{(K+2)r}^2}{3}.
\end{aligned}
$$

Here, the first inequality use the fact that $(\widehat{\lambda}_m - \xi_m)^2 \geq 1$ whenever $\widehat{\lambda}_m\xi_m \leq 0$ and $\langle e_0, e_{Nr}\rangle_{L^2} = 0$. The second inequality uses the fact that $r \in \mathcal{J}_M^N$ as long as $M, N > 1$ and that $\gamma_{(K+j)r}^2 \geq \gamma_{(K+2)r}^2$ for $j \in \{1, 2\}$.

Next, we establish the upper bound on the loss of the best-fixed operator. Given ξ, define an operator

$$T_\xi = \sum_{m\in\mathbb{Z}_{>0}^d}\xi_m\,\varphi_m \otimes \varphi_{-m}.$$

Clearly,

$$
\begin{aligned}
&\inf_{T\in\mathcal{T}}\mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\left\|Tv - w\right\|_{L^2}^2\right] \\
&\quad\leq \mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\left\|T_\xi v - w\right\|_{L^2}^2\right] \\
&\quad= \mathbb{E}\left[\left\|T_\xi v - w\right\|_{L^2}^2\right\|\left.v = \gamma_0\psi_0\right]\mathbb{P}[v = \gamma_0\psi_0] + \mathbb{E}\left[\left\|T_\xi v - w\right\|_{L^2}^2\right\|\left.v \neq \gamma_0\psi_0\right]\mathbb{P}[v \neq \gamma_0\psi_0] \\
&\quad\leq \|0 - \xi_{Nr}\,\gamma_{Nr}\,\psi_{Nr}\|_{L^2}^2\;\frac{1}{3} \\
&\quad\leq \frac{\gamma_{Nr}^2}{3},
\end{aligned}
$$

where we use the fact that $T_\xi v = 0$ whenever $v = \gamma_0 e_0$ and $T_\xi v = w$ otherwise. Overall, we have shown that

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\|\widehat{T}_K^N v - w\|_{L^2}^2\right] - \inf_{T\in\mathcal{T}}\mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\left\|Tv - w\right\|_{L^2}^2\right] \\
&\qquad\geq \frac{\gamma_r^2}{3|\mathcal{J}_M^N|}\,\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0] + \frac{\widehat{\lambda}_0^2\gamma_0^2 + \gamma_{Nr}^2}{3} + \frac{\gamma_t^2}{3} - \frac{\gamma_{Nr}^2}{3} \\
&\qquad\geq \frac{1}{3(s+1)}\left(\frac{\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0]}{|\mathcal{J}_M^N|} + \widehat{\lambda}_0^2 + \frac{B^2}{(K+2)^{2s}}\right),
\end{aligned}
$$

where the final inequality holds because $\gamma_0 = \gamma_r = \frac{B}{\sqrt{s+1}}$ and $\gamma_{(K+2)r} = \frac{B}{\sqrt{s+1}(K+2)^{2s}}$.

Next, we establish lowerbound of $\widehat{\lambda}_0^2$. To that end, let $S_n = \{(v_i, w_i)\}_{i=1}^n$ denote the n samples accessible to the learner over the uniform grid of size N. Recall our notation $v_i^N := \{v_i(x) : x \in \mathrm{G}\}$ and $w_i^N := \{w_i(x) : x \in \mathrm{G}\}$ for discretized samples. Take a sample $(v_i, w_i) \sim \mu_\xi$. Then, we must have $v_i = \gamma_k \psi_k$ for some $k \in \mathbb{Z}^d$. Consider the case that $k \neq 0$. Then, by definition of the distribution μ_ξ, it must be the case that $k \not\equiv 0 \pmod{\ } \mathrm{N}$. Then, Lemma 26 implies that

$$\mathrm{DFT}(v_i^N)(-0) = \frac{1}{N^d} \sum_{x \in \mathrm{G}} \gamma_k \, \psi_k(x) \, e^{-2\pi \, \mathrm{i} \langle x, 0 \rangle} = \frac{\gamma_k}{\sqrt{2} N^d} \left(\sum_{x \in \mathrm{G}} e^{-2\pi \, \mathrm{i} \langle k, x \rangle} + \sum_{x \in \mathrm{G}} e^{2\pi \, \mathrm{i} \langle k, x \rangle} \right) = 0.$$

On the other hand, if $v_i = \gamma_0 \psi_0$, then we have

$$\mathrm{DFT}(v_i^N)(-0) = \frac{1}{N^d} \sum_{x \in \mathrm{G}} \gamma_0 \psi_0(x) = \frac{\gamma_0}{N^d} \sum_{x \in \mathrm{G}} 1 = \gamma_0.$$

Additionally, when $v_i = \gamma_0 \psi_0$, we have $w_i = \gamma_{Nr} \, \psi_{Nr}$. In this case, Lemma 26 implies that

$$\mathrm{DFT}(w_i^N)(-0) = \frac{\gamma_{Nr}}{N^d} \sum_{x \in \mathrm{G}} \psi_{Nr}(x) = \frac{\gamma_{Nr}}{\sqrt{2} N^d} \left(\sum_{x \in \mathrm{G}} e^{-2\pi \, \mathrm{i} \langle Nr, x \rangle} + \sum_{x \in \mathrm{G}} e^{2\pi \, \mathrm{i} \langle Nr, x \rangle} \right) = \frac{\gamma_{Nr}}{\sqrt{2}} 2.$$

Using these facts, we can write the empirical least-square loss as

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \sum_{m \in \mathbb{Z}^d_{\leq K}} \left| \lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m) \right|^2 \\
&= \frac{|\lambda_0 - \sqrt{2} \, \gamma_{Nr}|^2}{n} \sum_{i=1}^n \mathbb{1}[v_i = \gamma_0 \psi_0] \ + \frac{1}{n} \sum_{i=1}^n \mathbb{1}[v_i \neq \gamma_0 \psi_0] \left| \mathrm{DFT}(w_i^N)(-m) \right|^2 \\
&+ \frac{1}{n} \sum_{i=1}^n \sum_{m \in \mathbb{Z}^d_{\leq K} \setminus \{\mathbf{0}\}} \left| \lambda_m \, \mathrm{DFT}(v_i^N)(-m) - \mathrm{DFT}(w_i^N)(-m) \right|^2
\end{aligned}$$

Thus, the least squares estimator for λ_0 must be $\widehat{\lambda}_0 = \sqrt{2} \gamma_{Nr}$. That is,

$$\widehat{\lambda}_0^2 = 2\gamma_{Nr}^2 = \frac{2B^2}{(s+1)|Nr|_\infty^s} = \frac{2B^2}{(s+1)N^{2s}}.$$

Note that this choice of $\widehat{\lambda}_0$ is valid as $\widehat{\lambda}_0 \leq 1$. Thus, so far, we have shown that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\|\widehat{T}_K^N v - w\|_{L^2}^2\right] - \mathop{\mathbb{E}}_{(v,w)\sim\mu_\xi}\left[\left\|T_\xi v - w\|_{L^2}^2\right\|\right] \\
&\geq \frac{B^2}{3(s+1)}\left(\frac{\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0]}{|\mathcal{J}_M^N|} + \frac{2}{N^{2s}} + \frac{1}{(K+2)^{2s}}\right)
\end{aligned}$$

Our proof will be complete upon establishing that

$$\frac{1}{|\mathcal{J}_M^N|}\mathop{\mathbb{E}}_{\xi}\left[\mathop{\mathbb{E}}_{S_n\sim\mu_\xi}\left[\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0]\right]\right] \geq \frac{1}{8n}$$

for an appropriate choice of M. To that end, let $\mu_\xi^{\mathcal{V}}$ be the marginal of μ_ξ on $\mathcal{V}$ and $S_n^{\mathcal{V}} \in \mathcal{V}^n$ denote the restriction of samples $S_n \in (\mathcal{V}\times\mathcal{W})^n$ to its first arguments. Then, we can change the order of expectations to write

$$\mathop{\mathbb{E}}_{\xi}\left[\mathop{\mathbb{E}}_{S_n\sim\mu_\xi}\left[\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0]\right]\right] = \mathop{\mathbb{E}}_{S_n^{\mathcal{V}}\sim\mu_n^{\mathcal{V}}}\left[\mathop{\mathbb{E}}_{\xi}\left[\mathbb{1}[\widehat{\lambda}_r\xi_r \leq 0]\right]\right] \geq \frac{1}{2}\,\mathbb{P}[\gamma_r\psi_r \notin S_n^{\mathcal{V}}]$$

To understand why the final inequality holds, observe that when the event $\gamma_r\psi_r \notin S_n^{\mathcal{V}}$ occurs, the learner has no information about ξ_r. This implies that ξ_r and $\widehat{\lambda}_r$ are independent. Consequently, given that $\gamma_r\psi_r \notin S_n^{\mathcal{V}}$, the event $\widehat{\lambda}_r\xi_r \leq 0$ has a probability of at least $1/2$ since ξ_r is sampled uniformly from $\{-1,+1\}$.

Next, it remains to pick M such that

$$\frac{\mathbb{P}[\gamma_r\psi_r \notin S_n^{\mathcal{V}}]}{|\mathcal{J}_M^N|} \geq \frac{1}{4n}.$$

To get this, we choose $M = 2n$. It is easy to verify that $|\mathcal{J}_M^N| \geq n$ whenever $N > 1$. This is true because no more than half of integers in $\{1, 2, \ldots, 2n\}$ are divisible by N. Thus, we have

$$\mathbb{P}[\gamma_r\psi_r \notin S_n^{\mathcal{V}}] = \left(1 - \frac{1}{3|\mathcal{J}_M^N|}\right)^n \geq \left(1 - \frac{1}{3n}\right)^n \geq \frac{1}{2}$$

for any $n \geq 1$. Noting that $|\mathcal{J}_M^N| \leq 2n$ completes our proof. ■

E.6 Additional Experiments

In this section, we present additional experiments with y-axis on log scale to illustrate the decay rate. Each plot includes the fitted slope s and the corresponding smoothness parameter γ. The data generation, training, and evaluation setup follow Section 6.5. For each input

function v sampled at smoothness γ, the noise term ε is sampled at smoothness $1.5{\cdot}\gamma$. Figures E.1, E.2, and E.3 plot the statistical, truncation, and discretization errors, respectively. In each panel the reported value m is the slope of a line fitted to log(error) versus $\log n$, $\log K$, or $\log N$, as appropriate.

The statistical error decays faster than our predicted $1/\sqrt{n}$ rate, with slopes close to or better than -1.0, even better than our lower bound. This may be due to the Gaussian distribution used in experiments, rather than worst-case inputs. Additionally, the error appears to vary slightly with γ, suggesting that smoother inputs might allow sharper statistical rates, at least for well-behaved distributions.

For truncation error, the observed rates are generally faster than predicted for $\gamma = 1.5, 2$, and slightly slower (but close) for $\gamma = 3$. Recall that our theory predicts m to be 2γ.

Interestingly, the discretization error remains nearly constant with a slope around -1.1 across all values of γ, whereas the theory predicts the estimated slope to be γ. This discrepancy may stem from the experimental setup: our theory considers the test resolution $N_2 \to \infty$, while the experiments fix a finite $N_2 = 512$. A more detailed analysis under finite training and testing resolution could help clarify this behavior.

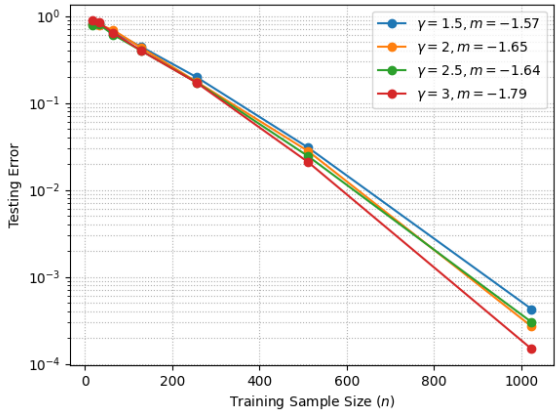


Figure E.1: Statistical error decay across sample sizes for different smoothness values γ.

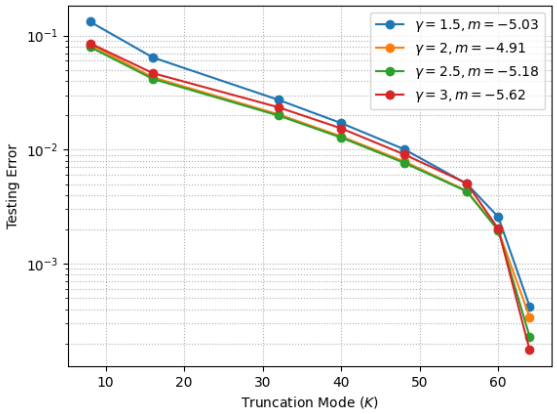


Figure E.2: Truncation error plotted against truncation mode for various values of γ.

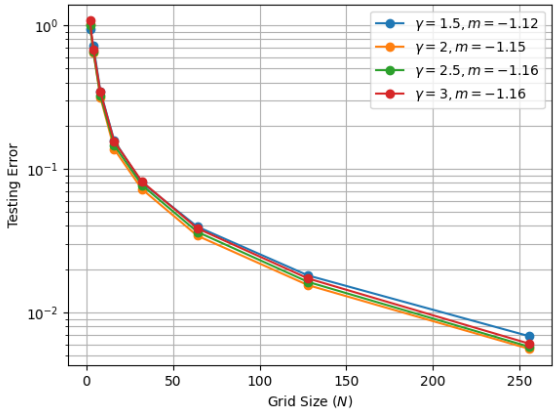


Figure E.3: Discretization error as a function of grid resolution for various smoothness levels γ.

APPENDIX F

On the Benefits of Active Data Collection in Operator Learning

F.1 Proof of Theorem 27

F.1.1 Specifying Data Collection Strategy and The Estimator

We first specify the estimator that achieves the claimed guarantee in Theorem 27. Let $\{\lambda_j, \varphi_j\}_{j=1}^{\infty}$ be the sequence of eigenpairs of K defined by solving the Feldholm integral equation

$$\int_{\mathcal{X}} K(y,x)\,\varphi_j(x)\,d\nu(x) = \lambda_j\,\varphi_j(y), \qquad y, x \in \mathcal{X}.$$

Given the Oracle call budget of n, the input functions that the learner selects are $\varphi_1, \varphi_2, \ldots, \varphi_n$ as source terms. For each $i \in [n]$, the learner makes an oracle call and obtain

$$w_i = \mathcal{O}(\varphi_i).$$

Consider the estimation rule

$$\underset{L \text{ is linear}}{\arg\min} \sum_{i=1}^{n} \|L\varphi_i - w_i\|_{L^2}^2 .$$

Solving this optimization problem boils down to solving the linear equation

$$\sum_{i=1}^{n} w_i \otimes \varphi_i = L \circ \left(\sum_{i=1}^{n} \varphi_i \otimes \varphi_i \right).$$

It is clear that this system is ill-posed and has infinitely many solutions. The family of solutions can be written as

$$L = \left(\sum_{i=1}^{n} w_i \otimes \varphi_i \right) \left(\sum_{i=1}^{n} \varphi_i \otimes \varphi_i \right)^{\dagger},$$

where † indicates the pseudoinverse. Each particular choice of pseudoinverse yields a distinct solution. Since φ_i's are orthonormal, a natural one is

$$\left(\sum_{i=1}^{n} \varphi_i \otimes \varphi_i\right)^{\dagger} = \sum_{i=1}^{n} \varphi_i \otimes \varphi_i.$$

This choice of pseudoinverse yields the estimator

$$\widehat{\mathrm{F}}_n := \sum_{i=1}^{n} w_i \otimes \varphi_i,$$

which will be our estimator interest.

F.1.2 Rewriting Risk using Karhunen–Loève Theorem

Next, we bound the risk of this estimator. Pick any $v \sim \mu$. Since v is defined using a centered and squared-integrable stochastic process with continuous covariance kernel K, the celebrated Karhunen–Loève Theorem [Hsing and Eubank, 2015, Theorem 7.3.5] states that

$$v(\cdot) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(\cdot),$$

where ξ_j's are random variables defined as

$$\xi_j := \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{X}} v_x(\omega)\, \varphi_j(x)\, d\nu(x).$$

Here, $v_x(\omega)$ is simply just $v(x)$, but we write the dependence on ω explicitly to highlight the fact that v is generated by stochastic process on the probability space $(\Omega, \Sigma, \mathbf{P})$.

It turns out that ξ_j's are uncorrelated random variables with mean 0 and variance 1. In particular, we have

$$\mathbb{E}[\xi_j] = 0 \quad \text{and} \quad \mathbb{E}[\xi_i\, \xi_j] = \mathbb{1}[i = j].$$

The precise convergence statement is

$$\lim_{m\to\infty} \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left|v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j \varphi_j(x)\right|^2\right] = 0. \tag{F.1}$$

We refer the reader to standard texts [Hsing and Eubank, 2015, Theorem 7.3.5] or [Lord et al., 2014, Theorem 7.52] for the full proof of Karhunen–Loève Theorem.

Fix $m \in \mathbb{N}$ such that $m > n$ and define Π_m to be a projection operator onto the first

m eigenfunctions of K. That is, for each v with Karhunen–Loève decomposition $v(\cdot) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(\cdot)$, we define

$$\Pi_m(v) := \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j \varphi_j(\cdot).$$

Since both $\widehat{\mathrm{F}}_n$ and F are linear operators, we can write

$$\begin{aligned}
&\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}(v)\right\|_{L^2}^2\right] \\
&= \mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right) + \Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}^2\right] \\
&\le \mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2}^2\right] \\
&+ 2\,\mathbb{E}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2} \left\|\Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}\right] \\
&+ \mathbb{E}\left[\left\|\Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}^2\right]
\end{aligned}$$

The inequality follows upon using triangle inequality and expanding the square. For the cross term, Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2} \left\|\Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}\right] \\
&\quad\le \sqrt{\mathbb{E}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2}^2\right]}\, \sqrt{\mathbb{E}\left[\left\|\Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}^2\right]}.
\end{aligned}$$

Thus, we can write

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}(v)\right\|_{L^2}^2\right] \le (\mathrm{I}) \;+\; 2\sqrt{(\mathrm{I})\,(\mathrm{II})} + (\mathrm{II}),$$

where we define

$$\begin{aligned}
(\mathrm{I}) &:= \mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2}^2\right] \\
(\mathrm{II}) &:= \mathbb{E}\left[\left\|\Big(\widehat{\mathrm{F}}_n - \mathrm{F}\Big)\Big(v - \Pi_m(v)\Big)\right\|_{L^2}^2\right].
\end{aligned}$$

Next, we will bound (I) and (II) separately.

F.1.3 Bounding (I).

Pick any $v \sim \mu$. Then, we know that there exists $\{\xi_j\}_{j\in\mathbb{N}}$ such that $v = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j \varphi_j$. So, $\Pi_m(v) = \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j \varphi_j$, which subsequently implies that

$$\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) = \widehat{\mathrm{F}}_n\left(\sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j\right) = \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \widehat{\mathrm{F}}_n(\varphi_j) \;= \sum_{j=1}^{n} \sqrt{\lambda_j}\, \xi_j\, w_j,$$

where the final equality uses the fact that $n < m$ and $\widehat{\mathrm{F}}_n(\varphi_j) = 0$ for all $j > n$. Defining $\delta_i := \mathcal{O}(\varphi_i) - \mathrm{F}(\varphi_i)$, we obtain $w_i = \mathrm{F}(\varphi_i) + \delta_i$. This allows us to write

$$\begin{aligned}
\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) &= \sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i\, (\mathrm{F}(\varphi_i) + \delta_i) \\
&= \mathrm{F}\left(\sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i \varphi_i\right) + \sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i \delta_i \\
&= \mathrm{F}\left(\Pi_m(v)\right) - \mathrm{F}\left(\sum_{j=n+1}^{m} \sqrt{\lambda_j}\, \xi_j \varphi_j\right) + \sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i \delta_i.
\end{aligned}$$

So, we can rewrite (I) as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n\Big(\Pi_m(v)\Big) - \mathrm{F}\left(\Pi_m(v)\right)\right\|_{L^2}^2\right] \\
&= \mathop{\mathbb{E}}_{\xi}\left[\left\|\sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i\, \delta_i - \mathrm{F}\left(\sum_{j=n+1}^{m} \sqrt{\lambda_j}\xi_j\, \varphi_j\right)\right\|_{L^2}^2\right] \\
&= \mathop{\mathbb{E}}_{\xi}\left[\left\|\sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i\, \delta_i\right\|_{L^2}^2\right] - 2\mathop{\mathbb{E}}_{\xi}\left[\left\langle \sum_{i=1}^{n} \sqrt{\lambda_i}\, \xi_i\, \delta_i, \mathrm{F}\left(\sum_{j=n+1}^{m} \sqrt{\lambda_j}\xi_j\, \varphi_j\right)\right\rangle\right] \\
&+ \mathop{\mathbb{E}}_{\xi}\left[\left\|\mathrm{F}\left(\sum_{j=n+1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j\right)\right\|_{L^2}^2\right].
\end{aligned}$$

The cross-term vanishes upon swapping sum and integral as ξ_i's are zero mean and uncorrelated. For the first term, note that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\xi}\left[\left\|\sum_{i=1}^{n}\sqrt{\lambda_i}\,\xi_i\,\delta_i\right\|_{L^2}^2\right] &= \mathop{\mathbb{E}}_{\xi}\left[\left\langle\sum_{i=1}^{n}\sqrt{\lambda_i}\,\xi_i\,\delta_i,\sum_{i=1}^{n}\sqrt{\lambda_i}\,\xi_i\,\delta_i\right\rangle_{L^2}\right] \\
&= \sum_{i=1}^{n}\lambda_i\,\mathbb{E}[\xi_i^2]\;\|\delta_i\|_{L^2} + 2\sum_{1\le i<j\le n}\sqrt{\lambda_i\lambda_j}\,\mathbb{E}[\xi_i\xi_j]\,\langle\delta_i,\delta_j\rangle_{L^2} \\
&= \sum_{i=1}^{n}\lambda_i\,\|\delta_i\|_{L^2}^2 + 0 \\
&\le \varepsilon^2\sum_{i=1}^{n}\lambda_i.
\end{aligned}$$

The final inequality uses the fact that $\mathcal{O}$ is ε-approximate for F. For the third term, similar arguments show that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\xi}\left[\left\|\mathrm{F}\left(\sum_{j=n+1}^{m}\sqrt{\lambda_j}\,\xi_j\,\varphi_j\right)\right\|_{L^2}^2\right] &\le \|\mathrm{F}\|_{\mathrm{op}}^2\;\mathop{\mathbb{E}}_{\xi}\left[\left\|\sum_{j=n+1}^{m}\sqrt{\lambda_j}\,\xi_j\,\varphi_j\right\|_{L^2}^2\right] \\
&= \|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=n+1}^{m}\lambda_j\,\|\varphi_j\|_{L^2}^2 \\
&= \|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=n+1}^{m}\lambda_j.
\end{aligned}$$

Thus, we have established that

$$(\mathrm{I}) \le \varepsilon^2\sum_{i=1}^{n}\lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=n+1}^{m}\lambda_j.$$

F.1.4 Bounding (II)

For any $v \sim \mu$, we have

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\left(\widehat{\mathrm{F}}_n - \mathrm{F}\right)\left(v - \Pi_m(v)\right)\right\|_{L^2}^2\right] \le \left\|\widehat{\mathrm{F}}_n - \mathrm{F}\right\|_{\mathrm{op}}^2\,\mathbb{E}\left[\|v - \Pi_m(v)\|_{L^2}^2\right]$$

Let $v = \sum_{j\geq 1} \sqrt{\lambda_j}\, \xi_j \varphi_j$. Then,

$$\begin{aligned}
\mathbb{E}\left[\|v - \Pi_m(v)\|_{L^2}^2\right] &= \mathbb{E}\left[\left\| v - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j \right\|_{L^2}^2\right] \\
&= \mathbb{E}\left[\int_{\mathcal{X}} \left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2 d\nu(x)\right] \\
&= \int_{\mathcal{X}} \mathbb{E}\left[\left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2\right] d\nu(x) \\
&\leq \nu(\mathcal{X}) \cdot \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2\right].
\end{aligned}$$

The third equality uses joint measurability, finiteness of ν, and Tonelli's theorem to exchange the order or integration. Therefore, we have established that

$$\text{(II)} \leq \left\|\widehat{\mathrm{F}}_n - \mathrm{F}\right\|_{\mathrm{op}}^2 \cdot \nu(\mathcal{X}) \cdot \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2\right].$$

F.1.5 Combining (I) and (II)

For each $m > n$, we have established

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}(v)\right\|_{L^2}^2\right] \leq \text{(I)} \;+\; 2\sqrt{\text{(I)}\,\text{(II)}} + \text{(II)},$$

where

$$\begin{aligned}
\text{(I)} &\leq \varepsilon^2 \sum_{i=1}^{n} \lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j=n+1}^{m} \lambda_j \\
\text{(II)} &\leq \left\|\widehat{\mathrm{F}}_n - \mathrm{F}\right\|_{\mathrm{op}}^2 \cdot \nu(\mathcal{X}) \cdot \sup_{x\in\mathcal{X}} \mathbb{E}\left[\left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2\right].
\end{aligned}$$

It now remains to choose m such that the upperbound is minimized. To that end, we will take $m \to \infty$. Since $w_1, \ldots, w_n \in L^2(\mathcal{X})$, we must have $\left\|\widehat{\mathrm{F}}_n\right\|_{\mathrm{op}} < \infty$. As F is also a bounded linear operator, for any $n \in \mathbb{N}$, we must have

$$\left\|\widehat{\mathrm{F}}_n - \mathrm{F}\right\|_{\mathrm{op}}^2 < \infty.$$

Importantly, the norm of $\widehat{\mathrm{F}}_n - \mathrm{F}$ may grow with n, but is independent of m and does not grow as $m \to \infty$. Moreover, as ν is a finite measure, we must have $\nu(\mathcal{X}) < \infty$. Therefore, Karhunen–Loève Theorem [Hsing and Eubank, 2015, Theorem 7.3.5] (also see Equation (F.1)) implies that

$$\text{(II)} \leq \left\|\widehat{\mathrm{F}}_n - \mathrm{F}\right\|_{\mathrm{op}}^2 \cdot \nu(\mathcal{X}) \cdot \sup_{x \in \mathcal{X}} \mathbb{E}\left[\left(v(x) - \sum_{j=1}^{m} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)^2\right] \xrightarrow{m \to \infty} 0.$$

On the other hand,

$$\text{(I)} \xrightarrow{m \to \infty} \varepsilon^2 \sum_{i=1}^{n} \lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j=n+1}^{\infty} \lambda_j.$$

Overall, we have shown that

$$\mathop{\mathbb{E}}_{v \sim \mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}(v)\right\|_{L^2}^2\right] \leq \varepsilon^2 \sum_{i=1}^{n} \lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j=n+1}^{\infty} \lambda_j.$$

This completes our proof of Theorem 27.

F.2 Examples of Covariance Kernels

In this section, we build upon and present a more rigorous analysis of the material discussed in Section 7.3.3 of the main text.

F.2.1 Fractional Inverse of Shifted Laplacian

Li et al. [2021] and Kovachki et al. [2023] generated input functions from $\mathrm{GP}(0, \alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma})$ for some constants $\alpha, \beta, \gamma > 0$. Here, ∇^2 is the Laplacian operator defined as

$$\nabla^2 v = \sum_{j=1}^{d} \frac{\partial^2 v}{\partial x_j^2}.$$

In this section, we will consider $\mathcal{X}$ to be a d-dimensional periodic torus $\mathbb{T}^d$ and the base measure ν is Lebesgue. We identify $\mathbb{T}^d$ by $[0,1]^d$ with periodic boundary conditions.

Let us define a function $\varphi_m : \mathbb{T}^d \to \mathbb{C}$ as $\varphi_m(x) = e^{2\pi\,\mathrm{i}\, m \cdot x}$ for every $m \in \mathbb{Z}^d$. Recall that φ_m is the eigenfunction of ∇^2 with eigenvalue $-4\pi^2|m|_2^2$. In particular,

$$\nabla^2 e^{2\pi\,\mathrm{i}\, m \cdot x} = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2}\, e^{2\pi\,\mathrm{i}\, m \cdot x} = \sum_{j=1}^{d} (2\pi\,\mathrm{i}\, m_j)^2 e^{2\pi\,\mathrm{i}\, m \cdot x} = -4\pi^2 |m|_2^2\, e^{2\pi\,\mathrm{i}\, m \cdot x}.$$

Since $\{\varphi_m : m \in \mathbb{Z}^d\}$ forms a complete orthonormal system in $L^2(\mathbb{T}^d)$, there are no other eigenfunctions of ∇^2. A simple algebra shows that φ_m's are also he eigenfunctions of shifted Laplacian $-\nabla^2 + \beta\mathbf{I}$ with eigenvalues being $(\beta + 4\pi^2|m|_2^2)$. Finally, the spectral mapping theorem implies that $\{(\lambda_m, \varphi_m) : m \in \mathbb{Z}^d\}$ is the sequence of eigenpairs of $\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$, where the eigenvalues are

$$\lambda_m = \alpha\left(\beta + 4\pi^2|m|_2^2\right)^{-\gamma}.$$

Next, we need to show that these eigenvalues are summable to use Theorem 27. Note that

$$\sum_{m\in\mathbb{Z}^d} \lambda_m = \sum_{m\in\mathbb{Z}^d} \alpha\left(\beta + 4\pi^2|m|_2^2\right)^{-\gamma} \leq \alpha\beta^{-\gamma} + \frac{\alpha}{(2\pi)^{2\gamma}} \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|_\infty^{-2\gamma}.$$

It is easy to see that

$$\sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|_\infty^{-2\gamma} \leq \sum_{j=1}^{\infty} j^{-2\gamma}\,(2j+1)^d \leq 3^d \sum_{j=1}^{\infty} j^{-2\gamma+d} < \infty$$

as long as $2\gamma > d$. The first inequality holds because $|\{m \in \mathbb{Z}^d\setminus\{0\} : |m|_\infty = j\}| \leq 2(2j+1)^{d-1}$. This is true because at least one of the entries has to be $\pm j$ and other $d-1$ entries could be anything in $\{0, \pm 1, \ldots, \pm j\}$. So we have $\sum_{m\in\mathbb{Z}^d} |\lambda_m| < \infty$, implying that the operator $\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$ is in trace class as long as $2\gamma > d$.

Finally, it is easy to see that the operator $\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma}$ is integral operator associated with the kernel

$$K(y,x) = \sum_{m\in\mathbb{Z}^d} \lambda_m\, \varphi_m(y)\, \overline{\varphi_m(x)}.$$

Upon writing the Fourier series of $v \in L^2(\mathbb{T})$, it is obvious that $\Big(\alpha(-\nabla^2 + \beta\mathbf{I})^{-\gamma} v\Big)(y) = \int_{\mathcal{X}} K(y,x)v(x)\,dx$ for all $y \in \mathcal{X}$. As $\sum_{m\in\mathbb{Z}} |\lambda_m| < \infty$, the convergence is absolute and uniform. Since K is a uniform limit of the sum of continuous functions, K must also be continuous. Moreover, as $\lambda_m = \lambda_{-m}$, the kernel K must be real-valued. In particular, we have

$$\begin{aligned}
K(y,x) &= \sum_{m\in\mathbb{Z}^d} \lambda_m\, \varphi_m(y)\, \overline{\varphi_m(x)} \\
&= \sum_{m\in\mathbb{Z}^d} \frac{\lambda_m}{2}\left(\varphi_m(y)\,\overline{\varphi_m(x)} + \varphi_{-m}(y)\,\overline{\varphi_{-m}(x)}\right) \\
&= \sum_{m\in\mathbb{Z}^d} \lambda_m \cos\Big(2\pi m\cdot(y-x)\Big) \\
&= \sum_{m\in\mathbb{Z}^d} \lambda_m \Big(\cos(2\pi m\cdot y)\,\cos(2\pi m\cdot x) + \sin(2\pi m\cdot y)\,\sin(2\pi m\cdot x)\Big)
\end{aligned}$$

This is a generalization of the cosine covariance kernel often considered in computational

PDE literature (see [Lord et al., 2014, Example 5.20]). Since $\cos(\theta) = \cos(-\theta)$, it is obvious that K is symmetric. As K is a continuous and real-valued covariance kernel defined on a bounded domain $\mathbb{T}^d$, we can use Theorem 27 for such K. In principle, we could use the Fourier modes φ_m's as source terms to define the estimator discussed in Section 7.3.1. However, the proof of Theorem 27 assumes that the eigenfunctions of the kernel are real-valued. So, we will first show that we can write the eigenfunctions of K solely using sine and cosine functions without having to use complex exponentials. This allows us to use these sine and cosine functions to define the estimator discussed in Section 7.3.1, and invoke results of Theorem 27.

F.2.1.1 Case $d = 1$

When $d = 1$, it is easy to see that $\{1\} \cup \{\sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx) \ : j \in \mathbb{N}\}$ are the eigenfunctions of K. Writing the expansion of K and using Fubini's to switch the sum and the integral, we get

$$\int_{\mathbb{T}} K(y,x)\sqrt{2}\cos(2\pi jx)\,dx = \lambda_j\sqrt{2}\,\cos(2\pi jy)\frac{1}{2} + \lambda_{-j}\sqrt{2}\,\cos(-2\pi jy)\frac{1}{2} = \lambda_j\sqrt{2}\cos(2\pi jy).$$

Note that the first equality holds because $\cos(2\pi jx)$ is orthogonal to all other cosine and sine functions except for $\cos(2\pi jx)$ and $\cos(-2\pi jx)$. The final equality holds because $\lambda_j = \lambda_{-j}$ and $\cos(\theta) = \cos(-\theta)$. A similar calculation shows that

$$\int_{\mathbb{T}} K(y,x)\sqrt{2}\sin(2\pi jx)\,dx = \lambda_j\sqrt{2}\,\sin(2\pi jy)\frac{1}{2} + \lambda_{-j}\sqrt{2}\,\sin(-2\pi jy)\frac{-1}{2} = \lambda_j\sqrt{2}\sin(2\pi jy).$$

Finally, we have $\int_{\mathbb{T}} K(y,x)\,1\,dx = \lambda_0 1$. Thus, λ_j for $j \in \mathbb{N}$ are the eigenvalues for sine/cosine functions and λ_0 for 1. Since $\{1\} \cup \{\sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx) \ : j \in \mathbb{N}\}$ forms a complete orthonormal system of $L^2(\mathbb{T}, \mathbb{R})$, there cannot be any more eigenfunctions of K. Next, we will plug in the values of λ_j's in Theorem 27 to get the precise rates.

Pick an odd $n \in \mathbb{N}$ and suppose the n input terms used to construct the estimator in Section 7.3.1 are $\{1\} \cup \{\sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx) \ : j \leq (n-1)/2\}$. Then, the upperbound

is

$$
\begin{aligned}
&\varepsilon^2\left(\lambda_0+\sum_{j=1}^{(n-1)/2}2\lambda_j\right)+\|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=(n+1)/2}^{\infty}2\lambda_j\\
&\leq\varepsilon^2\alpha\beta^{-\gamma}+\varepsilon^2 2\sum_{j=1}^{(n-1)/2}\alpha\left(\beta+4\pi^2j^2\right)^{-\gamma}+2\|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=(n+1)/2}^{\infty}\alpha\left(\beta+4\pi^2j^2\right)^{-\gamma}\\
&\leq\varepsilon^2\alpha\beta^{-\gamma}+\varepsilon^2\alpha\frac{2}{(4\pi^2)^\gamma}\sum_{j=1}^{(n-1)/2}\frac{1}{j^{2\gamma}}+\frac{2\alpha}{(4\pi^2)^\gamma}\|\mathrm{F}\|_{\mathrm{op}}^2\sum_{j=(n+1)/2}^{\infty}\frac{1}{j^{2\gamma}}\\
&\leq\varepsilon^2\alpha\beta^{-\gamma}+\varepsilon^2\alpha\frac{2}{(4\pi^2)^\gamma}+\varepsilon^2\alpha\frac{2}{(4\pi^2)^\gamma}\int_1^{(n-1)/2}t^{-2\gamma}\,dt+\frac{2\alpha}{(4\pi^2)^\gamma}\|\mathrm{F}\|_{\mathrm{op}}^2\int_{(n-1)/2}^{\infty}t^{-2\gamma}\,dt\\
&\leq\varepsilon^2\alpha\beta^{-\gamma}+\varepsilon^2\alpha\frac{2}{(4\pi^2)^\gamma}+\frac{2}{(4\pi^2)^\gamma}\frac{\varepsilon^2\alpha}{2\gamma-1}+\frac{2\alpha}{(4\pi^2)^\gamma}\|\mathrm{F}\|_{\mathrm{op}}^2\frac{1}{2\gamma-1}\frac{2^{2\gamma-1}}{(n-1)^{2\gamma-1}},\quad\forall\gamma>\frac{1}{2}.
\end{aligned}
$$

Since $2\cdot 2^{2\gamma-1}\leq(4\pi^2)^\gamma$ and $2^{2\gamma-1}\leq(4\pi^2)^\gamma$, the overall error is at most

$$
\varepsilon^2\left(\alpha\beta^{-\gamma}+\alpha+\frac{\alpha}{2\gamma-1}\right)+\frac{\alpha\,\|\mathrm{F}\|_{\mathrm{op}}^2}{2\gamma-1}\frac{1}{(n-1)^{2\gamma-1}}\qquad\text{for all }\gamma>\frac{1}{2}.
$$

Since $\gamma>1/2$, the reducible error goes to 0 as $n\to\infty$. As an example, Li et al. [2021] uses $\alpha=625$, $\beta=25$ and $\gamma=2$ in their experiment for $1d$-Burger's equation. In this case, we get the convergence rate of n^{-3} for the reducible error. Note that this rate of *cubic order* is faster than the usual passive statistical rate of $1/n$. In fact, for any value τ, one can take $\gamma=(\tau+1)/2$ to get the rate of $n^{-\tau}$. Thus, every polynomial rate is possible depending on the choice of γ.

F.2.1.2 Case $d>1$

Recall that $\{1\}\cup\{\sqrt{2}\cos(2\pi jx),\sqrt{2}\sin(2\pi jx)\ :j\in\mathbb{N}\}$ are the eigenvalues of K for $d=1$ with eigenvalues $\lambda_j:=\alpha\left(\beta+4\pi^2j^2\right)^{-\gamma}$. Define a set of functions

$$
\mathcal{E}=\prod_{i=1}^{d}\{1\}\cup\{\sqrt{2}\cos(2\pi jx_i),\sqrt{2}\sin(2\pi jx_i)\ :j\in\mathbb{N}\}.
$$

For each element $e\in\mathcal{E}$, there exists a tuple $j:=(j_1,\ldots,j_d)\in\mathbb{N}_0^d$ such that

$$
e(x)=\psi_{j_1}(x_1)\ldots\psi_{j_{d-1}}(x_{d-1})\cdot\psi_{j_d}(x_d),
$$

where $\psi_{j_i}(x_i) \in \{\sqrt{2}\cos(2\pi j_i x_i), \sqrt{2}\sin(2\pi j_i x_i)\}$ for $j_i > 0$ and $\sqrt{2}$ for $j_i = 0$. Let us denote the collection of all such functions by $\mathcal{E}_j$. Then, we have $\mathcal{E} = \cup_{j \in \mathbb{N}_0^d} \mathcal{E}_j$. We prove the following result on the eigenpairs of K.

Proposition 7. *For each $j \in \mathbb{N}_0^d$, define $\lambda_j = \alpha\left(\beta + 4\pi^2 |j|_2^2\right)^{-\gamma}$. Then,*

$$\bigcup_{j \in \mathbb{N}_0^d} \bigcup_{e \in \mathcal{E}_j} \{(\lambda_j, e)\}$$

is the set of eigenpairs of K on $\mathbb{T}^d$.

We defer the full proof of Proposition 7 to the end of this subsection. First, we use Proposition 7 and Theorem 27 to get the precise rate for kernel K. Pick r such that the source terms used to construct the estimator defined in Section 7.3.1 are

$$\bigcup_{j \in \mathbb{N}_0^d : |j|_\infty \leq r} \mathcal{E}_j$$

Note that $|\mathcal{E}_0| = 1$ and $|\mathcal{E}_j| \leq 2^d$ for all $|j|_\infty > 0$. Thus, there are $n \leq (r+1)^d 2^d$ source terms. Then, the upperbound is

$$\begin{aligned}
&\leq \varepsilon^2 \lambda_0 + \sum_{0 < |j|_\infty \leq r} 2^d \lambda_j + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{|j|_\infty > r} 2^d \lambda_j \\
&= \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 2^d \sum_{0 < |j|_\infty \leq r} \alpha \left(\beta + 4\pi^2 |j|_2^2\right)^{-\gamma} + 2^d \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{|j|_\infty > r} \alpha \left(\beta + 4\pi^2 |j|_2^2\right)^{-\gamma} \\
&= \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 2^d \sum_{0 < |j|_\infty \leq r} \alpha \left(\beta + 4\pi^2 |j|_\infty^2\right)^{-\gamma} + 2^d \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{|j|_\infty > r} \alpha \left(\beta + 4\pi^2 |j|_\infty^2\right)^{-\gamma} \\
&\leq \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 2^d \sum_{k=1}^{r} \alpha \left(\beta + 4\pi^2 k^2\right)^{-\gamma} (k+1)^{d-1} + 2^d \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{k > r} \alpha \left(\beta + 4\pi^2 k^2\right)^{-\gamma} (k+1)^{d-1} \\
&\leq \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 2^d \frac{\alpha 2^d}{(4\pi^2)^\gamma} \sum_{k=1}^{r} k^{d-1-2\gamma} + 2^d \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{\alpha\, 2^d}{(4\pi^2)^\gamma} \sum_{k > r} k^{d-1-2\gamma} \\
&\leq \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 \frac{\alpha 2^{2d}}{(4\pi^2)^\gamma} + \varepsilon^2 \frac{\alpha 2^{2d}}{(4\pi^2)^\gamma} \int_1^r t^{d-1-2\gamma}\, dt + \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{\alpha\, 2^{2d}}{(4\pi^2)^\gamma} \int_r^\infty t^{d-1-2\gamma}\, dt \\
&\leq \varepsilon^2 \alpha \beta^{-\gamma} + \varepsilon^2 \frac{\alpha 2^{2d}}{(4\pi^2)^\gamma} + \varepsilon^2 \frac{\alpha 2^{2d}}{(4\pi^2)^\gamma} \frac{1}{2\gamma - d} + \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{\alpha\, 2^{2d}}{(4\pi^2)^\gamma} \frac{1}{2\gamma - d} \frac{1}{r^{2\gamma - d}},
\end{aligned}$$

for all $2\gamma > d$. Recall that $n \leq (2r+2)^d$. So, we have $n^{1/d}/2 - 1 \leq r$. For $n^{1/d} \geq 4$, we have $r \geq n^{1/d}/4$. Thus,

$$\frac{1}{r^{2\gamma - d}} \leq \frac{4^{2\gamma - d}}{n^{\frac{2\gamma}{d} - 1}}$$

Note that $(4\pi^2)^\gamma = (2\pi)^{2\gamma} \geq 2^{2d} 4^{2\gamma - d}$. Moreover, as $2\gamma > d$, we also have $(4\pi^2)^\gamma \geq 2^{2d}$.

Therefore, our upper bound is at most

$$\varepsilon^2 \left(\alpha\beta^{-\gamma} + \alpha + \frac{\alpha}{2\gamma - d} \right) + \frac{\alpha \left\| \mathrm{F} \right\|_{\mathrm{op}}^2}{2\gamma - d} \frac{1}{n^{\frac{2\gamma}{d}-1}}.$$

Since $2\gamma/d - 1 > 0$, the reducible error above goes to 0 when $n \to \infty$. Again, as an example, Li et al. [2021] uses $\alpha = 7^{3/2}$, $\beta = 49$ and $\gamma = 2.5$ in their experiment for $2d$-Navier Stokes. In this case, $2\gamma/d = 2.5$, yielding the convergence rate of $n^{-1.5}$ for the reducible error. Note that this rate is faster than the usual passive statistical rate of $1/n$. However, as usual, for any value τ, one can take $\gamma = d(\tau + 1)/2$ to get the rate of $n^{-\tau}$. Thus, every polynomial rate is possible depending on the choice of γ.

We now end this section by providing the proof of Theorem 7.

Proof of Proposition 7. Since $\cup_{j \in \mathbb{N}_0^d} \cup_{e \in \mathcal{E}_j} \{e\}$ forms an orthonormal basis of $L^2(\mathbb{T}^d, \mathbb{R})$, there cannot be anymore eigenfunctions of K. Thus, it suffices to show that (λ_j, e) is an eigenpair for any $e \in \mathcal{E}_j$ and $j \in \mathbb{N}_0^d$. To prove this, we will establish that

$$\int_{\mathbb{T}^d} \sum_{m \in \mathbb{Z}^d} \mathbb{1}\{|m_i| = j_i \ \forall i \in [d]\} \cos\Big(2\pi m \cdot (y - x)\Big) e_j(x)\, dx = e_j(y), \tag{F.2}$$

where e_j is an arbitrary element of $\mathcal{E}_j$. Recall that

$$\int_{\mathbb{T}^d} \cos\Big(2\pi m \cdot (y - x)\Big) e_j(x)\, dx = 0 \text{ if } \exists i \text{ such that } |m_i| \neq j_i.$$

This is true because if $\exists i$ such that $|m_i| \neq j_i$, then we can write $\cos\Big(2\pi m \cdot (y - x)\Big) = \cos\Big(2\pi \sum_{\ell \neq i} m_\ell (y_\ell - x_\ell)\Big) \cos(2\pi m_i (y_i - x_i)) - \sin\Big(2\pi \sum_{\ell \neq i} m_\ell (y_\ell - x_\ell)\Big) \sin(2\pi m_i (y_i - x_i))$. Moreover, $e_j(x) = \psi_{j_1}(x_1) \ldots \psi_{j_{d-1}}(x_{d-1}) \cdot \psi_{j_d}(x_d)$, where ψ_{j_ℓ}'s are either sine, cosine, or a constant function. Our claim follows upon noting that $\psi_{j_i}(x_i)$ is orthogonal to both $\sin(2\pi m_i(y_i - x_i))$ and $\cos(2\pi m_i(y_i - x_i))$.

Thus, Equation (F.2) together with the fact that $\lambda_m = \lambda_j$ for all $m \in \{k \in \mathbb{Z}^d : |k_i| = j_i \quad \forall i \in [d]\}$ implies that (λ_j, e_j) is the eigenpair of K. As $j \in \mathbb{N}_0^d$ and $e_j \in \mathcal{E}_j$ are arbitrary, this completes our proof.

Now, it remains to prove Equation (F.2). We will proceed by induction on d. For the base case, take $d = 1$. If $j = 0$, $e_j = 1$ and our claim follows trivially. Suppose $j \neq 0$. Since $\cos(\theta) = \cos(-\theta)$, we have

$$\begin{aligned} \sum_{m \in \mathbb{Z}} \mathbb{1}\{|m| = j\} \cos\Big(2\pi m(y - x)\Big) &= 2\cos(2\pi j(y - x)) \\ &= 2\cos(2\pi jy)\ \cos(2\pi jx) + 2\ \sin(2\pi jy) \sin(2\pi jx). \end{aligned}$$

If $e_j(x) = \sqrt{2}\cos(2\pi jx)$, then

$$\int_{\mathbb{T}} \left(2\cos(2\pi jy)\,\cos(2\pi jx) + 2\,\sin(2\pi jy)\sin(2\pi jx)\right)\sqrt{2}\,\cos(2\pi jx)\,dx = \sqrt{2}\cos(2\pi jy).$$

If $e_j(x) = \sqrt{2}\sin(2\pi jx)$, a similar calculation shows that

$$\int_{\mathbb{T}} \left(2\cos(2\pi jy)\,\cos(2\pi jx) + 2\,\sin(2\pi jy)\sin(2\pi jx)\right)\sqrt{2}\sin(2\pi jx)\,dx = \sqrt{2}\sin(2\pi jy).$$

This completes our proof of the base case.

Suppose (F.2) is true for $d-1$. We will now prove it for d. Note that

$$\begin{aligned}
&\cos\Big(2\pi m\cdot(y-x)\Big) = \cos\left(2\pi\sum_{i=1}^{d} m_i(y_i-x_i)\right)\\
&= \cos\left(2\pi\sum_{i=1}^{d-1} m_i(y_i-x_i)\right)\,\cos\left(2\pi m_d(y_d-x_d)\right)\\
&- \sin\left(2\pi\sum_{i=1}^{d-1} m_i(y_i-x_i)\right)\,\sin\left(2\pi m_d(y_d-x_d)\right).
\end{aligned}$$

First, observe that when summed over all $m\in\mathbb{Z}^d$ such that $|m_i| = j_i$ for all $i\in[d]$, the sine term vanishes. That is,

$$\begin{aligned}
&\sum_{m\in\mathbb{Z}^d} \mathbb{1}\Big\{|m_i| = j_i\ \forall i\in[d]\Big\}\left(\sin\left(2\pi\sum_{i=1}^{d-1} m_i(y_i-x_i)\right)\,\sin\left(2\pi m_d(y_d-x_d)\right)\right)\\
&= \left(\sum_{m\in\mathbb{Z}^{d-1}} \mathbb{1}\Big\{|m_i| = j_i\Big\}\,\sin\left(2\pi\sum_{i=1}^{d-1} m_i(y_i-x_i)\right)\right)\\
&\times\left(\sum_{m_d\in\mathbb{Z}} \mathbb{1}\{|m_d| = j_d\}\,\sin\left(2\pi m_d(y_d-x_d)\right)\right)\\
&= 0.
\end{aligned}$$

The final step follows here because the term in the second parenthesis above is always 0. There are two cases to consider. If $j_d = 0$, the summand only has one term and our claim holds as $\sin(0) = 0$. On the other hand, if $j_d\neq 0$, then we are have $\sin(\theta)+\sin(-\theta) = 0$.

Therefore, we obtain

$$\begin{aligned}
&\sum_{m\in\mathbb{Z}^d} \mathbb{1}\{|m_i| = j_i \ \forall i \in [d]\} \cos\Big(2\pi m\cdot(y-x)\Big) \\
&= \left(\sum_{m\in\mathbb{Z}^{d-1}} \mathbb{1}\{|m_i| = j_i \ \forall i \in [d-1]\} \ \cos\left(2\pi \sum_{i=1}^{d-1} m_i(y_i - x_i)\right)\right) \\
&\times \left(\sum_{m_d\in\mathbb{Z}} \mathbb{1}\{|m_d| = j_d\} \ \cos\left(2\pi m_d(y_d - x_d)\right)\right)
\end{aligned}$$

A similar factorization can be done for e_j to write

$$e_j(x) = \psi_{j_1}(x_1)\dots\psi_{j_d}(x_d),$$

where ψ_{j_i}'s are either sine, cosine, or a constant function.

However, ψ_{j_d} is some e_{j_d} defined on $\mathbb{T}$. Thus, using the base case, we have

$$\int_{\mathbb{T}} \sum_{m_d\in\mathbb{Z}} \mathbb{1}\{|m_d| = j_d\} \ \cos\left(2\pi m_d(y_d - x_d)\right) \ \psi_{j_d}(x_d)\, dx_d = \psi_{j_d}(y_d).$$

Similarly, using the induction hypothesis, we have

$$\begin{aligned}
&\int_{\mathbb{T}^{d-1}} \left(\sum_{m\in\mathbb{Z}^{d-1}} \mathbb{1}\{|m_i| = j_i \, \forall i < d\} \cos\left(2\pi \sum_{i=1}^{d-1} m_i(y_i - x_i)\right)\right) \prod_{i=1}^{d-1} \psi_{j_i}(x_i) \ d(x_1,\dots,x_{d-1}) \\
&= \prod_{i=1}^{d-1} \psi_{j_i}(y_i).
\end{aligned}$$

Combining everything, we obtain

$$\int_{\mathbb{T}^d} \sum_{m\in\mathbb{Z}^d} \mathbb{1}\{|m_i| = j_i \ \forall i \in [d]\} \cos\Big(2\pi m\cdot(y-x)\Big) \prod_{i=1}^{d} \psi_{j_i}(x_i) \ dx = \prod_{i=1}^{d} \psi_{j_i}(y_i).$$

The final step requires using the factorization of cosine and writing integral over $\mathbb{T}^d$ as product of integral over $\mathbb{T}^{d-1}$ and $\mathbb{T}$. This completes our induction step, and thus the proof. ■

F.2.2 RBF Kernel on $\mathbb{R}$.

Let K be the RBF kernel on $\mathbb{R}$. That is, $K(x,y) = \exp\left(-\frac{1}{2\ell^2}|x-y|^2\right)$ for all $x, y \in \mathbb{R}$. For now, let ν is a Gaussian measure with mean 0 and variance σ^2 on $\mathbb{R}$. Then, it is known

[Williams and Rasmussen, 2006, Section 4.3.1] that $K(x,y) = \sum_{j=0}^{\infty} \lambda_j \, \varphi_j(x) \, \varphi_j(y)$, where

$$\lambda_j := \sqrt{\frac{2a}{a+b+c}} \left(\frac{b}{a+b+c}\right)^j$$
$$\varphi_j(x) := \exp(-(c-a)x^2) \, H_j(\sqrt{2c}x).$$

Here, $a = (4\sigma^2)^{-1}$, $b = (2\ell^2)^{-1}$, $c = \sqrt{a^2 + 2ab}$, and $H_j(\cdot)$ is the Hermite polynomial of order j defined as

$$H_j(x) = (-1)^j \exp(x^2) \, \frac{d^j}{dx^j} \exp(-x^2).$$

Note that this is the eigenpairs of $K(y,x)$ over the entire $\mathbb{R}$, whereas we need eigenpairs over some compact domain $\mathcal{X} \subseteq \mathbb{R}$. The eigenpairs of $K(y,x)$ are generally not available in closed form for arbitrary $\mathcal{X}$. However, the variance of the Gaussian measure σ^2 can be tuned appropriately to localize the domain $\mathbb{R}$ to appropriate $\mathcal{X}$ of interest. For example, let $\mathcal{X} = [-1,1]$. Then,

$$\int_{-1}^{1} K(y,x)\varphi_j(x) \, d\nu(x) = \int_{\mathbb{R}} K(y,x) \, \varphi_j(x) \, d\nu(x) - \int_{|x|>1} K(y,x) \, \varphi_j(x) \, d\nu(x).$$

Since $\int_{\mathbb{R}} K(y,x) \, \varphi_j(x) \, d\nu(x) = \lambda_j \varphi_j(y)$, we have

$$\begin{aligned}
\left| \int_{-1}^{1} K(y,x)\varphi_j(x) \, d\nu(x) - \lambda_j \varphi_j(y) \right| &\leq \int_{|x|>1} |K(y,x)| \, |\varphi_j(x)| \, d\nu(x) \\
&\leq \sqrt{\int_{|x|>1} |K(y,x)|^2 \, d\nu(x)} \sqrt{\int_{|x|>1} |\varphi_j(x)|^2 \, d\nu(x)} \\
&\leq \sqrt{\int_{|x|>1} \exp\left(-\frac{|x-y|^2}{\ell^2}\right) \, d\nu(x)},
\end{aligned}$$

where the second term is upper bounded by 1 as φ_j^2 integrates to 1 over the whole domain $\mathbb{R}$. Note that $\exp\left(-\frac{|x-y|^2}{\ell^2}\right) \leq 1$ and $\sqrt{\nu([-1,1]^c)} \leq 3.9 \times 10^{-12}$ when $\sigma = 0.1$. So, σ can be appropriately tuned such that $(\lambda_j, \varphi_j)_{j \geq 1}$ is a good approximation of the eigenpair of K for our domain $\mathcal{X}$ of interest. Next, we use these eigenvalues to study how the upper bound in Theorem 27 decays as $n \to \infty$.

Let $\gamma := b/(a+b+c)$. It is clear that $\gamma \in (0,1)$. Since $c = \sqrt{a^2 + 2ab} \geq a$, we also have $\sqrt{\frac{2a}{a+b+c}} \leq 1$. Thus, we obtain $\lambda_j \leq \gamma^j$. Plugging this estimate in the upperbound of

Theorem 27, we obtain

$$\begin{aligned}\varepsilon^2 \sum_{i=0}^{n-1} \lambda_i + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{i=n}^{\infty} \lambda_i &\leq \varepsilon^2 \sum_{i=0}^{n-1} \gamma^j \; + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{i=n}^{\infty} \gamma^j \\ &= \frac{1-\gamma^n}{1-\gamma}\varepsilon^2 + \|\mathrm{F}\|_{\mathrm{op}}^2 \; \frac{\gamma^n}{1-\gamma} \\ &\leq \frac{1}{(1-\gamma)} \left(\varepsilon^2 + \|\mathrm{F}\|_{\mathrm{op}}^2 \, \gamma^n\right).\end{aligned}$$

Therefore, the reducible error vanishes exponentially fast as $n \to \infty$.

F.2.3 RBF Kernel on $\mathbb{R}^d$

Let $K(y,x) = \exp(-|x-y|_2^2/(2\ell^2))$, where $x, y \in \mathbb{R}^d$. Then, it is clear that

$$K(y,x) = \prod_{i=1}^{d} \exp(-|x_i - y_i|^2/(2\ell^2)) =: \prod_{i=1}^{d} K_i(y_i, x_i).$$

If $(\lambda_{ij}, \varphi_{ij})_{j\in\mathbb{N}}$ are the eigenpairs of K_i under the weighted measure standard Gaussian measure on $\mathbb{R}$, then

$$\left\{ \left(\prod_{i=1}^{d} \lambda_{ij_i} \; , \; \prod_{i=1}^{d} \varphi_{ij_i} \right) \,\middle|\, (j_1, j_2, \ldots, j_d) \in \mathbb{N}_0^d \right\}$$

are the eigenpairs of K when ν is multivariate Gaussian with mean 0 and covariance $\sigma^2\mathbf{I}$. This follows immediately upon noting that

$$\int_{\mathbb{R}^d} K(y,x) \prod_{i=1}^{d} \varphi_{ij_i}(x_i) \, d\nu(x) = \prod_{i=1}^{d} \int_{\mathbb{R}} K_i(y_i, x_i) \, \varphi_{ij_i}(x_i) \, d\nu(x_i) = \prod_{i=1}^{d} \lambda_{ij_i} \, \varphi_{ij_i}(x_i).$$

Finally, these are the only eigenpairs because the product functions $\prod_{i=1}^{d} \varphi_{ij_i}$ for all possible $j_1, \ldots, j_d \in \mathbb{N}_0$ form a complete orthonormal system of $L^2(\mathbb{R}^d)$ under the base measure ν.

Pick m such that $m > d$, and suppose the n source terms in Theorem 27 are $\{\varphi_{ij} : i \in [d] \text{ and } 0 \leq j \leq m-1\}$. That is, we have $n = m^d$ source terms. So, the upperbound is

$$\varepsilon^2 \sum_{j_1=0}^{m-1} \cdots \sum_{j_d=0}^{m-1} \prod_{i=1}^{d} \lambda_{ij_i} \; + \; \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{\substack{(j_1,\ldots,j_d)\in\mathbb{N}_0^d \\ \max\{j_1,\ldots,j_d\}\geq m}} \prod_{i=1}^{d} \lambda_{ij_i}$$

The first summation is

$$\sum_{j_1=0}^{m-1} \cdots \sum_{j_d=0}^{m-1} \prod_{i=1}^{d} \lambda_{ij_i} = \prod_{i=1}^{d} \sum_{j_i=0}^{m-1} \lambda_{ij_i} \leq \prod_{i=1}^{d} \sum_{j_i=0}^{m-1} \gamma^{j_i} \leq \left(\frac{1-\gamma^m}{1-\gamma}\right)^d \leq \frac{1}{(1-\gamma)^d}.$$

On the other hand,

$$\sum_{\substack{(j_1,\ldots,j_d)\in\mathbb{N}_0^d \\ \max\{j_1,\ldots,j_d\}\geq m}} \prod_{i=1}^{d} \lambda_{ij_i} \leq \sum_{\substack{(j_1,\ldots,j_d)\in\mathbb{N}_0^d \\ \max\{j_1,\ldots,j_d\}\geq m}} \gamma^{j_1+\ldots+j_d} \leq \sum_{r=m}^{\infty} r^d \gamma^r \leq \int_{m-1}^{\infty} r^d \gamma^r \, dr.$$

The second inequality follows because the number of tuple $(j_1, \ldots, j_d)$ that sum to r is $\leq r^d$. It is easy to see that the integral converges faster than $1/n^t$ for every $t \geq 1$. To see this, pick $t \geq 1$. Then, there exists $c > 0$ such that $\gamma^r \leq c\, r^{-dt-1-d}$. Note that c may depend on γ, d, and t, but it does not depend on r. Thus, we obtain

$$\int_{m-1}^{\infty} r^d \gamma^r \, dr \leq c \int_{m-1}^{\infty} r^{-dt-1} \, dr = \frac{c}{(m-1)^{dt}}.$$

Since $m = n^{1/d}$, this rate is c'/n^t for some c'. That is, our overall upper bound is

$$\varepsilon^2 \frac{1}{(1-\gamma)^d} + \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{c'}{n^t}.$$

for some c' for every $t \geq 1$. Therefore, the reducible error vanishes at a rate faster than every polynomial function of $1/n$.

F.2.4 Brownian Motion

Let us consider the case where $\mathcal{X} = [0, 1]$, the base measure ν is Lebegsue, and the stochastic process in Section 7.2.2 is Brownian motion. Recall that the Brownian motion is a Gaussian process with covariance kernel

$$K(s, t) = \min(s, t) \qquad s, t \in [0, 1].$$

It is well-known [Hsing and Eubank, 2015, Example 4.6.3] that the eigenpairs of K is given by

$$\lambda_j := \frac{1}{\left(j - \frac{1}{2}\right)^2 \pi^2} \qquad \text{and} \qquad \varphi_j(t) := \sqrt{2} \sin\left(\left(j - \frac{1}{2}\right) \pi t\right) \qquad \forall j \in \mathbb{N}.$$

Plugging this in the upperbound of Theorem 27 yields the bound

$$\begin{aligned}
&\varepsilon^2 \sum_{j=1}^{n} \frac{1}{\left(j-\frac{1}{2}\right)^2 \pi^2} + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j=n+1}^{\infty} \frac{1}{\left(j-\frac{1}{2}\right)^2 \pi^2} \\
&= \varepsilon^2 \frac{\pi^2}{2} \frac{1}{\pi^2} + \|\mathrm{F}\|_{\mathrm{op}}^2 \sum_{j=n+1}^{\infty} \frac{1}{\left(j-\frac{1}{2}\right)^2 \pi^2} \\
&\leq \frac{\varepsilon^2}{2} + \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{1}{\pi^2} \int_n^{\infty} \frac{1}{(t-1/2)^2}\, dt \\
&= \frac{\varepsilon^2}{2} + \|\mathrm{F}\|_{\mathrm{op}}^2 \frac{1}{\pi^2} \frac{2}{2n-1}.
\end{aligned}$$

Therefore, the reducible error vanishes at rate $\sim \frac{1}{n}$.

F.3 Numerical Approximation of Eigenfunctions

In Section 7.3.3, we provided analytic expressions for the eigenfunctions of certain covariance kernels. However, for some kernels of interest, closed-form expressions for the eigenfunctions are generally not available. In such cases, numerical approximation is necessary. Here, we will briefly mention some key concepts behind the numerical approximation of eigenfunctions of kernels. The material presented here is based on [Williams and Rasmussen, 2006, Section 4.3.2], so we refer the reader to that text for a more detailed discussion and relevant references.

Let $d\nu(x) \propto p(x)\, dx$ for some density function p. For example, if ν is Lebesgue measure on $[-1,1] \times [-1,1]$, then $p(x) = 1/4$. Then, the solution of Feldolm integral

$$\int_{\mathcal{X}} K(y,x)\, \varphi_j(x)\, d\nu(x) = \lambda_j \varphi(y)$$

is approximated using the equation

$$\frac{1}{N} \sum_{i=1}^{N} K(y,x_i)\, \varphi_j(x_i) = \lambda_j \varphi_j(y).$$

Here, $x_1, x_2, \ldots, x_N$ are iid samples from p. Taking $y = x_1, \ldots, x_N$, we obtain a matrix eigenvalue equation

$$\mathbf{K} u_j = \gamma_j\, u_j,$$

where $\mathbf{K}$ is a $N \times N$ matrix such that $[\mathbf{K}] = K(x_i, x_j)$. The sequence $(\gamma_j, u_j)_{j \geq 1}$ is the

eigenpair of $\mathbf{K}$. Then, the estimator for eigenfunctions φ_j's and eigenvalues λ_j's are

$$\varphi_j(x_i) \sim \sqrt{N}\,[u_j]_i \qquad \lambda_j \sim \frac{\gamma_j}{N}.$$

The $\sqrt{N}$ normalization for eigenfunction is to ensure that the squared integral of φ_j on the observed samples is 1. That is,

$$\int_{\mathcal{X}} \varphi_j(x)\,\varphi_j(x)\,d\nu(x) \approx \frac{1}{N}\sum_{i=1}^{N} \varphi_j(x_i)\,\varphi_j(x_i) = \frac{1}{N}\sum_{i=1} \sqrt{N}[u_j]_i \cdot \sqrt{N}\,[u_j]_i = u_j^{\intercal} u_j = 1.$$

As for the eigenvalues, the proposed estimator is consistent. That is, $\gamma_j/N \to \lambda_j$ when $N \to \infty$ [Baker and Taylor, 1979, Theorem 3.4].

The estimator for eigenfunction only allows evaluation on points $x_1, \ldots, x_N$ used to solve the matrix eigenvalue equation. To evaluate the eigenfunction on arbitrary input, one can use a generalized Nyström-type estimator, defined as

$$\varphi_j(y) \sim \frac{\sqrt{N}}{\gamma_j} \sum_{i=1}^{N} K(y, x_i)\,[u_j]_i.$$

F.4 Proof of Lower Bound

Proof. Let $\{\varphi_j\}_{j\in\mathbb{N}}$ be the eigenfunctions of K. That is,

$$\int_{\mathcal{X}} K(y,x)\,\varphi_i(x)\,dx = \lambda_i\,\varphi_i(y) \quad \forall i \in \mathbb{N}.$$

We now construct a hard distribution for the learner. Fix some $p \in (0,1)$ and let $\xi_1, \xi_2, \ldots$ denote the sequence of pairwise independent random variables such that

$$\xi_j = \begin{cases} -\sqrt{1/p} & \text{with probability } \frac{p}{2} \\ 0 & \text{with probability } 1-p \\ \sqrt{1/p} & \text{with probability } \frac{p}{2} \end{cases}.$$

Given such sequence, define a function v such that

$$v(\cdot) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\,\xi_j\,\varphi_j(\cdot).$$

Note that

$$\|v\|_{L^2} = \sum_{j=1}^{\infty} \lambda_j \xi_j^2 < \infty$$

as $\sup_{j\in\mathbb{N}} |\xi_j|^2 \leq 1/p$ and $\sum_{j=1}^{\infty} \lambda_j < \infty$. Thus, v is a random element in $L^2(\mathcal{X})$. Let μ denote the probability measure over $L^2(\mathcal{X})$ induced by the random sequence $\{\xi_j\}_{j\in\mathbb{N}}$. It is easy to see that $\mathbb{E}[v(x)] = 0$ for each $x \in \mathcal{X}$. Moreover, for every $x, y \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}[v(x)\, v(y)] &= \mathbb{E}\left[\left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(x)\right)\left(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \xi_j\, \varphi_j(y)\right)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{\infty} \lambda_j\, \xi_j^2 \varphi_j(x)\, \varphi_j(y) + 2\sum_{i<j} \sqrt{\lambda_i \lambda_j}\, \xi_i \xi_j\, \varphi_i(x)\, \varphi_j(y)\right] \\
&= \sum_{j=1}^{\infty} \lambda_j\ \mathbb{E}[\xi_j^2]\, \varphi_j(x)\, \varphi_j(y) \\
&= \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)\, \varphi_j(y) \\
&= K(y, x),
\end{aligned}$$

where the final equality holds due to Mercer's theorem and the convergence is uniform over $x, y \in \mathcal{X}$. Therefore, we have shown that $\mu \in \mathcal{P}(K)$. Let $\sigma := \{\sigma_j\}_{j\geq 1}$ be a sequence of iid random variables such that $\sigma_j \sim \text{Uniform}(\{-1, 1\})$. Fix $c > 0$ and for each such $\sigma \in \{-1, 1\}^{\mathbb{N}}$, define

$$\mathrm{F}_\sigma := c \sum_{j=1}^{\infty} \sigma_j\, \varphi_j \otimes \varphi_j.$$

For each $m \in \mathbb{N}$, we will show that

$$\mathop{\mathbb{E}}_{\sigma}\left[\mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}_\sigma(v)\right\|_{L^2}^2\right]\right]\right] \geq \frac{c^2}{2}\sum_{j=1}^{m} \lambda_j.$$

Since this holds in expectation, using the probabilistic method, there must be a $\sigma^\star$ such that

$$\mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}_{\sigma^\star}(v)\right\|_{L^2}^2\right]\right] \geq \frac{c^2}{2}\sum_{j=1}^{m} \lambda_j.$$

Noting that $\|\mathrm{F}_{\sigma^\star}\|_{\mathrm{op}} = c$ completes our proof. The rest of the proof will establish this inequality.

Since $\{\varphi_j\}_{j\in\mathbb{N}}$ is the orthonormal bases of $L^2(\mathcal{X})$, Parseval's identity implies that

$$\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}_\sigma(v)\right\|_{L^2}^2 = \sum_{j=1}^{\infty} \left|\left\langle \widehat{\mathrm{F}}_n(v) - \mathrm{F}_\sigma(v), \varphi_j \right\rangle\right|^2 = \sum_{j=1}^{\infty} \left|\left\langle \widehat{\mathrm{F}}_n(v), \varphi_j \right\rangle - \langle \mathrm{F}_\sigma(v), \varphi_j\rangle\right|^2.$$

Recall that $\mathrm{F}^\star(\varphi_j) = c\sigma_j\varphi_j$, where $\mathrm{F}^\star$ is the adjoint of F. Thus, for any $v \sim \mu$, we have

$$\langle \mathrm{F}(v), \varphi_j\rangle = \langle v, \mathrm{F}^\star(\varphi_j)\rangle = \langle v, c\,\sigma_j\,\varphi_j\rangle = c\,\sigma_j\sqrt{\lambda_j}\,\xi_j,$$

which subsequently implies

$$\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}_\sigma(v)\right\|_{L^2}^2 = \sum_{j=1}^{\infty} \left|\left\langle \widehat{\mathrm{F}}_n(v), \varphi_j \right\rangle - c\,\sigma_j\,\sqrt{\lambda_j}\,\xi_j\right|^2.$$

Using this fact, we can write

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma}\left[\mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\left\|\widehat{\mathrm{F}}_n(v) - \mathrm{F}_\sigma(v)\right\|_{L^2}^2\right]\right]\right] \\
&\quad= \mathop{\mathbb{E}}_{\sigma}\left[\mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\sum_{j=1}^{\infty}\left|\left\langle \widehat{\mathrm{F}}_n(v), \varphi_j \right\rangle - c\,\sigma_j\,\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right]\right] \\
&\quad= \mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{j=1}^{\infty}\left|\left\langle \widehat{\mathrm{F}}_n(v), \varphi_j \right\rangle - c\,\sigma_j\,\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right]\right].
\end{aligned}$$

In the final step, we changed the order of integration. Note that drawing n samples of $v_1, \ldots, v_n$ and drawing σ can be done in any order, as they are interchangeable. Finally, the draw of $v \sim \mu$ occurs during the test phase, independent of the previously drawn samples $v_{1:n}$ and σ.

Next, let $E_{n,m}$ denote the event such that

$$\langle v_i, \varphi_j\rangle = 0 \qquad \forall 1 \le i \le n \text{ and } 1 \le j \le m.$$

Then, we will lowerbound

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{j=1}^{\infty}\left|\left\langle \widehat{\mathrm{F}}_n(v), \varphi_j \right\rangle - c\,\sigma_j\,\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right]$$

conditioned on the event $E_{n,m}$. First, note that

$$\begin{aligned}\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{j=1}^{\infty}\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right] &\geq \mathop{\mathbb{E}}_{v\sim\mu}\left[\sum_{j=1}^{m}\mathop{\mathbb{E}}_{\sigma}\left[\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right] \\ &\geq \mathop{\mathbb{E}}_{v\sim\mu}\left[\sum_{j=1}^{m}\left(\mathop{\mathbb{E}}_{\sigma}\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|\right)^2\right],\end{aligned}$$

where the final step uses Jensen's inequality. Next, we use the fact that when the event $E_{n,m}$ occurs, the learner has no information about $\sigma_1,\ldots,\sigma_m$. This is because the input data shows no variation along the directions spanned by $\varphi_1,\ldots,\varphi_m$. Given that $\mathcal{O}$ is the perfect oracle for F_σ, any information provided by the oracle $\mathcal{O}$ must be independent of how F_σ operates on the subspace spanned by $\varphi_1,\ldots,\varphi_m$. Specifically, for every $1\leq i\leq n$ and $1\leq j\leq m$, the output of the oracle $\mathcal{O}(v_i)$ must be independent of σ_j. If this condition holds, then the estimator $\widehat{\mathrm{F}}_n$ must also be independent of $\sigma_1,\ldots,\sigma_m$. Thus, conditioned on the event $E_{n,m}$, for any $1\leq j\leq m$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma}\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\sigma_j\sqrt{\lambda_j}\,\xi_j\right| &= \mathbb{E}\left[\mathop{\mathbb{E}}_{\sigma_j}\left[\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|\,\middle|\,\sigma\backslash\{\sigma_j\}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{2}\left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle - c\sqrt{\lambda_j}\,\xi_j\right| + \left|\left\langle\widehat{\mathrm{F}}_n(v),\varphi_j\right\rangle + c\sqrt{\lambda_j}\,\xi_j\right|\right] \\ &\geq \frac{1}{2}\left|2\,c\sqrt{\lambda_j}\xi_j\right| \\ &= |c\sqrt{\lambda_j}\,\xi_j|.\end{aligned}$$

The first equality uses the fact that conditioned on $\sigma\backslash\{\sigma_j\}$, the function $\widehat{\mathrm{F}}_n(v)$ is independent of σ_j. Thus, conditioned on the event $E_{n,m}$, we have shown that

$$\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{j=1}^{\infty}\left|\left\langle\widehat{\mathrm{F}}(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right] \geq \mathop{\mathbb{E}}_{v\sim\mu}\left[\sum_{j=1}^{m}c^2\,\lambda_j\xi_j^2\right] = c^2\sum_{j=1}^{m}\lambda_j\,\mathbb{E}[\xi_j^2] = c^2\sum_{j=1}^{m}\lambda_j.$$

Therefore, our overall lowerbound is

$$\begin{aligned}&\mathop{\mathbb{E}}_{v_{1:n}\sim\mu^n}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\sigma}\left[\sum_{j=1}^{\infty}\left|\left\langle\widehat{\mathrm{F}}(v),\varphi_j\right\rangle - c\,\sigma_j\sqrt{\lambda_j}\,\xi_j\right|^2\right]\right]\right] \\ &\geq c^2\left(\sum_{j=1}^{m}\lambda_j\right)\,\mathbb{P}\left[E_{n,m}\right] = c^2(1-p)^{n\cdot m}\sum_{j=1}^{m}\lambda_j.\end{aligned}$$

The final step uses the fact that $\mathbb{P}[E_{n,m}] = (1-p)^{n\cdot m}$. It now remains to pick p to obtain

the claimed lowerbound. Let us pick $p = \frac{1}{2mn}$. Then, we have $(1-p)^{mn} \geq 1/2$ as long as $n \geq 1$, yielding the lowerbound of

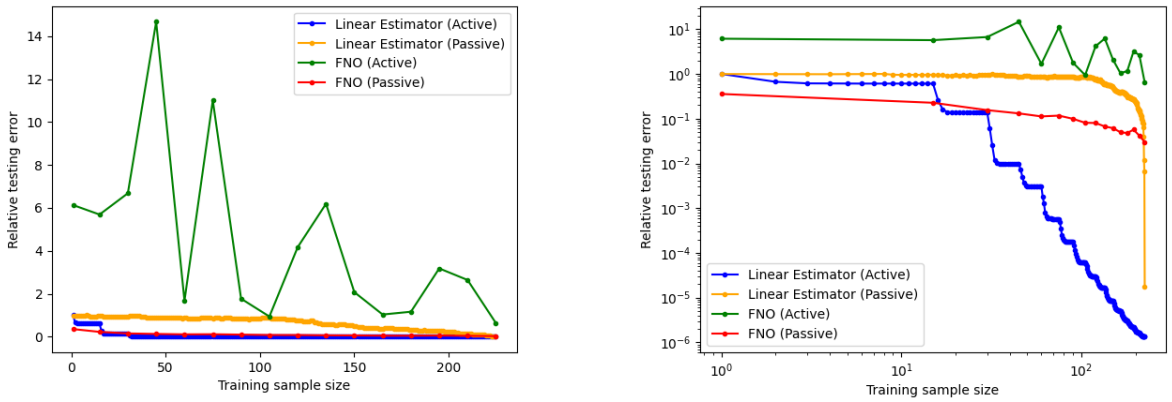
$$\frac{c^2}{2} \sum_{j=1}^{m} \lambda_j.$$

Since $m \in \mathbb{N}$ is arbitrary, our lowerbound holds for every fixed m. Noting that $\|\mathrm{F}_\sigma\|_{\mathrm{op}} = c$ for every σ completes our proof. ■

F.5 Experiments

This section presents additional experimental results using the same setup as described in Section 7.5. The results show that the Fourier Neural Operator (FNO) performs poorly with actively collected data. This is likely because the training data are not i.i.d. samples from the test distribution, requiring FNO to generalize out of distribution when trained on actively collected data.

F.5.1 Poisson Equation



(a) (b)

Figure F.1: Error Plots for various estimators for Poisson Equation. The plot on the right shows the same plot in log scale.

F.5.2 Heat Equation

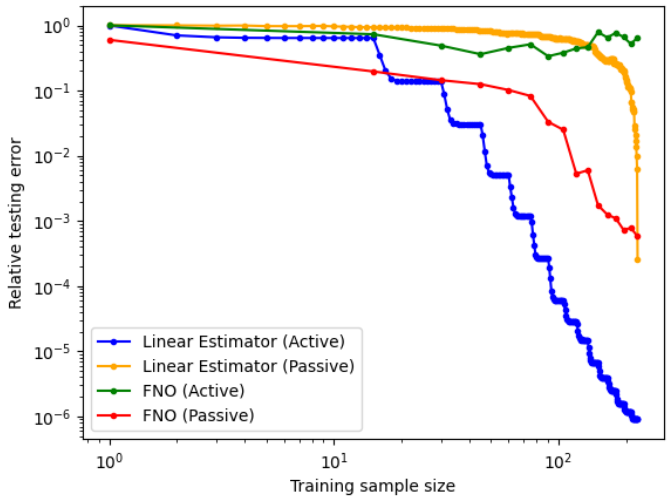


Figure F.2: Error Plots for various estimators for Heat Equations in log-log scale.

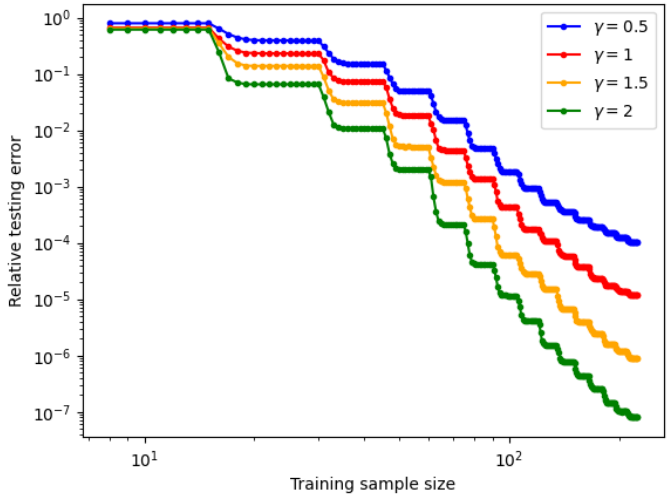


Figure F.3: Convergence rate of the active linear estimator for Heat equation with actively collected data for different values of γ.

APPENDIX G

Is Zero-Shot Super-Resolution Possible In Operator Learning?

G.1 Proof of Theorem 29

G.1.1 Constructing the Class $\mathcal{G}$

Define a function $g : [0,1)^2 \to [0,1]$ such that

$$g(y,x) = yx.$$

Recall that the integral operator of the function $g(y,x)$ is an operator G_{base} such that

$$\Big(\mathrm{G}_{\text{base}}(v)\Big)(y) = \int_0^1 g(y,x)\, v(x)\, dx.$$

We will first modify g for certain values of y and take our ground truth to be the integral operator of that modified g.

For any sequence $\xi := \{\xi_y\}_{y \in \mathbb{Q}\cap[0,1)}$, where each $\xi_y \sim \mathrm{Unif}(\{-1,1\})$ independently, define a function $f_\xi : [0,1) \to [-1,1]$ by

$$f_\xi(y) := \begin{cases} \xi_y, & \text{if } y \in \mathbb{Q}\cap[0,1), \\ y, & \text{otherwise.} \end{cases}$$

Since $f_\xi(y) = y$ almost everywhere and the identity function $y \mapsto y$ is measurable, it follows by Folland [1999, Proposition 2.11] that f_ξ is also measurable.

Now, for each sequence $\xi \in \{-1,1\}^{\mathbb{Q}\cap[0,1)}$, define a function $g_\xi : [0,1)\times[0,1) \to \mathbb{R}$ by

$$g_\xi(y,x) := f_\xi(y)\, x.$$

Note that $g(y,x) = g_\xi(y,x)$ almost everywhere.

Let G_ξ denote the integral operator associated with this function. By construction, we have

$$\mathrm{G}_\xi(v)(y) = f_\xi(y) \cdot \int_0^1 x\, v(x)\, dx,$$

which implies that $\mathrm{G}_\xi(v) \in \mathrm{span}(f_\xi)$ for all $v \in L^2([0,1))$. That is, G_ξ is a rank-one operator. Moreover,

$$\|\mathrm{G}_\xi\|_{\mathrm{op}}^2 = \|\mathrm{G}_\xi\|_{\mathrm{HS}}^2 = \int_0^1 \int_0^1 |g_\xi(y,x)|^2\, dy\, dx = \int_0^1 \int_0^1 |g(y,x)|^2\, dy\, dx = \left(\int_0^1 r^2\, dr\right)^2 = \frac{1}{9}.$$

Thus, $\|\mathrm{G}_\xi\|_{\mathrm{op}} \leq 1$. Define a class of operators

$$\mathcal{G} := \left\{\mathrm{G}_\xi \mid \xi \in \{-1,1\}^{\mathbb{Q}\cap[0,1)}\right\}.$$

G.1.2 Proof of part (i)

Let μ be any probability measure supported on $L^2(\mathcal{X})$, such that every $v \sim \mu$ satisfies $0 < a \leq |v(x)| \leq 1$ for all $x \in \mathcal{X}$. Suppose the ground truth operator $\mathrm{G} \in \mathcal{G}$. By definition of $\mathcal{G}$, there exists a sequence $\xi := \{\xi_y\}_{y\in\mathbb{Q}\cap[0,1)}$ such that $\mathrm{G} = \mathrm{G}_\xi$.

Consider any sample (v_1, w_1). Then,

$$w_1(y) = \mathrm{G}_\xi(v_1)(y) = \int_0^1 g_\xi(y,x)\, v_1(x)\, dx = f_\xi(y) \int_0^1 x\, v_1(x)\, dx.$$

Since $f_\xi(y) = \xi_y$ for all $y \in \mathcal{Y}_{\mathrm{train}} \subseteq \mathbb{Q} \cap [0,1)$, and

$$\int_0^1 x v_1(x)\, dx \geq c \int_0^1 x\, dx = \frac{c}{2} > 0,$$

we can recover ξ_y for all $y \in \mathcal{Y}_{\mathrm{train}}$ via

$$\xi_y = \frac{w_1(y)}{\int_0^1 x v_1(x)\, dx}.$$

Now define an estimator $\widehat{\mathrm{F}}' := \mathrm{G}_{\xi'}$ such that

$$\xi'_y = \xi_y \quad \text{for all } y \in \mathcal{Y}_{\mathrm{train}}.$$

Then, for any $v \in L^2(\mathcal{X})$,

$$\widehat{\mathrm{F}}'(v)(y) = \xi'_y \int_0^1 xv(x)\,dx = \xi_y \int_0^1 xv(x)\,dx = \mathrm{G}_\xi(v)(y), \quad \forall y \in \mathcal{Y}_{\text{train}}.$$

Therefore,

$$\mathbb{E}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{train}}|}\sum_{y\in\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}'(v)(y) - \mathrm{G}(v)(y)\right)^2\right] = 0.$$

This completes our proof of part (i).

G.1.3 Proof of part (ii)

Let μ be any probability measure supported on $L^2(\mathcal{X})$ such that every $v \sim \mu$ satisfies $0 < a \le |v(x)| \le 1$ for all $x \in \mathcal{X}$.

Our proof is based on probabilistic method. In particular, we show that for any estimation rule $\widehat{\mathrm{F}}_n$, the following bound holds

$$\mathop{\mathbb{E}}_{\xi}\left[\mathop{\mathbb{E}}_{v_1,\ldots,v_n\sim\mu}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}(v)(y) - \mathrm{G}_\xi(v)(y)\right)^2\right]\right]\right] \ge \frac{a^2}{4}\left(1 - \frac{N_1}{N_2}\right). \tag{G.1}$$

By a standard probabilistic argument (see Alon and Spencer [2016]), it follows that there exists a particular sequence $\xi^\star$ and the corresponding operator $\mathrm{G} := \mathrm{G}_{\xi^\star}$ such that

$$\mathop{\mathbb{E}}_{v_1,\ldots,v_n\sim\mu}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}(v)(y) - \mathrm{G}^\star(v)(y)\right)^2\right]\right] \ge \frac{a^2}{4}\left(1 - \frac{N_1}{N_2}\right).$$

Using the fact that $N_2 = mN_1$ completes our proof.

We now proceed to prove Equation (G.1). We can rewrite the left-hand side of Equation (G.1) and restrict the summation to points in the test grid not included in the training grid as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\xi}\left[\mathop{\mathbb{E}}_{v_1,\ldots,v_n\sim\mu}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}(v)(y) - \mathrm{G}_\xi(v)(y)\right)^2\right]\right]\right] \\
&\ge \mathop{\mathbb{E}}_{v_1,\ldots,v_n\sim\mu}\left[\mathop{\mathbb{E}}_{v\sim\mu}\left[\mathop{\mathbb{E}}_{\xi}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}_\xi(v)(y)\right)^2\right]\right]\right],
\end{aligned}$$

where the final step follows by Fubini's theorem, since the random variables $v_1, \ldots, v_n \sim \mu$, the test sample $v \sim \mu$, and the sequence $\xi \sim \mathrm{Unif}(\{-1,1\})^{\mathbb{Q}\cap[0,1)}$ are all drawn independently. In particular, the order of integration over ξ, the training data, and the test data can be

exchanged freely.

Note that for any $y \in \mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}$, the squared error can be lower bounded as

$$\begin{aligned}\left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}_\xi(v)(y)\right)^2 &= \left(\widehat{\mathrm{F}}_n(v)(y) - \xi_y \int_0^1 e^{-x} v(x)\,dx\right)^2 \\ &= \left(\widehat{\mathrm{F}}_n(v)(y)\right)^2 - 2\widehat{\mathrm{F}}_n(v)(y) \cdot \xi_y \int_0^1 x v(x)\,dx + \left(\int_0^1 x v(x)\,dx\right)^2 \\ &\geq -2\widehat{\mathrm{F}}_n(v)(y) \cdot \xi_y \int_0^1 x v(x)\,dx + \left(\int_0^1 x v(x)\,dx\right)^2.\end{aligned}$$

Next, we show that the first term vanishes in expectation. More precisely, for fixed training samples $v_1, \ldots, v_n$ and a test point $v \sim \mu$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\xi}\left[\widehat{\mathrm{F}}_n(v)(y) \cdot \xi_y \int_0^1 x v(x)\,dx\right] &= \mathop{\mathbb{E}}_{\xi \setminus \xi_y}\left[\mathop{\mathbb{E}}_{\xi_y}\left[\widehat{\mathrm{F}}_n(v)(y) \cdot \xi_y \int_0^1 x v(x)\,dx \,\middle|\, \xi \setminus \xi_y\right]\right] \\ &= \mathop{\mathbb{E}}_{\xi \setminus \xi_y}\left[\widehat{\mathrm{F}}_n(v)(y) \cdot \left(\int_0^1 x v(x)\,dx\right) \cdot \mathop{\mathbb{E}}_{\xi_y}\left[\xi_y \,\middle|\, \xi \setminus \xi_y\right]\right] \\ &= 0,\end{aligned}$$

since $\mathbb{E}_{\xi_y}[\xi_y \mid \xi \setminus \xi_y] = 0$. The first step uses the law of iterated expectation. The second step follows from the fact that, for any fixed $v_1, \ldots, v_n$ and v, the random variable ξ_y is conditionally independent of $\widehat{\mathrm{F}}_n(v)(y)$ given $\xi \setminus \xi_y$. This is because the estimator $\widehat{\mathrm{F}}_n$ only has access to the values of ξ_y at the training grid $\mathcal{Y}_{\text{train}}$. For any $y \notin \mathcal{Y}_{\text{train}}$, the learner receives no information about ξ_y, and therefore the estimator's prediction at such y must be independent of ξ_y conditioned on $\xi \setminus \xi_y$.

To bound the second term, observe that since $v(x) \geq a$ for all $x \in [0, 1)$, we have

$$\left(\int_0^1 x v(x)\,dx\right)^2 \geq \left(\int_0^1 x \cdot a\,dx\right)^2 = a^2\left(\int_0^1 x\,dx\right)^2 = \frac{a^2}{4}.$$

Combining this bound with the previous steps, we conclude that

$$\begin{aligned}&\mathop{\mathbb{E}}_{\xi}\left[\mathop{\mathbb{E}}_{v_1,\ldots,v_n \sim \mu}\left[\mathop{\mathbb{E}}_{v \sim \mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y \in \mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}(v)(y) - \mathrm{G}_\xi(v)(y)\right)^2\right]\right]\right] \\ &\geq \frac{1}{\mathcal{Y}_{\text{test}}}\sum_{y \in \mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}} \frac{a^2}{4} \\ &= \frac{a^2}{4} \cdot \frac{|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}.\end{aligned}$$

Noting that $|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}| \geq |\mathcal{Y}_{\text{test}}| - |\mathcal{Y}_{\text{train}}|$ completes our proof of part (ii).

G.2 Proof of Theorem 30

Proof. We begin by decomposing the error over the test grid as

$$\begin{aligned}
&\mathbb{E}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}}\left(\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right)^2\right]\\
&=\mathbb{E}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right)^2\right]\\
&+\mathbb{E}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right)^2\right].
\end{aligned}$$

For each $y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}$, let $\mathrm{nn}(y)\in\mathcal{Y}_{\text{train}}$ denote a nearest neighbor of y in the training grid, that is

$$|y-\mathrm{nn}(y)|_2\le|y-y'|_2\quad\text{for all } y'\in\mathcal{Y}_{\text{train}}.$$

Here, ties may be broken arbitrarily. Then for any such y, we have:

$$\begin{aligned}
&\left|\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right|\\
&=\left|\widehat{\mathrm{F}}_n(v)(y)-\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))+\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))+\mathrm{G}(v)(\mathrm{nn}(y))-\mathrm{G}(v)(y)\right|\\
&\le\left|\widehat{\mathrm{F}}_n(v)(y)-\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))\right|+\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|+|\mathrm{G}(v)(\mathrm{nn}(y))-\mathrm{G}(v)(y)|\\
&\le c|y-\mathrm{nn}(y)|_2^\alpha+\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|+c|y-\mathrm{nn}(y)|_2^\alpha\\
&=2c|y-\mathrm{nn}(y)|_2^\alpha+\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|,
\end{aligned}$$

where we have used the assumption that both $\widehat{\mathrm{F}}_n(v)$ and $\mathrm{G}(v)$ are uniformly Hölder continuous with exponent α and constant c. Using the inequality $(a+b)^2\le 2(a^2+b^2)$, we have

$$\begin{aligned}
\left|\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right|^2&\le 8c^2|y-\mathrm{nn}(y)|_2^{2\alpha}+2\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|^2\\
&\le 8c^2\beta^{2\alpha}+2\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}}\left(\widehat{\mathrm{F}}_n(v)(y)-\mathrm{G}(v)(y)\right)^2\right]\\
&\le\frac{|\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}\cdot 8c^2\beta^{2\alpha}+2\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|}\sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}}\left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y))-\mathrm{G}(v)(\mathrm{nn}(y))\right|^2.\right]
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}} \left|\widehat{\mathrm{F}}_n(v)(\mathrm{nn}(y)) - \mathrm{G}(v)(\mathrm{nn}(y))\right|^2 \\
&\qquad = \sum_{z\in\mathcal{Y}_{\text{train}}} \sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}} \left|\widehat{\mathrm{F}}_n(v)(z) - \mathrm{G}(v)(z)\right|^2 \mathbb{1}[\mathrm{nn}(y) = z] \\
&\qquad \leq \left(\sup_{z\in\mathcal{Y}_{\text{train}}} \sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}} \mathbb{1}[\mathrm{nn}(y) = z]\right) \cdot \sum_{z\in\mathcal{Y}_{\text{train}}} \left|\widehat{\mathrm{F}}_n(v)(z) - \mathrm{G}(v)(z)\right|^2 \\
&\qquad = (\nu - 1) \cdot \sum_{z\in\mathcal{Y}_{\text{train}}} \left|\widehat{\mathrm{F}}_n(v)(z) - \mathrm{G}(v)(z)\right|^2.
\end{aligned}$$

Note that the final equality holds because the point itself is the nearest neighbor for all $y \in \mathcal{Y}_{\text{train}}$. Thus, we have shown that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{y\in\mathcal{Y}_{\text{test}}\setminus\mathcal{Y}_{\text{train}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}(v)(y)\right)^2\right] \\
&\leq \frac{|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} \cdot 8c^2\beta^{2\alpha} + 2(\nu - 1) \cdot \mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{z\in\mathcal{Y}_{\text{train}}} \left|\widehat{\mathrm{F}}_n(v)(z) - \mathrm{G}(v)(z)\right|^2\right] \\
&= \frac{|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} \cdot 8c^2\beta^{2\alpha} + \frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} \cdot 2(\nu - 1) \cdot \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}).
\end{aligned}$$

Therefore, by combining everything, we have shown that

$$\begin{aligned}
&\mathbb{E}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\text{test}}|} \sum_{y\in\mathcal{Y}_{\text{test}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \mathrm{G}(v)(y)\right)^2\right] \\
&\qquad \leq \frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} \cdot (2\nu - 1) \cdot \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}) + \frac{|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} \cdot 8c^2\beta^{2\alpha}.
\end{aligned}$$

Here, $2(\nu-1)$ becomes $2\nu-1$ as we also account for extra term from the decomposition in the first step of this proof. Finally, noting that $|\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}| = |\mathcal{Y}_{\text{test}}| - |\mathcal{Y}_{\text{train}}|$ for $\mathcal{Y}_{\text{train}} \subseteq \mathcal{Y}_{\text{test}}$ completes our proof.

■

G.3 Proof of Corollary 3

Proof. Let $y := (y_1, \ldots, y_d) \in \mathcal{Y}_{\text{test}}$. By construction of the test grid, each coordinate can be written as $y_j = \frac{r_j}{N_2}$ for some $r_j \in \{0, 1, \ldots, N_2 - 1\}$. Since $N_2 = mN_1$, we can further express

$$y_j = \frac{r_j}{m} \cdot \frac{1}{N_1}.$$

Now define a point $y' := (y'_1, \ldots, y'_d) \in \mathcal{Y}_{\text{train}}$ by rounding each r_j/m to its nearest integer:

$$y'_j := \frac{[r_j/m]}{N_1},$$

where $[r_j/m]$ denotes the closest integer to r_j/m in $\{0, 1, \ldots, N_1 - 1\}$, with ties rounded down. Note that although r_j/m can exceed $N_1 - 1$, we always map it to $N_1 - 1$, as $1 = N_1/N_1$ is not in the training grid.

By this construction, $y' \in \mathcal{Y}_{\text{train}}$ is the nearest neighbor of y, and the Euclidean distance satisfies

$$\beta := |y - y'|_2 \leq \frac{1}{N_1} \sqrt{\sum_{j=1}^{d} \left| \frac{r_j}{m} - \left[\frac{r_j}{m} \right] \right|^2} \leq \frac{\sqrt{d}}{N_1}.$$

Next, to bound ν, consider how many test points $y \in \mathcal{Y}_{\text{test}}$ can be mapped to a fixed training point $y' \in \mathcal{Y}_{\text{train}}$. Fix a coordinate $j \in \{1, \ldots, d\}$, and suppose $y'_j = \frac{t_j}{N_1}$, where $t_j \in \{0, \ldots, N_1 - 1\} \setminus \{N_1 - 1\}$. Then $y_j = \frac{r_j}{m} \cdot \frac{1}{N_1}$ is mapped to y'_j if and only if

$$\frac{r_j}{m} \in \left(t_j - \frac{1}{2}, t_j + \frac{1}{2} \right] \quad \Longleftrightarrow \quad r_j \in \left(mt_j - \frac{m}{2}, mt_j + \frac{m}{2} \right].$$

Since $r_j \in \{0, \ldots, N_2 - 1\}$, this interval contains at most m integer values. However, in the edge case when $t_j = N_1 - 1$, the interval becomes

$$r_j \in \left(m(N_1 - 2) + \frac{m}{2}, mN_1 - 1 \right],$$

which contains at most $1.5m$ integer values.

Therefore, for each coordinate, there are at most $1.5m$ admissible values of r_j, so the total number of test points assigned to a single training point is bounded by

$$\nu \leq (1.5m)^d.$$

Finally, noting

$$\frac{|\mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|} = \frac{1}{m^d}$$

and

$$2\nu - 1 = 2(1.5m)^d - 1 \leq 2^{d+1}$$

completes our proof. ■

G.4 Refined Bound for Uniform Grid

Recall that in the proof of Corollary 30, each test point $y \in \mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}$ is assigned to its nearest neighbor in $\mathcal{Y}_{\text{train}}$. In a uniform grid, this nearest-neighbor mapping can lead to edge cases near the boundary (e.g., near $y_j = 1$), where a disproportionate number of test points may be mapped to a single training point. To address this, one can modify the assignment rule to use a slightly adjusted rounding scheme rather than strict nearest-neighbor mapping, ensuring a more balanced reuse across the grid.

To that end, $y := (y_1, \ldots, y_d) \in \mathcal{Y}_{\text{test}}$ such that $y_j = \frac{r_j}{mN_1}$ for some $r_j \in \{0, 1, \ldots, N_2 - 1\}$. Now define a point $y' := (y'_1, \ldots, y'_d) \in \mathcal{Y}_{\text{train}}$ by mapping each y_j to

$$y'_j := \frac{\lfloor r_j/m \rfloor}{N_1}.$$

Instead of mapping r_j/m to the closest integer, we map it to its floor value, which is close but not the nearest neighbor. Clearly,

$$|y - y'|_2 \leq \frac{1}{N_1} \sqrt{\sum_{j=1}^{d} \left| \frac{r_j}{m} - \left\lfloor \frac{r_j}{m} \right\rfloor \right|^2} \leq \frac{\sqrt{d}}{N_1}.$$

Next, we bound the number of test points $y \in \mathcal{Y}_{\text{test}}$ can be mapped to a fixed training point $y' \in \mathcal{Y}_{\text{train}}$. Fix a coordinate $j \in \{1, \ldots, d\}$, and suppose $y'_j = \frac{t_j}{N_1}$, where $t_j \in \{0, \ldots, N_1 - 1\}$. Then $y_j = \frac{r_j}{m} \cdot \frac{1}{N_1}$ is mapped to y'_j if and only if

$$\frac{r_j}{m} \in [t_j, t_j + 1) \quad \iff \quad r_j \in [mt_j, mt_j + m)\,.$$

Thus, r can take $\leq m$ values.

Therefore, for each coordinate, there are at most m admissible values of r_j, so the total

number of test points assigned to a single training point is bounded by

$$m^d.$$

Redoing the proof of Theorem 30 where each $y \in \mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}}$ is mapped to $y' \in \mathcal{Y}_{\text{train}}$ yields the claimed bound of

$$2\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{Y}_{\text{train}}) + \left(1 - \frac{1}{m^d}\right) \cdot 8c^2 \cdot \left(\frac{\sqrt{d}}{N_1}\right)^\alpha.$$

G.5 Proof of Proposition 2

Let

$$u(y) := \int_\Omega g(y, x)\, f(x)\, dx.$$

Then for any $y_1, y_2 \in \Omega$, we have

$$\begin{aligned}
|u(y_1) - u(y_2)| &= \left| \int_\Omega \Big(g(y_1, x) - g(y_2, x) \Big)\, f(x)\, dx \right| \\
&\leq \int_\Omega |g(y_1, x) - g(y_2, x)|\, |f(x)|\, dx \\
&\leq \left(\int_\Omega |g(y_1, x) - g(y_2, x)|^2 dx \right)^{1/2} \cdot \|f\|_{L^2} \\
&\leq c\, |y_1 - y_2|^\alpha \cdot \sqrt{\mathrm{vol}(\Omega)} \cdot \|f\|_{L^2}.
\end{aligned}$$

Here, c is a uniform bound on the Hölder coefficient of $y \mapsto g(y, x)$, valid for almost every $x \in \Omega$. Since Ω is bounded, we conclude that $u \in C^{0,\alpha}(\Omega)$, completing the proof.

G.6 Proof of Proposition 4

Proof. Let

$$w(y) := \sigma\Big(\int_{\mathcal{X}} k(y, x)\, v(x)\, dx + b(y) \Big).$$

Then, using the fact that $\sigma(\cdot)$ is 1-Lipschitz, we have

$$\begin{aligned}
|w(y_1) - w(y_2)| &= \left|\sigma\Big(\int_{\mathcal{X}} k(y_1, x)\, v(x)\, dx + b(y_1)\Big) - \sigma\Big(\int_{\mathcal{X}} k(y_2, x)\, v(x)\, dx + b(y_2)\Big)\right| \\
&\leq \left|\int_{\mathcal{X}} k(y_1, x)\, v(x)\, dx + b(y_1) - \int_{\mathcal{X}} k(y_2, x)\, v(x)\, dx - b(y_2)\right| \\
&= \left|\int_{\mathcal{X}} (k(y_1, x) - k(y_2, x))\, v(x)\, dx + b(y_1) - b(y_2)\right| \\
&\leq \left|\int_{\mathcal{X}} (k(y_1, x) - k(y_2, x))\, v(x)\, dx\right| + |b(y_1) - b(y_2)|
\end{aligned}$$

Note that $|b(y_1) - b(y_2)| \leq [b]\, |y_1 - y_2|^\alpha$. Similarly,

$$\begin{aligned}
&\left|\int_{\mathcal{X}} (k(y_1, x) - k(y_2, x))\, v(x)\, dx\right| \\
&= \sqrt{\sup_x |k(y_1, x) - k(y_2, x)|^2 \ \mathrm{vol}(\mathcal{X})}\ \ \|v\|_{L^2} \\
&= [k]\, |y_1 - y_2|^\alpha \sqrt{\mathrm{vol}(\mathcal{X})}\ \ \|v\|_{L^2}\,,
\end{aligned}$$

where $[k]$ is supremum over all Holder coefficient $[k(\cdot, x)]$. Recall that $[k] < \infty$ by assumption. Thus, combining everything, we have

$$|w(y_1) - w(y_2)| \leq \Big([k] \sqrt{\mathrm{vol}(\mathcal{X})}\ \|v\|_{L^2} + [b]\Big)\, |y_1 - y_2|^\alpha.$$

This completes our proof. ■

G.7 Proof of Theorem 31

Note that

$$\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{test}}, \mathcal{Y}_{\mathrm{test}}) = \mathop{\mathbb{E}}_{v\sim\mu}\left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} \Big(\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \mathrm{G}(v)(y)\Big)^2\right].$$

We can write

$$\begin{aligned}
&|\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \mathrm{G}(v)(y)| \\
&= |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y) + \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y) - \mathrm{G}(v)(y)| \\
&\leq |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y)| + |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y) - \mathrm{G}(v)(y)|
\end{aligned}$$

Thus, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$\begin{aligned}
&\mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{test}}, \mathcal{Y}_{\mathrm{test}}) \\
&= 2 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y)|^2 \right] \\
&+ 2 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y) - \mathrm{G}(v)(y)|^2 \right]
\end{aligned}$$

For the second term, using the same arguments as in the proof of Theorem 30, we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} \left|\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y) - \mathrm{G}(v)(y)\right|^2 \right] &\leq \frac{|\mathcal{Y}_{\mathrm{train}}|}{|\mathcal{Y}_{\mathrm{test}}|} \cdot (2\nu - 1) \cdot \mathcal{E}_\mu(\widehat{\mathrm{F}}_n, \mathrm{G}, \mathcal{X}_{\mathrm{train}}, \mathcal{Y}_{\mathrm{train}}) \\
&+ \frac{|\mathcal{Y}_{\mathrm{test}} \setminus \mathcal{Y}_{\mathrm{train}}|}{|\mathcal{Y}_{\mathrm{test}}|} \cdot 8c^2\beta^{2\alpha}.
\end{aligned}$$

This completes the contribution of the second term in the bound stated in Theorem 31.

So, it remains to bound

$$2 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} |\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y)|^2 \right].$$

Again using $(a+b)^2 \leq 2a^2 + 2b^2$, we can further write

$$\begin{aligned}
&2 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} \left(\widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y)\right)^2 \right] \\
&\leq 4 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{test}}})(y)\right)^2 \right] \\
&\quad + 4 \mathop{\mathbb{E}}_{v\sim\mu} \left[\frac{1}{|\mathcal{Y}_{\mathrm{test}}|} \sum_{y\in\mathcal{Y}_{\mathrm{test}}} \left(\widehat{\mathrm{F}}_n(v)(y) - \widehat{\mathrm{F}}_n(v_{|\mathcal{X}_{\mathrm{train}}})(y)\right)^2 \right] \\
&\leq 4(\varepsilon_{k_1} + \varepsilon_{k_2}).
\end{aligned}$$

The final step uses the fact that $\mathcal{X}_{\mathrm{train}} = \mathcal{X}_{k_1}$ and $\mathcal{X}_{\mathrm{test}} = \mathcal{X}_{k_2}$.

APPENDIX H

Operator Learning for Schrödinger Equation: Unitarity, Error Bounds, and Time Generalization

H.1 Extended Related Works

In recent years, there has been considerable work on using machine learning methods to solve the static (time-independent) Schrödinger equation for many-body electronic systems. See the review article by Hermann et al. [2023] for an overview. These methods typically parametrize the ground state wave function ψ_θ using a neural network and optimize the parameters by minimizing the energy functional $\langle \psi_\theta, \mathrm{H}\,\psi_\theta \rangle_{L^2}$. This framework has also been extended to the time-dependent Schrödinger equation for many-electron systems by Nys et al. [2024]. This line of work is closely related to Physics-Informed Neural Networks (PINNs), which approximate solutions to PDEs by fitting a neural network ansatz that satisfies the variational form of the governing equations; see Shah et al. [2022] and [Cuomo et al., 2022, Section 3.2.2.3]. However, these methods effectively act as solvers, requiring optimization for each new instance, and thus do not amortize computational costs. In contrast, our focus is on learning the global evolution operator directly from data, enabling fast and efficient evaluation for new initial conditions without retraining, thereby significantly reducing downstream computational cost.

An early work in learning solution operators for the Schrödinger equation was by Mills et al. [2017], who trained a neural network to predict ground-state wave functions from potentials for the time-independent Schrödinger equation. More recently, Stepaniants [2023] proposed an operator learning approach that models the solution operator mapping potentials to ground state wave functions by learning the associated Green's functions in a reproducing kernel Hilbert space (RKHS). A similar strategy was studied by Boullé et al. [2022], who used rotational neural networks to learn Green's functions for static Schrödinger equations. In addition, Boullé et al. [2022] also considered learning the green functions associated with time dependent propagator for 1-dimensional Harmonic oscillator.

A slightly more general framework was studied by Mizera [2023], who used Fourier Neural Operators (FNOs) [Li et al., 2021] to estimate the time evolution operator for simple quantum systems, such as random potentials and the double-slit potential. They also studied the ability of the learned operator to generalize across time, extrapolating beyond the training time range. Their learned operator is more flexible than ours in that it takes both the initial wave and the potential function as inputs, rather than assuming a fixed Hamiltonian. Relatedly, Niarchos and Papageorgakis [2024] studied learning the phases of amplitudes in scattering problems. Beyond isolated systems, Zhang et al. [2024] and Zhang et al. [2025a] extended FNO-based architectures to model dissipative quantum systems that interact with an environment and are possibly driven by external fields, again evaluating time generalization. Most recently, Shah et al. [2024] trained FNOs to learn the evolution operator for relatively larger quantum spin systems (up to 8-qubit systems), studying both single-step and multi-step time extrapolation.

H.2 Extensions to Non-Periodic Domains

Extending the results from Sections 9.3 and 9.5 to general *bounded* domain $\Omega \subset \mathbb{R}^d$ is straightforward. This requires choosing an orthonormal basis of $L^2(\Omega)$ and defining a corresponding Sobolev-type space. While any orthonormal basis of $L^2(\Omega)$ could be used, the most natural choice is the eigenfunctions of the Laplacian, which satisfy the eigenvalue problem,

$$-\Delta u = \lambda u, \quad \text{subject to appropriate boundary conditions.}$$

Common boundary conditions include Dirichlet, Neumann, and Robin.

For the special case $\Omega = \mathbb{T}^d$, the Laplacian eigenvalues are $\{4\pi^2|k|_2^2 : k \in \mathbb{Z}^d\}$, and the corresponding eigenfunctions are Fourier modes $\{\varphi_k\}_{k\in\mathbb{Z}^d}$. This motivates defining a more general Sobolev-type space using the eigenpairs of the Laplacian. Let $\{\lambda_j\}_{j=1}^\infty$ denote the eigenvalues such that $0 < \lambda_1 \le \lambda_2 \le \ldots$, and let $\{\phi_j\}_{j=1}^\infty$ be the corresponding eigenfunctions. Then, the Sobolev-type space is defined as

$$\mathcal{H}^s(\Omega) = \left\{ f \in L^2(\Omega) \,\middle|\, \sum_{j=1}^{\infty} (1 + |\lambda_j|)^s \, |\langle f, \phi_j \rangle_{L^2}|^2 < \infty \right\}.$$

For specific choices of Ω, this space might be defined more naturally using a weight function $\zeta(\lambda_j)^s$ for some function $\zeta : (0,\infty) \to (0,\infty)$, or by indexing the eigenvalues with another countable set, such as $\mathbb{N}^d$ or $\mathbb{Z}^d$. Nevertheless, this general formulation captures the essential structure of the space. To avoid such indexing issue, one can define this space more

implicitly as

$$\mathcal{H}^s(\Omega) := \Big\{ f \in L^2(\Omega) \,|\, (\mathbf{I} - \Delta)^{s/2} f \in L^2(\Omega) \Big\}.$$

The operator $(\mathbf{I} - \Delta)^{s/2}$ is called Bessel potential, and this space is also referred to as Bessel potential space. It is important to note that, unlike in the case of the torus, the equivalence between this Sobolev-type space and the classical Sobolev space defined using differential operators does not generally hold for arbitrary domains Ω. However, in applied operator learning, sample functions are typically generated using their spectral representation. Thus, we argue that the spectral definition of smoothness is arguably more natural from a practical perspective than the one based on differentials.

To construct the estimator from Section 9.3.1, given a sample budget of n, one queries the first n eigenfunctions $\phi_1, \phi_2, \ldots, \phi_n$ instead of Fourier modes. The proofs then remain valid without modification, as they only rely on the fact that Fourier modes form an orthonormal basis of $L^2(\mathbb{T}^d)$. However, because the eigenvalues are indexed by $\mathbb{N}$ in $\mathcal{H}^s(\Omega)$, the parameter s in this setting serves as an analog of s/d in $\mathcal{H}^s(\mathbb{T}^d)$.

In the experiments presented in Section 9.6, we also consider the case where Ω is a sphere and use spherical harmonics, which are the eigenfunctions of the Laplacian on a sphere. Similarly, Bessel functions serve as the Laplacian eigenfunctions in cylindrical domains. However, for general domains Ω, explicit eigenfunctions may not be available in closed form. In these cases, alternative bases such as orthonormal polynomials or wavelets, which are defined algebraically rather than through an eigenvalue problem, may be used. These bases form a complete system for sufficiently regular Ω' that contains Ω. A Sobolev-type space can then be defined using these algebraic bases and retain the theoretical guarantees established in this work.

H.2.1 Comparison to Function Generation in Applied Literature

Next, we discuss how our assumption that the initial wave ψ_0 lies in $\mathcal{H}^s(\Omega)$ is implicit in the function generation strategies commonly used in applied operator learning. In the applied literature, input functions are typically sampled from a Gaussian measure, $\mathrm{N}(0, (-\Delta + I)^{-\beta})$, or through some elementary push-forward of this distribution. This distribution, widely used in the applied stochastic PDE literature [Lord et al., 2014], was first introduced in the operator learning setting by Bhattacharya et al. [2021] and has since been implemented in works such as [Li et al., 2021, Kovachki et al., 2023].

Let $(\lambda_j, \phi_j)_{j=1}^{\infty}$ be the eigenpairs of $-\Delta$ in Ω with the given boundary conditions. By the Spectral Mapping Theorem, the eigenvalues of the covariance operator $(-\Delta + I)^{-\beta}$ are $(\lambda_j + 1)^{-\beta}$, while the eigenfunctions remain ϕ_j's. Applying the Karhunen-Loève Theorem

[Hsing and Eubank, 2015, Theorem 7.3.5], a sample $u \sim \text{Normal}(0, (-\Delta + I)^{-\beta})$ drawn from this distribution has the decomposition

$$u(x) = \sum_{j=1}^{\infty} (\lambda_j + 1)^{-\beta/2} \, \xi_j \, \phi_j(x),$$

where $\{\xi_j\}_{j=1}^{\infty}$ are uncorrelated standard Gaussian random variables, meaning $\xi_j \sim \text{Normal}(0, 1)$ and $\mathbb{E}[\xi_i \xi_j] = \mathbf{1}[i = j]$. Thus, sampling u is reduced to generating a sequence of independent Gaussian random variables $(\xi_j)_{j=1}^{\infty}$. In practical implementations, this is done by truncating the sequence to $(\xi_j)_{j=1}^{M}$.

Using this decomposition, we compute

$$\mathbb{E}[|\langle u, \phi_j \rangle_{L^2}|^2] = (\lambda_j + 1)^{-\beta} \, \mathbb{E}[|\xi_j|^2] = (\lambda_j + 1)^{-\beta}.$$

Thus, for any $s > 0$, we have

$$\mathbb{E}\left[\sum_{j=1}^{\infty} (1 + \lambda_j)^s |\langle u, \phi_j \rangle_{L^2}|^2\right] = \sum_{j=1}^{\infty} (1 + \lambda_j)^s (\lambda_j + 1)^{-\beta} = \sum_{j=1}^{\infty} (\lambda_j + 1)^{s-\beta}.$$

Recall that $\lambda_j \to \infty$ as $j \to \infty$. Thus, for large enough $\beta > 0$, there always exists $s < \beta$ such that $\sum_{j=1}^{\infty} (\lambda_j + 1)^{s-\beta} < \infty$. This shows that β controls the expected smoothness of samples from this distribution. More importantly, on average, these sampled functions belong to a Sobolev-type space. For a more detailed discussion of how β and s relate when $\Omega = \mathbb{T}^d$, we refer the reader to [Subedi and Tewari, 2025a, Section B.1].

A similar strategy is used in Lu et al. [2021], where functions are sampled from a Gaussian process with a covariance kernel given by the radial basis function (RBF) kernel, $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$. Since the eigenvalues of the RBF kernel decay exponentially fast, a similar analysis shows that the sampled functions belong to an extremely smooth space, essentially corresponding to the limiting case $s = \infty$. This argument is not specific to the RBF kernel– any kernel with sufficiently fast eigenvalue decay produces functions with high regularity.

Therefore, the requirement that input wave functions ψ belong to a smooth space $\mathcal{H}^s(\Omega)$ is both reasonable and consistent with what is often an implicit assumption in applied operator learning literature.

H.3 Proof of Proposition 5

Proof. Let $u, v \in L^2(\Omega)$. Expanding the inner product,

$$\langle \widehat{\mathrm{F}}_n(u), \widehat{\mathrm{F}}_n(v)\rangle_{L^2} = \left\langle \sum_{|k|_\infty \le K_n} w_k \langle u, \varphi_k\rangle_{L^2}, \sum_{|k|_\infty \le K_n} w_k \langle v, \varphi_k\rangle_{L^2} \right\rangle_{L^2}$$
$$= \sum_{|k|_\infty, |\ell|_\infty \le K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_\ell\rangle}_{L^2} \, \langle w_k, w_\ell\rangle_{L^2}.$$

Using the assumption on the PDE solver, we have

$$\langle w_k, w_\ell\rangle_{L^2} = \langle \mathrm{P}(\varphi_k), \mathrm{P}(\varphi_\ell)\rangle_{L^2} = \langle \varphi_k, \varphi_\ell\rangle_{L^2} = \mathbf{1}[k = \ell].$$

Substituting this into the sum,

$$\langle \widehat{\mathrm{F}}_n(u), \widehat{\mathrm{F}}_n(v)\rangle_{L^2} = \sum_{|k|_\infty, |\ell|_\infty \le K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_\ell\rangle}_{L^2} \, \mathbf{1}[k = \ell]$$
$$= \sum_{|k|_\infty \le K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_k\rangle}_{L^2}.$$

Using Parseval's identity, we can rewrite this as

$$\sum_{|k|_\infty \le K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_k\rangle}_{L^2} = \sum_{k \in \mathbb{Z}^d} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_k\rangle}_{L^2} - \sum_{|k|_\infty > K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_k\rangle}_{L^2}$$
$$= \langle u, v\rangle_{L^2} - \sum_{|k|_\infty > K_n} \langle u, \varphi_k\rangle_{L^2} \, \overline{\langle v, \varphi_k\rangle}_{L^2}.$$

Property (i) follows since the second summation vanishes when u, v belong to the span of $\{\varphi_k : k \in \mathbb{Z}^d, |k|_\infty \le K_n\}$. To establish property (ii), setting $u = v$ in the above expression,

$$\|\widehat{\mathrm{F}}_n(u)\|_{L^2}^2 = \|u\|_{L^2}^2 - \sum_{|k|_\infty > K_n} |\langle u, \varphi_k\rangle_{L^2}|^2 \le \|u\|_{L^2}^2.$$

This completes the proof. ■

H.4 Proof of Theorem 32

Proof. Recall that

$$\widehat{\mathrm{F}}_n = \sum_{|k|_\infty \le K_n} w_k \otimes \varphi_k.$$

For each k such that $|k|_\infty \leq K_n$, define the error term

$$\delta_k := w_k - \mathrm{F}(\varphi_k).$$

By Assumption 5, it follows that $\|\delta_k\|_{L^2} \leq \varepsilon$. So, for any wave function $\psi \in \mathcal{H}^s(\mathbb{T}^d)$, we can expand

$$\begin{aligned}
\widehat{\mathrm{F}}_n(\psi) &= \sum_{|k|_\infty \leq K_n} w_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \\
&= \sum_{|k|_\infty \leq K_n} \mathrm{F}(\varphi_k) \left\langle \psi, \varphi_k \right\rangle_{L^2} + \sum_{|k|_\infty \leq K_n} \delta_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \\
&= \mathrm{F}\left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \right) + \sum_{|k|_\infty \leq K_n} \delta_k \left\langle \psi, \varphi_k \right\rangle_{L^2},
\end{aligned}$$

where the last equality follows from the linearity of F. Then, applying the triangle inequality,

$$\begin{aligned}
\left\| \widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi) \right\|_{L^2} &= \left\| \mathrm{F}\left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \right) + \sum_{|k|_\infty \leq K_n} \delta_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \mathrm{F}(\psi) \right\|_{L^2} \\
&= \left\| \mathrm{F}\left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \psi \right) + \sum_{|k|_\infty \leq K_n} \delta_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2} \\
&\leq \left\| \mathrm{F}\left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \psi \right) \right\|_{L^2} + \left\| \sum_{|k|_\infty \leq K_n} \delta_k \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2},
\end{aligned}$$

To bound the first term, note that the operator norm of F from a L^2 to L^2 is 1 as F is a unitary operator. So,

$$\begin{aligned}
\left\| \mathrm{F}\left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \psi \right) \right\|_{L^2} &\leq \left\| \sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \psi \right\|_{L^2} \\
&= \sqrt{\sum_{|k|_\infty > K_n} |\left\langle \psi, \varphi_k \right\rangle_{L^2}|^2} \\
&= \sqrt{\sum_{|k|_\infty > K_n} \frac{(1+|k|_2^2)^s}{(1+|k|_2^2)^s} |\left\langle \psi, \varphi_k \right\rangle_{L^2}|^2} \\
&\leq \sqrt{\frac{1}{(1+K_n^2)^s}} \sqrt{\sum_{|k|_\infty > K_n} (1+|k|_2^2)^s |\left\langle \psi, \varphi_k \right\rangle_{L^2}|^2} \\
&\leq K_n^{-s} \|\psi\|_{\mathcal{H}^s}.
\end{aligned}$$

The first equality follows from Parseval's identity.

To bound the contribution from the PDE solver error, we use triangle inequality to get

$$\begin{aligned}
\left\| \sum_{|k|_\infty \le K_n} \delta_k \, \langle \psi, \varphi_k \rangle_{L^2} \right\|_{L^2} &\le \sum_{|k|_\infty \le K_n} \|\delta_k\|_{L^2} \, | \langle \psi, \varphi_k \rangle_{L^2} | \\
&\le \varepsilon \sum_{|k|_\infty \le K_n} | \langle \psi, \varphi_k \rangle_{L^2} | \\
&\le \varepsilon \sum_{|k|_\infty \le K_n} \sqrt{\frac{(1+|k|_2^2)^s}{(1+|k|_2^2)^s}} \, | \langle \psi, \varphi_k \rangle_{L^2} | \\
&\le \varepsilon \sqrt{\sum_{|k|_\infty \le K_n} \frac{1}{(1+|k|_2^2)^s}} \sqrt{\sum_{|k|_\infty \le K_n} (1+|k|_2^2)^s \, | \langle \psi, \varphi_k \rangle_{L^2} |^2} \\
&\le \varepsilon \, \|\psi\|_{\mathcal{H}^s} \sqrt{\sum_{|k|_\infty \le K_n} \frac{1}{(1+|k|_2^2)^s}} \\
&= \varepsilon \, \|\psi\|_{\mathcal{H}^s} \, \gamma_n,
\end{aligned}$$

where

$$\gamma_n := \sqrt{\sum_{|k|_\infty \le K_n} \frac{1}{(1+|k|_2^2)^s}}.$$

Thus, we have established that

$$\left\| \widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi) \right\|_{L^2} \le \|\psi\|_{\mathcal{H}^s} \left(\varepsilon \, \gamma_n + K_n^{-s} \right).$$

Note that $K_n = (n^{1/d} - 1)/2 \ge n^{1/d}/3$ as long as $n \ge 3^d$. This yields that $K_n^{-s} \le 3^s \, n^{-s/d}$.

To bound γ_n, recall that $|\{k \in \mathbb{Z}^d \, : \, |k|_\infty = j\}| = 2(2j+1)^{d-1}$. This is because one of the entry of m has to be $\pm j$ and other $d-1$ entries could be anything in $\{-j \ldots, -1, 0, 1, \ldots, j\}$.

Thus,

$$\begin{aligned}
\gamma_n^2 = \sum_{|k|_\infty \le K_n} \frac{1}{(1+|k|_2^2)^s} \le \sum_{|k|_\infty \le K_n} \frac{1}{(1+|k|_\infty^2)^s} &\le 1 + \sum_{1<|k|_\infty \le K_n} \frac{1}{(1+|k|_\infty^2)^s} \\
&\le 1 + \sum_{j=1}^{K_n} \frac{2(2j+1)^{d-1}}{(1+j^2)^s} \\
&\le 1 + \sum_{j=1}^{K_n} \frac{2(2j+1)^{d-1}}{j^{2s}} \\
&\le 1 + 2 \cdot 3^{d-1} \sum_{j=1}^{K_n} \frac{1}{j^{2s-d+1}} \\
&\lesssim \int_1^{K_n} \frac{1}{t^{2s-d+1}}\, dt \\
&\lesssim \begin{cases} 1, & \text{if } 2s > d, \\ \log(K_n), & \text{if } 2s = d, \\ K_n^{d-2s} & \text{if } 2s < d. \end{cases}
\end{aligned}$$

Our proof completes upon noting that $K_n \lesssim n^{1/d}$. ■

H.5 Proof of Theorem 33

Proof. We will break down the proof into multiple steps.

Defining the Hamiltonian: Considering the case where the potential is zero, meaning $V(x) = 0$ for all $x \in \mathbb{T}^d$. In this case, the Hamiltonian simplifies to

$$\mathrm{H} = -\frac{\hbar^2}{2m}\Delta.$$

The corresponding solution operator is given by

$$\mathrm{F} = \exp\left(-\frac{i}{\hbar} T\,\mathrm{H}\right) = \exp\left(i\frac{\hbar T}{2m}\Delta\right).$$

Note that the Fourier modes are eigenfunctions of this operator. Specifically, for any wave vector k, we have

$$\mathrm{F}(\varphi_k) = \exp\left(-i\frac{4\pi^2|k|_2^2\hbar T}{2m}\right)\varphi_k.$$

This follows from expanding the operator exponential and using the property

$$\Delta\varphi_k = \sum_{j=1}^{d} \frac{\partial^2 e^{2\pi i k\cdot x}}{\partial x_j^2} = \sum_{j=1}^{d} (2\pi i)^2 k_j^2 e^{2\pi i k\cdot x} = -4\pi^2 |k|_2^2 \varphi_k.$$

Defining

$$\eta_k := \exp\left(-i\frac{4\pi^2|k|_2^2\hbar T}{2m}\right),$$

we can express the action of the solution operator as

$$\mathrm{F}(\varphi_k) = \eta_k \varphi_k.$$

Specifying a PDE Solver. Our next step is to specify the exact details of the PDE solver that satisfies the ε-approximate assumption while also allowing us to construct hard instances to establish the lower bound. To that end, let P be the PDE solver defined as

$$\mathrm{P}(u) = \mathrm{F}(u) + \varepsilon\varphi_0, \quad \text{for every } u \in L^2(\Omega).$$

Here, φ_0 is simply the constant function 1 on the domain. That is, our PDE solver oracle returns the true solution shifted by $\varepsilon\varphi_0$ noise. While such a PDE solver is not practical, it is still a valid ε-approximate oracle since $\|\varphi_0\|_{L^2} = 1$, and thus our upper bound in Theorem 32 applies.

Writing out the Estimator. For this solution operator and the PDE solver specified above, our estimator has a more concrete form. In particular, we can write

$$\widehat{\mathrm{F}} := \sum_{|k|_\infty \le K_n} \eta_k \, \varphi_k \otimes \varphi_k + \varepsilon \sum_{|k|_\infty \le K_n} \varphi_0 \otimes \varphi_k.$$

Defining the Test Function. Given a sample size budget of n, we now construct a hard test wave function ψ_{test} to establish the claimed lower bound. To do this, choose a large $M \gg n$, which will be specified later, and define ψ_{test} as

$$\psi_{\text{test}} = \sum_{|k|_\infty \le M} c_k \varphi_k,$$

for some coefficients $c_k \ge 0$.

Establishing the Lower Bound. For the wave function ψ_{test} defined above, the true evolution under F is given by

$$\mathrm{F}(\psi_{\text{test}}) = \sum_{|k|_\infty \le M} c_k \, \eta_k \, \varphi_k.$$

On the other hand, the estimator $\widehat{\mathrm{F}}_n$ produces

$$\widehat{\mathrm{F}}_n(\psi_{\text{test}}) = \sum_{|k|_\infty \le K_n} c_k \, \eta_k \, \varphi_k + \varepsilon \sum_{|k|_\infty \le K_n} c_k \, \varphi_0.$$

Rewriting the second term,

$$\widehat{\mathrm{F}}_n(\psi_{\text{test}}) = \sum_{|k|_\infty \le K_n} c_k \, \eta_k \, \varphi_k + \varepsilon \left(\sum_{|k|_\infty \le K_n} c_k \right) \varphi_0.$$

Thus, the difference between the estimated and true evolution is

$$\widehat{\mathrm{F}}_n(\psi_{\text{test}}) - \mathrm{F}(\psi_{\text{test}}) = \varepsilon \left(\sum_{|k|_\infty \le K_n} c_k \right) \varphi_0 - \sum_{K_n < |k|_\infty \le M} c_k \, \eta_k \, \varphi_k.$$

Using Parseval's identity, we obtain

$$\begin{aligned} \left\| \widehat{\mathrm{F}}_n(\psi_{\text{test}}) - \mathrm{F}(\psi_{\text{test}}) \right\|_{L^2}^2 &= \varepsilon^2 \left| \sum_{|k|_\infty \le K_n} c_k \right| + \sum_{K_n < |k|_\infty \le M} |c_k \eta_k|^2 \\ &= \varepsilon^2 \sum_{|k|_\infty \le K_n} |c_k| + \sum_{K_n < |k|_\infty \le M} |c_k|^2, \end{aligned}$$

where the last step follows from the assumptions that $c_k \ge 0$ and $|\eta_k| = 1$ for all $k \in \mathbb{Z}^d$. To establish the claimed rate, we now choose c_k appropriately while ensuring that $\|\psi_{\text{test}}\| = 1$ and $\|\psi_{\text{test}}\|_{\mathcal{H}^s} \le 2$. To that end, we fix an index ℓ such that $|\ell|_\infty = \lceil K_n + 1 \rceil$ and set

$$c_\ell = \frac{1}{(1 + |\ell|_2^2)^{s/2}}.$$

This ensures that

$$\sum_{K_n < |k|_\infty \le M} |c_k|^2 \ge |c_\ell|^2 = \frac{1}{(1 + |\ell|_2^2)^s} \gtrsim \frac{1}{K_n^{2s}} \gtrsim \frac{1}{n^{2s/d}}.$$

It now remains to bound $\sum_{|k|_\infty \le K_n} |c_k|$, for which we proceed with a case analysis.

Case (I): $s \geq d/2$. In this case, we can simply set

$$c_0 := \sqrt{1 - c_\ell^2}$$

and assign $c_k = 0$ for all $k \notin \{0, \ell\}$. It is straightforward to verify that

$$\|\psi_{\text{test}}\|_{L^2}^2 = (1 - c_\ell^2) + c_\ell^2 = 1,$$

and

$$\|\psi_{\text{test}}\|_{\mathcal{H}^s}^2 = c_0^2 + (1 + |\ell|_2^2)^s c_\ell^2 = 1 - c_\ell^2 + 1 \leq 2.$$

Thus, we obtain

$$\sum_{|k|_\infty \leq K_n} |c_k| \geq |c_0| = \sqrt{1 - c_\ell^2} \geq \frac{1}{\sqrt{2}},$$

since $c_\ell \leq 1/2$ for sufficiently large K_n.

Case (II): $s < d/2$. Define

$$R_n := \sum_{0 < |k|_\infty \leq K_n} 1.$$

Now, set

$$c_k := \sqrt{\frac{1}{R_n(1 + K_n^{2s})}}$$

for all $0 < |k|_\infty \leq K_n$, and let $c_k = 0$ for all $|k|_\infty > K_n$ such that $k \neq \ell$. It follows that

$$\begin{aligned} \|\psi_{\text{test}}\|_{L^2}^2 &= |c_0|^2 + \sum_{|k|_\infty \leq K_n} \frac{1}{R_n(1 + K_n^2)^s} + \frac{1}{(1 + |\ell|_2^2)^s} \\ &= |c_0|^2 + \frac{1}{(1 + K_n^2)^s} + \frac{1}{(1 + |\lceil K_n + 1 \rceil|_2^2)^s}. \end{aligned}$$

For sufficiently large n, the second and third terms can each be made at most $1/3$, allowing $1 \geq c_0 > 0$ to be chosen appropriately so that $\|\psi_{\text{test}}\|_{L^2} = 1$ and ψ_{test} is a valid wave function.

Next, note that

$$\begin{aligned}\|\psi_{\text{test}}\|_{\mathcal{H}^s}^2 &= |c_0|^2 + \sum_{0<|k|_\infty \le K_n} \frac{(1+|k|_2^2)^s}{R_n(1+K_n^2)^s} + (1+|\ell|_2^2)^s c_\ell^2 \\ &\le |c_0|^2 + \sum_{0<|k|_\infty \le K_n} \frac{1}{R_n} + 1 \\ &\le 3.\end{aligned}$$

Thus, the sum is

$$\sum_{|k|_\infty \le K_n} |c_k| \ge \sum_{0<|k|_\infty \le K_n} \sqrt{\frac{1}{R_n(1+K_n^2)^s}} = \sqrt{\frac{1}{R_n(1+K_n^2)^s}} \cdot R_n = \sqrt{\frac{R_n}{(1+K_n^2)^s}}.$$

Since $R_n \gtrsim K_n^d$, we obtain

$$\sum_{|k|_\infty \le K_n} |c_k| \gtrsim \sqrt{K_n^{d-2s}} \gtrsim \sqrt{n^{1-\frac{2s}{d}}} = n^{\frac{1}{2}-\frac{s}{d}}.$$

Thus, we have established the lower bound of

$$\gtrsim \begin{cases} \varepsilon^2 + n^{-2s/d}, & \text{if } 2s \ge d, \\ \varepsilon^2\ n^{\frac{1}{2}-\frac{s}{d}} + n^{-2s/d} & \text{if } 2s < d. \end{cases}.$$

However, this is a lower bound for the squared norm. Using the inequality $\sqrt{a^2+b^2} \ge \frac{1}{\sqrt{2}}(|a|+|b|)$ completes our proof. ■

H.6 Refined Upper Bound Under Stronger Assumptions on PDE Solver.

Proof of Theorem 34. Our proof here largely follow the proof of Theorem 32 provided in Appendix H.4. Recall that, for any wave function ψ, we established in the proof of Theorem 32 that

$$\left\|\widehat{\mathrm{F}}_n(\psi) - \mathrm{F}(\psi)\right\|_{L^2} \le \left\|\mathrm{F}\left(\sum_{|k|_\infty \le K_n} \varphi_k \langle \psi, \varphi_k \rangle_{L^2} - \psi\right)\right\|_{L^2} + \left\|\sum_{|k|_\infty \le K_n} \delta_k \ \langle \psi, \varphi_k \rangle_{L^2}\right\|_{L^2}.$$

The first term does not have any randomness. So, following the same argument as in that proof, we can show that

$$\left\| \mathrm{F} \left(\sum_{|k|_\infty \leq K_n} \varphi_k \left\langle \psi, \varphi_k \right\rangle_{L^2} - \psi \right) \right\|_{L^2} \leq K_n^{-s} \ \|\psi\|_{\mathcal{H}^s} \leq 3^s \, c \, n^{-\frac{s}{d}}.$$

Here, we used the definition of K_n and the fact that $\|\psi\|_{\mathcal{H}^s} \leq c$. Now, it remains to bound the term with δ_k's. Since this is a random variable, we want to bound its expectation. To that end, Jensen's inequality implies

$$\mathbb{E}\left[\left\| \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2} \right] \leq \sqrt{\mathbb{E}\left[\left\| \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2}^2 \right]}.$$

Note that

$$\begin{aligned}
\left\| \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2}^2 &= \left\langle \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2}, \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\rangle_{L^2} \\
&= \sum_{|k|_\infty, |\ell|_\infty \leq K_n} \left\langle \psi, \varphi_k \right\rangle_{L^2} \overline{\left\langle \psi, \varphi_\ell \right\rangle_{L^2}} \left\langle \delta_k, \delta_\ell \right\rangle_{L^2} \\
&= \sum_{|k|_\infty \leq K_n} |\left\langle \psi, \varphi_k \right\rangle_{L^2}|^2 \|\delta_k\|_{L^2}^2 + \sum_{k \neq \ell} \left\langle \psi, \varphi_k \right\rangle_{L^2} \overline{\left\langle \psi, \varphi_\ell \right\rangle_{L^2}} \left\langle \delta_k, \delta_\ell \right\rangle
\end{aligned}$$

Note that the cross terms $k \neq \ell$ vanishes in expectation due to part (ii) of Assumption 6. Using part (i) of Assumption 6 yields

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2}^2 \right] &= \sum_{|k|_\infty \leq K_n} |\left\langle \psi, \varphi_k \right\rangle_{L^2}|^2 \, \mathbb{E}[\|\delta_k\|_{L^2}^2] \\
&\leq \varepsilon^2 \sum_{|k|_\infty \leq K_n} \left\langle \psi, \varphi_k \right\rangle_{L^2} \\
&\leq \varepsilon^2,
\end{aligned}$$

where the final step uses the fact that $\|\psi\|_{L^2}^2 = 1$. This shows that

$$\mathbb{E}\left[\left\| \sum_{|k|_\infty \leq K_n} \delta_k \ \left\langle \psi, \varphi_k \right\rangle_{L^2} \right\|_{L^2} \right] \leq \varepsilon.$$

This completes our proof. ■

H.7 Proof of Theorem 35

Proof. Note that

$$\widehat{\mathrm{F}}_n^q \psi - \mathrm{F}^q \, \psi = (\widehat{\mathrm{F}}_n^q - \mathrm{F}^q)\psi = \sum_{j=0}^{q-1} \widehat{\mathrm{F}}_n^{q-1-j} (\widehat{\mathrm{F}}_n - \mathrm{F}) \, \mathrm{F}^j \, \psi.$$

Applying the triangle inequality,

$$\|\widehat{\mathrm{F}}_n^q \psi - \mathrm{F}^q \, \psi\|_{L^2} \leq \sum_{j=0}^{q-1} \left\| \widehat{\mathrm{F}}_n^{q-1-j} (\widehat{\mathrm{F}}_n - \mathrm{F}) \, \mathrm{F}^j \, \psi \right\|_{L^2}.$$

Using property (ii) of Proposition 5 iteratively $q-j-1$ times, we obtain

$$\left\| \widehat{\mathrm{F}}_n^{q-1-j} (\widehat{\mathrm{F}}_n - \mathrm{F}) \, \mathrm{F}^j \, \psi \right\|_{L^2} \leq \left\| (\widehat{\mathrm{F}}_n - \mathrm{F}) \, \mathrm{F}^j \, \psi \right\|_{L^2}.$$

Furthermore, applying Theorem 32, we obtain the bound

$$\left\| (\widehat{\mathrm{F}}_n - \mathrm{F}) \, \mathrm{F}^j \, \psi \right\|_{L^2} \leq \| \, \mathrm{F}^j \, \psi \|_{\mathcal{H}^s} \left(\varepsilon \gamma_n + 3^s n^{-s/d} \right).$$

Thus, we conclude that

$$\|\widehat{\mathrm{F}}_n^q \psi - \mathrm{F}^q \, \psi\|_{L^2} \leq \left(\varepsilon \gamma_n + 3^s n^{-s/d} \right) \sum_{j=0}^{q-1} \| \, \mathrm{F}^j \, \psi \|_{\mathcal{H}^s}.$$

■

H.8 Proof of Corollary 4

H.8.1 Proof of Part (i)

Proof. Let $V(x) = a$ for all $x \in \mathbb{T}^d$. Then, for every Fourier mode φ_k, the Hamiltonian acts as

$$\mathrm{H} \, \varphi_k = \left(-\frac{\hbar^2}{2m} \Delta + V(\cdot) \right) \varphi_k = \left(\frac{\hbar^2}{2m} \, 4\pi^2 |k|_2^2 + a \right) \varphi_k.$$

The second equality holds because φ_k is an eigenfunction of $-\Delta$ with eigenvalue $4\pi^2|k|_2^2$. Next, applying the time evolution operator, we get

$$\mathrm{F}(\varphi_k) = e^{-\frac{\mathrm{i}}{\hbar} T \, \mathrm{H}} \varphi_k = e^{-\frac{\mathrm{i}}{\hbar} T \left(\frac{\hbar^2}{2m} \, 4\pi^2 |k|_2^2 + a \right)} \varphi_k.$$

Since the modulus of the complex exponential factor is always one, we can use this identity to establish that

$$\begin{aligned}
\|\,\mathrm{F}(\psi)\|_{\mathcal{H}^s} &= \sqrt{\sum_{k\in\mathbb{Z}^d} (1+|k|_2^2)^s\,|\langle \mathrm{F}(\psi),\varphi_k\rangle_{L^2}|^2} \\
&= \sqrt{\sum_{k\in\mathbb{Z}^d} (1+|k|_2^2)^s \left|\left\langle \sum_{\ell\in\mathbb{Z}^d} \langle\psi,\varphi_\ell\rangle\,\mathrm{F}(\varphi_\ell),\varphi_k\right\rangle_{L^2}\right|^2} \\
&= \sqrt{\sum_{k\in\mathbb{Z}^d} (1+|k|_2^2)^s \left|\left\langle \sum_{\ell\in\mathbb{Z}^d} \langle\psi,\varphi_\ell\rangle\, e^{-\frac{\mathrm{i}}{\hbar}T\left(\frac{\hbar^2}{2m}4\pi^2|\ell|_2^2+a\right)}\varphi_\ell,\varphi_k\right\rangle_{L^2}\right|^2} \\
&= \sqrt{\sum_{k\in\mathbb{Z}^d} (1+|k|_2^2)^s\,|\,\langle\psi,\varphi_k\rangle\,|^2},
\end{aligned}$$

where the final equality uses the fact that $\langle\varphi_\ell,\varphi_k\rangle = \mathbb{1}[k=\ell]$ and $\left|e^{-\frac{\mathrm{i}}{\hbar}T\left(\frac{\hbar^2}{2m}4\pi^2|k|_2^2+a\right)}\right| = 1$. Applying this iteratively for j steps, we obtain $\|\,\mathrm{F}^j(\psi)\|_{\mathcal{H}^s} = \|\psi\|_{\mathcal{H}^s}$ for all $j\in\mathbb{N}$. ■

H.8.2 Proof Part (ii)

Proof. Our result follows directly from the bound in [Delort, 2010, Theorem 1], originally established by Bourgain [1999], which states that

$$\left\|\mathrm{F}^j\,\psi\right\|_{\mathcal{H}^s} = \|\psi(\cdot,jT)\|_{\mathcal{H}^s} \le c\,(1+jT)\;\|\psi\|_{\mathcal{H}^s}\,.$$

This can be further refined using [Delort, 2010, Equation 1.3], yielding the bound

$$\left\|\mathrm{F}^j\,\psi\right\|_{\mathcal{H}^s} \le c\,(1+jT)^{\varepsilon}\;\|\psi\|_{\mathcal{H}^s}$$

for any fixed $\varepsilon>0$. Substituting this into our generalization bound gives

$$\left\|\widehat{\mathrm{F}}_n^q(\psi)-\mathrm{F}^q(\psi)\right\|_{L^2} \le \|\psi\|_{\mathcal{H}^s}\left(\varepsilon\gamma_n + 3^s n^{-s/d}\right)\cdot c\,q(1+Tq)^{\varepsilon}.$$

■

H.8.3 Proof Part (iii)

Proof. Since the Hamiltonian H is time-independent, the evolution operator satisfies

$$\mathrm{F}^j\,\psi = e^{-ijT\,\mathrm{H}/\hbar}\psi.$$

Defining $\psi(t)$ as the wave function at time t with initial condition $\psi(0) = \psi$, we write

$$\mathrm{F}^j \, \psi = \psi(jT).$$

Thus, bounding the Sobolev norm of $\mathrm{F}^j \, \psi$ reduces to bounding $\|\psi(t)\|_{\mathcal{H}^s}$ in terms of $\|\psi(0)\|_{\mathcal{H}^s}$ for all $t > 0$. To proceed, define the operator

$$\Lambda^s := \left(\mathrm{I} - (4\pi^2)^{-1} \, \Delta\right)^{s/2}.$$

Note that

$$\begin{aligned} \|\Lambda^s \psi\|_{L^2}^2 = \left\| \sum_{k \in \mathbb{Z}^d} \langle \psi, \varphi_k \rangle_{L^2} \, \Lambda^s \varphi_k \right\|_{L^2}^2 &= \left\| \sum_{k \in \mathbb{Z}^d} \langle \psi, \varphi_k \rangle_{L^2} \, (1 + |k|_2^2)^{s/2} \varphi_k \right\|_{L^2}^2 \\ &= \sum_{k \in \mathbb{Z}^d} (1 + |k|_2^2)^s |\, \langle \psi, \varphi_k \rangle_{L^2} \,|^2 \\ &= \|\psi\|_{\mathcal{H}^s}^2 . \end{aligned}$$

Thus, we focus on bounding $\|\Lambda^s \psi(t)\|_{L^2}$.

Energy Functional. Define the energy functional

$$E_s(t) := \|\Lambda^s \psi(t)\|_{L^2}^2.$$

Using the product rule rule in a Hilbert space, we obtain

$$\frac{d}{dt} E_s(t) = \langle \Lambda^s(\partial_t \psi), \Lambda^s \psi \rangle_{L^2} + \langle \Lambda^s \psi, \Lambda^s(\partial_t \psi) \rangle_{L^2} = 2 \, \mathrm{Re} \left(\langle \Lambda^s(\partial_t \psi), \Lambda^s \psi \rangle_{L^2} \right).$$

Since the Schrödinger equation states

$$\partial_t \psi = \mathrm{i} \, \frac{\hbar}{2m} \Delta \psi - \frac{\mathrm{i}}{\hbar} V \psi,$$

applying Λ^s to both sides yields

$$\Lambda^s(\partial_t \psi) = \mathrm{i} \, \frac{\hbar}{2m} \Lambda^s(\Delta \psi) - \frac{\mathrm{i}}{\hbar} \Lambda^s(V \psi).$$

Thus, we obtain the energy functional equation

$$\frac{d}{dt} E_s(t) = 2 \, \mathrm{Re} \left(\left\langle \mathrm{i} \, \frac{\hbar}{2m} \Lambda^s(\Delta \psi) - \frac{\mathrm{i}}{\hbar} \Lambda^s(V \psi), \Lambda^s \psi \right\rangle_{L^2} \right).$$

Note that

$$\operatorname{Re}\left\langle \mathrm{i}\,\frac{\hbar}{2m}\Lambda^s(\Delta\psi), \Lambda^s\psi \right\rangle = \operatorname{Re}\left(\mathrm{i}\,\frac{\hbar}{2m}\,\langle \Lambda^s(\Delta\psi), \Lambda^s\psi\rangle \right) = 0.$$

This follows because $\langle \Lambda^s(\Delta\psi), \Lambda^s\psi\rangle$ is a real number. To see why, observe that we can rewrite

$$\langle \Lambda^s(\Delta\psi), \Lambda^s\psi\rangle = \langle (\Lambda^s\Delta\Lambda^{-s})\Lambda^s\psi, \Lambda^s\psi\rangle.$$

Since $(\Lambda^s\Delta\Lambda^{-s})$ is a self-adjoint operator on L^2, the inner product must be real. So, the only contribution comes from

$$-\frac{\mathrm{i}}{\hbar}\Lambda^s(V\psi).$$

Thus, we obtain

$$\frac{d}{dt}E_s(t) = -\frac{2}{\hbar}\operatorname{Im}\langle \Lambda^s(V\psi), \Lambda^s\psi\rangle_{L^2}.$$

Applying the Cauchy–Schwarz inequality,

$$\left|\frac{d}{dt}E_s(t)\right| \leq \frac{2}{\hbar}\|\Lambda^s(V\psi)\|_{L^2}\|\Lambda^s\psi\|_{L^2} = \frac{2}{\hbar}\|\Lambda^s(V\psi)\|_{L^2}\sqrt{E_s(t)}.$$

Bounding the Sobolev Norm of $V\psi$. By assumption, V belongs to $\mathcal{H}^r(\mathbb{T}^d)$. We will now establish

$$\|V\psi\|_{\mathcal{H}^s} \leq a\|V\|_{\mathcal{H}^r}\|\psi\|_{\mathcal{H}^s}$$

for some universal $a > 0$ that only depends on s, d, r. This is a Hölder-type inequality for the Sobolev norm of a product of two functions, commonly known as a Sobolev multiplication inequality. This inequality is established in the proof of [Behzadan and Holst, 2021, Theorem 5.1] for the domain $\mathbb{R}^d$ (take $p_1 = p_2 = p = 2$, $s_1 = r$, and $s_2 = s$). The proof works verbatim for $\mathbb{T}^d$ as it only uses the Sobolev Embedding Theorems, which continue to hold on $\mathbb{T}^d$.

Thus, rewriting in terms of Λ^s yields

$$\|\Lambda^s(V\psi)\|_{L^2} \leq a\|V\|_{\mathcal{H}^r}\|\psi\|_{\mathcal{H}^s},$$

which upon using the definition of energy functional implies

$$\|\Lambda^s(V\psi)\|_{L^2} \leq a\|V\|_{\mathcal{H}^r}\sqrt{E_s(t)}.$$

Applying Grönwall's Inequality. Substituting this inequality into our bound for $\frac{d}{dt}E_s(t)$, we obtain

$$\left|\frac{d}{dt}E_s(t)\right| \leq \frac{2a}{\hbar}\|V\|_{\mathcal{H}^r}E_s(t).$$

Applying Grönwall's inequality on $[0,t]$, we obtain

$$E_s(t) \leq E_s(0)\exp\left(\frac{2a}{\hbar}\|V\|_{\mathcal{H}^r}\cdot t\right).$$

Finally, using the equivalence of norms,

$$\|\psi(t)\|_{\mathcal{H}^s}^2 \leq \|\psi(0)\|_{\mathcal{H}^s}^2\exp\left(\frac{2a}{\hbar}\|V\|_{\mathcal{H}^r}\cdot t\right).$$

Taking square roots on both sides and defining $c := \frac{2a}{\hbar}$, we conclude

$$\|\psi(t)\|_{\mathcal{H}^s} \leq \|\psi(0)\|_{\mathcal{H}^s}\exp\left(c\,\|V\|_{\mathcal{H}^r}\cdot t\right).$$

■

H.9 Experimental Potentials Details

We here provided more detailed descriptions of the potentials studied in the main text.

Free Particle If a particle is not exposed to an external potential, $V(x) = 0$ for all $x \in \Omega$.

Harmonic Oscillator Molecular vibrations are naturally modeled with a potential $V(x) = \frac{1}{2}m\,\omega^2\,|x|_2^2$, where m is the particle mass and ω the angular frequency of the oscillation.

Double Slit For a particle traveling across a barrier of potential V_0 at $x = x_0$ with two slits centered at y_1 and y_2 each with width w that are sufficiently far apart such that $|y_1 - y_2| \gg w$, the system potential is given by $V(x,y) = V_0$ when $x = x_0$ and $|y - y_1| > \frac{w}{2}$ and $|y - y_2| > \frac{w}{2}$, whereas $V(x,y) = 0$ otherwise.

Random Potentials To demonstrate robustness over arbitrary smooth potentials, a random potential $V(x)$ was drawn from a Gaussian Random Field identically to how such draws were made to define initial conditions, with $\alpha = 1$, $\beta = 1$, and $\gamma = 4$.

Coloumb Potential For a particle exposed to a radially symmetric electric field, such as in a Hydrogen atom, the potential is given by $V(x) = -\frac{ke^2}{r^2}$. We specifically focus on the case of a fixed radius of $r = 1$, for which the system can modeled as a uniform field in spherical coordinates. As discussed, both the pseudospectral solver and estimator were computed using spherical harmonics for this setup.

Paul Trap for Qubit Design A Paul trap is a device that confines charged particles, such as ions, using oscillating electric fields. Notably, therefore, such a potential is time-dependent. For a detailed mathematical treatment of the Paul trap, see [Major et al., 2005, Chapter 2]. A broader discussion on how Paul traps are used to localize charged ions for qubit encoding in their energy states can be found in the review article by [Bernardini et al., 2023]. In 2D, the potential function is given by $V(x, y, t) = \frac{U_0 + V_0 \cos(\omega t)}{r_0^2}(x^2 + y^2)$.

Shaken Lattice Optical lattices are a common design pattern for trapping neutral atoms with laser interferometry Deutsch et al. [2000]. The promise in certain applications, such as quantum computing, is subsequent manipulation of such trapped atoms Zhang et al. [2006]. One mechanism of control is known as "shaking," in which the phase of the potentials is manipulated to affect the momenta of the trapped particles Zheng and Zhai [2014], Kiely et al. [2016], Weidner et al. [2017]. If the shaking is restricted to a single axis, the potential is then given by $V(x, y, t) = V_0 \cos[k(x - A\sin(\omega t))] + V_0 \cos(ky)$.

Pulsed Gaussian Recent works have begun investigating the stability of bound states under pulsed external potentials, such as that of deuterons as studied in Rais et al. [2022]. In particular, stability was assessed in the presence of external Gaussian pulses, given by the potential $V(x, y, t) = V_0 \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right) \sum_{t_0} e^{-\frac{(t-t_0)^2}{2\sigma_t^2}}$.

We further provide the choices of parameters used for the experiments in H.1.

Table H.1: Parameter values used in the implementation of each potential.

Potential Name	**Parameter Values**
Free Particle	—
Barrier	$V_0 = 50.0$, $w = 0.2$
Harmonic Oscillator	$m = 1.0$, $\omega = 2.0$
Random Field (GRF)	$\alpha = 1, \beta = 1, \gamma = 4$
Paul Trap	$U_0 = 10.0$, $V_0 = 15.0$, $\omega = 3.0$, $r_0 = 2.0$
Shaken Lattice	$V_0 = 4.0$, $k_{\text{lat}} = 4\pi$, $A = 0.08$, $\omega_{\text{sh}} = 15.0$
Gaussian Pulse	$V = 100.0$, $x_0 = 0.0$, $y_0 = 0.0$, $\sigma_x = \sigma_y = 1.2$, $\sigma_t = 1.0$, $t_{\text{centers}} = \{0.0\}$
Coulomb	$k = 1.0$, $e = 1.0$
Coulomb Dipole	$V_0 = 1.0$

H.10 Experiment Results Over Noise Levels

We below present the additional results to accompany those presented in 9.6.

Table H.2: Average relative errors across different Hamiltonians for a relative noise level of 0.01%, computed over 100 i.i.d. test samples, with standard deviations in parentheses. Note that, for the Coulomb and dipole potential, the FNO columns instead refer to SFNO models. Dashes for DeepONet and UNO indicate that they do not handle functions on a spherical domain.

	FNO	UNO	DeepONet	Linear
Barrier	5.111e-02 (2.543e-02)	3.160e-02 (1.611e-02)	1.661e-01 (8.485e-02)	**1.955e-04 (5.894e-05)**
Coulomb	4.746e-02 (9.560e-03)	—	—	**1.550e-04 (6.582e-06)**
Dipole	4.362e-02 (9.411e-03)	—	—	**1.549e-04 (7.103e-06)**
Free	1.995e-02 (1.001e-02)	1.848e-02 (6.146e-03)	1.306e-01 (7.648e-02)	**1.904e-04 (3.065e-05)**
Gaussian Pulse	5.024e-02 (2.948e-02)	4.531e-02 (2.219e-02)	2.284e-01 (1.022e-01)	**2.012e-04 (6.267e-05)**
Harmonic Oscillator	6.899e-02 (3.591e-02)	5.559e-02 (1.578e-02)	1.544e-01 (8.427e-02)	**1.954e-04 (4.911e-05)**
Paul Trap	1.294e-01 (6.423e-02)	8.915e-02 (2.669e-02)	5.267e-01 (6.064e-02)	**1.982e-04 (5.174e-05)**
Random	2.526e-02 (1.576e-02)	7.334e-02 (1.656e-02)	3.048e-01 (1.081e-01)	**1.962e-04 (5.490e-05)**
Shaken Lattice	7.083e-02 (3.959e-03)	1.148e-02 (4.810e-03)	2.384e-01 (9.088e-02)	**1.921e-04 (4.245e-05)**

Table H.3: Average relative errors across different Hamiltonians for a relative noise level of 1.0%, computed over 100 i.i.d. test samples, with standard deviations in parentheses. Note that, for the Coulomb and dipole potential, the FNO columns instead refer to SFNO models. Dashes for DeepONet and UNO indicate that they do not handle functions on a spherical domain.

	FNO	UNO	DeepONet	Linear
Barrier	5.634e-02 (2.388e-02)	2.879e-02 (9.208e-03)	1.54e-01 (6.566e-02)	**1.591e-02 (1.321e-04)**
Coulomb	5.463e-02 (1.069e-02)	—	—	**1.462e-02 (1.673e-04)**
Dipole	7.903e-02 (2.250e-02)	—	—	**1.458e-02 (1.609e-04)**
Free	5.906e-02 (3.533e-02)	2.294e-02 (6.852e-03)	1.401e-01 (8.664e-02)	**1.591e-02 (1.281e-04)**
Gaussian Pulse	6.296e-02 (2.55e-02)	3.595e-02 (1.105e-02)	2.547e-01 (7.022e-02)	**1.594e-02 (1.272e-04)**
Harmonic Oscillator	3.560e-02 (1.146e-02)	3.66e-02 (1.233e-02)	4.123e-01 (8.527e-02)	**1.592e-02 (1.467e-04)**
Paul Trap	1.12e-01 (4.197e-02)	9.392e-02 (2.574e-02)	6.134e-01 (9.994e-02)	**1.592e-02 (1.369e-04)**
Random	1.924e-02 (5.013e-03)	2.66e-02 (6.725e-03)	2.005e-01 (7.011e-02)	**1.591e-02 (1.318e-04)**
Shaken Lattice	7.168e-02 (3.097e-03)	2.905e-02 (8.274e-03)	2.090e-01 (9.046e-02)	**1.590e-02 (1.297e-04)**

H.11 Additional Experiments for Partial Observation

Here we present the additional results for partial observation under a masking probability of 20% to accompany those presented in 9.6.3.

Table H.4: Average relative errors across different Hamiltonians for a masking probability of 20%, computed over 100 i.i.d. test samples, with standard deviations in parentheses. Note that, for the Coulomb and dipole potential, the FNO columns instead refer to SFNO models. Dashes for DeepONet and UNO indicate that they do not handle functions on a spherical domain.

	FNO	UNO	DeepONet	Linear
Barrier	2.282e-01 (3.934e-02)	1.634e-01 (4.013e-02)	4.574e-01 (1.088e-01)	**1.596e-03 (1.508e-05)**
Coulomb	2.483e-01 (4.045e-02)	—	—	**1.463e-03 (1.643e-05)**
Dipole	2.224e-01 (4.358e-02)	—	—	**1.464e-03 (1.736e-05)**
Free	1.767e-01 (2.721e-02)	1.522e-01 (2.136e-02)	4.242e-01 (1.127e-01)	**1.597e-03 (1.673e-05)**
Gaussian Pulse	2.796e-01 (5.107e-02)	5.404e-01 (7.378e-02)	3.879e-01 (1.053e-01)	**1.598e-03 (1.965e-05)**
Harmonic Oscillator	2.020e-01 (5.281e-02)	2.082e-01 (3.244e-02)	4.182e-01 (6.639e-02)	**1.595e-03 (1.416e-05)**
Paul Trap	2.826e-01 (5.004e-02)	1.000e+00 (8.714e-04)	6.165e-01 (4.276e-02)	**1.594e-03 (1.458e-05)**
Random	2.074e-01 (3.237e-02)	1.889e-01 (3.114e-02)	3.61e-01 (9.507e-02)	**4.640e-03 (1.095e-04)**
Shaken Lattice	1.224e-01 (1.065e-02)	1.468e-01 (1.739e-02)	3.066e-01 (1.020e-01)	**1.596e-03 (1.711e-05)**

H.12 Additional Experiment for Time Generalization

In Table H.5, we present additional results for time generalization under a relative noise of 1% to accompany those presented in 9.6.4.

Table H.5: Average relative time-generalization errors across different Hamiltonians for a relative noise level of 1%, computed over 100 i.i.d. test samples. Standard deviations are shown in parentheses.

Hamiltonian	$j = 1$	$j = 2$	$j = 4$	$j = 8$	$j = 16$
Barrier	1.590e-02 (1.215e-04)	2.412e-02 (2.718e-03)	2.429e-02 (2.911e-03)	2.326e-02 (2.419e-03)	2.288e-02 (2.404e-03)
Coulomb	1.461e-02 (1.574e-04)	1.464e-02 (1.744e-04)	1.459e-02 (1.559e-04)	1.460e-02 (1.726e-04)	1.458e-02 (1.681e-04)
Dipole	1.461e-02 (1.716e-04)	1.465e-02 (1.662e-04)	1.461e-02 (1.646e-04)	1.460e-02 (1.553e-04)	1.466e-02 (1.725e-04)
Free	1.593e-02 (1.493e-04)	1.592e-02 (1.395e-04)	1.593e-02 (1.294e-04)	1.594e-02 (1.196e-04)	1.590e-02 (1.300e-04)
Gaussian Pulse	1.592e-02 (1.097e-04)	2.245e-02 (3.847e-03)	2.435e-02 (5.065e-03)	2.466e-02 (5.202e-03)	2.563e-02 (5.738e-03)
Harmonic Oscillator	1.593e-02 (1.380e-04)	1.592e-02 (1.248e-04)	1.585e-02 (1.322e-04)	1.581e-02 (1.371e-04)	1.576e-02 (1.182e-04)
Paul Trap	1.593e-02 (1.178e-04)	1.519e-01 (2.011e-02)	4.538e-01 (4.887e-02)	6.345e-01 (4.912e-02)	6.671e-01 (4.591e-02)
Random Lattice	1.590e-02 (1.320e-04)	1.590e-02 (1.349e-04)	1.589e-02 (1.376e-04)	1.591e-02 (1.377e-04)	1.588e-02 (1.407e-04)
Shaken Lattice	1.591e-02 (1.359e-04)	1.677e-02 (2.702e-04)	1.677e-02 (2.449e-04)	1.627e-02 (1.478e-04)	1.703e-02 (2.380e-04)

H.13 Additional Experimental Results

We present in H.6 the results for the case where the test functions have a spectrum to match that of the linear estimator, i.e. where $|k|_\infty \leq K_n$. Noise was disabled for this test, i.e. $\sigma = 0$ was used in the data generation.

Table H.6: Average relative errors across Hamiltonians assessed over a batch of 100 i.i.d. test samples with a restricted spectrum.

	FNO	UNO	DeepONet	Linear
Barrier	4.231e-02 (1.711e-02)	6.767e-02 (3.416e-02)	2.616e-01 (7.999e-02)	**3.085e-15 (2.219e-16)**
Coulomb	8.033e-02 (1.926e-02)	—	—	**4.639e-05 (1.447e-05)**
Dipole	5.362e-02 (1.438e-02)	—	—	**4.660e-05 (1.706e-05)**
Free	1.45e-02 (7.094e-03)	2.167e-02 (1.021e-02)	1.962e-01 (8.468e-02)	**2.57e-15 (1.713e-16)**
Gaussian Pulse	5.412e-02 (3.173e-02)	1.062e-01 (6.663e-02)	2.788e-01 (8.848e-02)	**3.129e-15 (1.119e-16)**
Harmonic Oscillator	3.48e-02 (1.887e-02)	4.439e-02 (1.78e-02)	1.855e-01 (7.109e-02)	**3.068e-15 (1.028e-16)**
Paul Trap	1.264e-01 (5.449e-02)	5.246e-02 (2.538e-02)	6.663e-01 (9.779e-02)	**3.394e-15 (4.389e-17)**
Random	1.483e-02 (6.859e-03)	5.872e-02 (2.138e-02)	2.093e-01 (6.123e-02)	**3.014e-15 (1.708e-16)**
Shaken Lattice	7.036e-02 (2.764e-03)	1.17e-02 (4.428e-03)	2.031e-01 (6.624e-02)	**3.055e-15 (1.347e-16)**

H.14 Compute Resources

All experiments involving the linear estimator were run on a standard-grade CPU. The deep learning-based approaches, namely the FNO and DeepONet, were trained on an Nvidia RTX 2080 Ti.

BIBLIOGRAPHY

Jacob Abernethy, Francis Bach, Theodoros Evgeniou, and Jean-Philippe Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(3), 2009.

Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4), 1997. ISSN 0004-5411. doi: 10.1145/263867.263927.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

Noga Alon, Zhihan Jin, and Benny Sudakov. The helly number of hamming balls and related problems. *arXiv preprint*, 2024.

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Dana Angluin. Queries and concept learning. *Machine learning*, 2:319–342, 1988.

Dana Angluin. Queries revisited. In *Algorithmic Learning Theory: 12th International Conference, ALT 2001 Washington, DC, USA, November 25–28, 2001 Proceedings 12*, pages 12–31. Springer, 2001.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511624216.

Patrick Assouad. Densité et dimension. In *Annales de l'institut Fourier*, volume 33, pages 233–282, 1983.

Les Atlas, David Cohn, and Richard Ladner. Training connectionist networks with queries and selective sampling. *Advances in neural information processing systems*, 2, 1989.

Idan Attias and Steve Hanneke. Adversarially robust pac learnability of real-valued functions. In *International Conference on Machine Learning*, pages 1172–1199. PMLR, 2023.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: Pac learning and online learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Yannick Augenstein, Taavi Repan, and Carsten Rockstuhl. Neural operator-based surrogate solver for free-form electromagnetic inverse design. *ACS Photonics*, 10(5):1547–1557, 2023.

Kamyar Azizzadenesheli. Neural operator learning. *International Conference on Machine Learning (Tutorial)*, 2024.

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine learning*, 43:211–246, 2001.

Christopher TH Baker and RL Taylor. The numerical treatment of integral equations. *Journal of Applied Mechanics*, 46(4):969, 1979.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.

Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1996.0033.

Francesca Bartolucci, Emmanuel de Bezenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. *Advances in Neural Information Processing Systems*, 36: 69661–69672, 2023.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Ali Behzadan and Michael Holst. Multiplication in sobolev spaces, revisited. *Arkiv för Matematik*, 59(2):275–306, 2021.

Mikhail Gennadievich Belov and Vladislav Gennadievich Malyshkin. Partially unitary learning. *Physical Review E*, 110(5):055306, 2024.

Shai Ben-David, Nicolò Cesa-Bianchi, and Philip M. Long. Characterizations of learnability for classes of $\{0, ..., n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1): 74–86, 1995.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48, 2019.

Francesco Bernardini, Abhijit Chakraborty, and Carlos R Ordóñez. Quantum computing with trapped ions: a beginner's guide. *European Journal of Physics*, 45(1):013001, 2023.

Rajendra Bhatia and John AR Holbrook. On the Clarkson-McCarthy inequalities. *Mathematische Annalen*, 281:7–12, 1988.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric pdes. *The SMAI journal of computational mathematics*, 7:121–157, 2021.

Alessandro Bisio, Giulio Chiribella, Giacomo Mauro D'Ariano, Stefano Facchini, and Paolo Perinotti. Optimal quantum learning of a unitary transformation. *Physical Review A—Atomic, Molecular, and Optical Physics*, 81(3):032324, 2010.

Moise Blanchard. Universal online learning: An optimistically universal learning rule. In *Conference on Learning Theory*, pages 1077–1125. PMLR, 2022.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023.

Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5):216–233, 2015.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Nicolas Boullé and Alex Townsend. A mathematical guide to operator learning. In *Handbook of Numerical Analysis*, volume 25, pages 83–125. Elsevier, 2024.

Nicolas Boullé, Christopher J Earls, and Alex Townsend. Data-driven discovery of green's functions with human-understandable deep learning. *Scientific reports*, 12(1):4824, 2022.

Nicolas Boullé, Diana Halikias, and Alex Townsend. Elliptic pde learning is provably data-efficient. *Proceedings of the National Academy of Sciences*, 120(39):e2303904120, 2023.

Jean Bourgain. On growth of sobolev norms in linear schrödinger equations with smooth time dependent potential. *Journal d'Analyse Mathématique*, 77(1):315–348, 1999.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 09–12 Jul 2020.

Philip J Brown and James V Zidek. Adaptive multivariate ridge regression. *The Annals of Statistics*, 8(1):64–74, 1980.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

Róbert Busa-Fekete, Heejin Choi, Krzysztof Dembczynski, Claudio Gentile, Henry Reeve, and Balazs Szorenyi. Regret bounds for multilabel classification in sparse label regimes. *Advances in Neural Information Processing Systems*, 35:5404–5416, 2022.

T. Ceccherini-Silberstein, M. Salvatori, and E. Sava-Huss. *Groups, Graphs and Random Walks*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2017. doi: 10.1017/9781316576571.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Lena Chekina, Dan Gutfreund, Aryeh Kontorovich, Lior Rokach, and Bracha Shapira. Exploiting label dependencies for improved sample complexity. *Machine learning*, 91:1–42, 2013.

Dong Chen, Peter Hall, and Hans-Georg Müller. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, pages 1720–1747, 2011.

Xiao Chen, Silas Hoffman, James N Fry, and Hai-Ping Cheng. Simulating decoherence of coupled two spin qubits using generalized cluster correlation expansion. *arXiv preprint arXiv:2402.18722*, 2024.

John B Conway. A course in functional analysis (1990). *Graduate Texts in Mathematics*, 1990.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. *Advances in Neural Information Processing Systems*, 29, 2016.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104. PMLR, 2013.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19, pages 207–232. PMLR, 2011.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Constantinos Daskalakis and Noah Golowich. Fast rates for nonparametric online learning: from realizability to learning in games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 846–859, 2022.

Maarten V de Hoop, Nikola B Kovachki, Nicholas H Nelsen, and Andrew M Stuart. Convergence rates for learning linear operators from noisy data. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):480–513, 2023.

Jean-Marc Delort. Growth of sobolev norms of solutions of linear schrödinger equations on some compact manifolds. *International Mathematics Research Notices*, 2010(12):2305–2328, 2010.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 280–295. Springer, 2010.

Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88:5–45, 2012.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Ivan H Deutsch, Gavin K Brennen, and Poul S Jessen. Quantum computing with neutral atoms in an optical lattice. *Fortschritte der Physik: Progress of Physics*, 48(9-11):925–943, 2000.

Phil Diamond. Least squares fitting of compact set-valued data. *Journal of Mathematical Analysis and Applications*, 147(2):351–362, 1990.

Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

Jürgen Eckhoff. Helly, Radon, and Carathéodory Type Theorems. In *Handbook of Convex Geometry*, pages 389–448. North-Holland, Amsterdam, 1993. ISBN 978-0-444-89596-7.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

SİNAN Eyi, JO Hager, and KD Lee. Airfoil design optimization using the navier-stokes equations. *Journal of Optimization Theory and Applications*, 83:447–461, 1994.

Frédéric Ferraty. *Nonparametric functional data analysis*. Springer, 2006.

Yuval Filmus, Steve Hanneke, Idan Mehalel, and Shay Moran. Optimal prediction using expert advice and randomized littlestone dimension. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 773–836. PMLR, 2023.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Dylan J. Foster and Alexander Rakhlin. ℓ_∞ vector contraction for rademacher complexity, 2019.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pages 341–358. JMLR Workshop and Conference Proceedings, 2011.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? on the discretization mismatch errors in neural operators. In *The Thirteenth International Conference on Learning Representations*, 2025.

Carl Friedrich Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Perthes et Besser, Hamburg, 1809. Translated as "Theory of Motion of the Heavenly Bodies Moving About the Sun in Conic Sections" by C. H. Davis, Little, Brown, Boston, 1857. Reprinted by Dover, New York, 1963.

Carl Friedrich Gauss. *Theoria Combinationis Observationum Erroribus Minimis Obnoxia*. Dieterich, Göttingen, 1823. Translated as "Theory of the Combination of Observations Least Subject to Errors" by G. W. Stewart, SIAM, Philadelphia, 1995.

Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Claudio Gentile and Francesco Orabona. On multilabel classification and ranking with partial feedback. *Advances in Neural Information Processing Systems*, 25, 2012.

María Angeles Gil, María Asunción Lubiano, Manuel Montenegro, and María Teresa López. Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika*, 56:97–111, 2002.

David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 1977.

Giorgio Gnecco and Marcello Sanguineti. Estimates of the approximation error using rademacher complexity: learning vector-valued functions. *Journal of Inequalities and Applications*, 2008:1–16, 2008.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Loukas Grafakos. *Classical fourier analysis*, volume 2. Springer, 2008.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. *Advances in neural information processing systems*, 34:24048–24062, 2021.

Steve Hanneke. A statistical theory of active learning. *Foundations and Trends in Machine Learning*, pages 1–212, 2013.

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

Steve Hanneke and Liu Yang. Bandit learnability can be undecidable. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5813–5849. PMLR, 2023.

Steve Hanneke, Roi Livni, and Shay Moran. Online learning with simple predictors and a combinatorial characterization of minimax in 0/1 games. In *Conference on Learning Theory*, pages 2289–2314. PMLR, 2021.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. *Proceedings of the 36th Annual Conference on Learning Theory (COLT)*, 2023.

Friedel Hartmann. *Green's functions and finite elements*. Springer Science & Business Media, 2012.

Ed Helly. Über mengen konvexer körper mit gemeinschaftlichen punkte. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923.

Jan Hermann, James Spencer, Kenny Choo, Antonio Mezzacapo, W Matthew C Foulkes, David Pfau, Giuseppe Carleo, and Frank Noé. Ab initio quantum chemistry with neural-network wavefunctions. *Nature Reviews Chemistry*, 7(10):692–709, 2023.

Max Hopkins, Daniel M. Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3015–3069. PMLR, 02–05 Jul 2022.

Siegfried Hörmann and Łukasz Kidziński. A note on estimation in hilbertian linear models. *Scandinavian journal of statistics*, 42(1):43–62, 2015.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Yi-Ming Huang, Xiao-Yu Li, Yi-Xuan Zhu, Hang Lei, Qing-Sheng Zhu, and Shan Yang. Learning unitary transformation by quantum machine learning model. *Computers, Materials & Continua*, 68(1), 2021.

Catherine Huber, Valentin Solev, and Filia Vonta. Interval censored and truncated data: Rate of convergence of npmle of the density. *Journal of Statistical Planning and Inference*, 139(5):1734–1749, 2009.

John K. Hunter. Notes on partial differential equations, 2023. URL https://www.math.ucdavis.edu/~hunter/pdes/pde_notes.pdf. Available at: https://www.math.ucdavis.edu/~hunter/pdes/pde_notes.pdf.

Stephanie Hyland and Gunnar Rätsch. Learning unitary operators with help from u (n). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944, 2016.

Peishi Jiang, Zhao Yang, Jiali Wang, Chenfu Huang, Pengfei Xue, TC Chakraborty, Xingyuan Chen, and Yun Qian. Efficient super-resolution of near-surface climate modeling using the fourier neural operator. *Journal of Advances in Modeling Earth Systems*, 15(7): e2023MS003800, 2023.

J Robert Johansson, Paul D Nation, and Franco Nori. Qutip: An open-source python framework for the dynamics of open quantum systems. *Computer physics communications*, 183(8):1760–1772, 2012.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Daniel Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. In *Conference on Learning Theory*, pages 1903–1943. PMLR, 2019.

Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 382–391. IEEE, 1990.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.

Johannes Kepler. *Harmonices Mundi*. 1619. English translation: J. V. Field, The Harmony of the World, American Philosophical Society, 1997.

Anthony Kiely, Albert Benseny, Thomas Busch, and Andreas Ruschhaupt. Shaken not stirred: creating exotic angular momentum states by shaking an optical lattice. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(21):215003, 2016.

Taeyoung Kim and Myungjoo Kang. Bounding the rademacher complexity of fourier neural operators. *Machine Learning*, 113(5):2467–2498, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andreas Kirsch. *An introduction to the mathematical theory of inverse problems*, volume 120. Springer, 2011.

Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416, 2020.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22 (290):1–76, 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Nikola B Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data complexity estimates for operator learning. *arXiv preprint arXiv:2405.15992*, 2024a.

Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024b.

Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. *Advances in Neural Information Processing Systems*, 28, 2015.

Samuel Lanthaler. Operator learning with pca-net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023.

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for deeponets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):tnac001, 2022.

Samuel Lanthaler, Andrew M Stuart, and Margaret Trautner. Discretization error of fourier neural operators. *arXiv preprint arXiv:2405.02221*, 2024.

Claude Leforestier, RH Bisseling, Charly Cerjan, MD Feit, Rich Friesner, A Guldberg, A Hammerich, G Jolicard, W Karrlein, H-D Meyer, et al. A comparison of different propagation schemes for the time dependent schrödinger equation. *Journal of Computational Physics*, 94(1):59–80, 1991.

Adrien-Marie Legendre. Sur la méthode des moindres carrés. In *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, 1805. Translated from French by H. A. Ruger and H. M. Walker.

DD Lewis and WA Gale. A sequential algorithmfor training text classifiers. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

Shibo Li, Xin Yu, Wei Xing, Robert Kirby, Akil Narayan, and Shandian Zhe. Multi-resolution active learning of fourier neural operators. In *International Conference on Artificial Intelligence and Statistics*, pages 2440–2448. PMLR, 2024a.

Xiaosong Li, Niranjan Govind, Christine Isborn, A Eugene DePrince III, and Kenneth Lopata. Real-time time-dependent electronic structure theory. *Chemical Reviews*, 120 (18):9951–9993, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020b.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*, 2021.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024b.

Chensen Lin, Zhen Li, Lu Lu, Shengze Cai, Martin Maxey, and George Em Karniadakis. Operator learning for predicting multiscale bubble growth dynamics. *The Journal of Chemical Physics*, 154(10), 2021.

HQ Lin, JE Gubernatis, Harvey Gould, and Jan Tobochnik. Exact diagonalization methods for quantum systems. *Computers in Physics*, 7(4):400–407, 1993.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1987.

Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research*, 25(24):1–67, 2024.

Hao Liu, Biraj Dahal, Rongjie Lai, and Wenjing Liao. Generalization error guaranteed auto-encoder-based nonlinear model reduction for operator learning. *Applied and Computational Harmonic Analysis*, 74:101717, 2025.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Weiwei Liu and Ivor Tsang. On the optimality of classifier chain for multi-label classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Wen-Hao Liu, Zhi Wang, Zhang-Hui Chen, Jun-Wei Luo, Shu-Shen Li, and Lin-Wang Wang. Algorithm advances and applications of time-dependent first-principles simulations for ultrafast dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(3):e1577, 2022.

Gabriel J Lord, Catherine E Powell, and Tony Shardlow. *An introduction to computational stochastic PDEs*, volume 50. Cambridge University Press, 2014.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

Xihaier Luo, Xiaoning Qian, and Byung-Jun Yoon. Hierarchical neural operator transformer with learnable frequency-aware loss prior for arbitrary-scale super-resolution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 33466–33485, 2024.

Fouad G. Major, Viorica N. Gheorghe, and Günther Werth. *Charged Particle Traps: Physics and Techniques of Charged Particle Field Confinement*. Springer, 2005. ISBN 978-3-540-40710-5. doi: 10.1007/b137836. URL https://link.springer.com/book/10.1007/b137836.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990): 80–85, 2023.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Kyle Mills, Michael Spanner, and Isaac Tamblyn. Deep learning and the schrödinger equation. *Physical Review A*, 96(4):042113, 2017.

Sebastian Mizera. Scattering with neural operators. *Physical Review D*, 108(10):L101701, 2023.

Wael W Mohammed, Naveed Iqbal, S Bourazza, and Elsayed M Elsayed. The optical structures for the fractional chiral nonlinear schrödinger equation with time-dependent coefficients. *Optical and Quantum Electronics*, 56(9):1476, 2024.

Mattes Mollenhauer, Nicole Mücke, and TJ Sullivan. Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem. *arXiv preprint arXiv:2211.08875*, 2022.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Shay Moran, Ohad Sharon, Iska Tsubari, and Sivan Yosebashvili. List online classification. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1885–1913. PMLR, 2023.

Daniel Musekamp, Marimuthu Kalimuthu, David Holzmüller, Makoto Takamoto, and Mathias Niepert. Active learning for neural pde solvers. *arXiv preprint arXiv:2408.01536*, 2024.

Chris Nagele, Oliver Janssen, and Matthew Kleban. Decoherence: a numerical study. *Journal of Physics A: Mathematical and Theoretical*, 56(8):085301, 2023.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

B. K. Natarajan. On learning sets and functions. *Mach. Learn.*, 4(1): 67–97, oct 1989a. ISSN 0885-6125. doi: 10.1023/A:1022605311895. URL https://doi.org/10.1023/A:1022605311895.

Balaubramaniam Kausik Natarajan. Some results on learning. Technical Report CMU-RI-TR-89-06, The Robotics Institute, Carnegie Mellon University, 1989b.

Nicholas H Nelsen and Andrew M Stuart. The random feature model for input-output maps between banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

Francisco Duarte Moura Neto and Antônio José da Silva Neto. *An introduction to inverse problems with applications*. Springer Science & Business Media, 2012.

Vasilis Niarchos and Constantinos Papageorgakis. Learning s-matrix phases with neural operators. *Physical Review D*, 110(4):045020, 2024.

Jannes Nys, Gabriel Pescia, Alessandro Sinibaldi, and Giuseppe Carleo. Ab-initio variational wave functions for the time-dependent many-electron schrödinger equation. *Nature communications*, 15(1):9404, 2024.

Junhyung Park and Krikamol Muandet. Towards empirical process theory for vector-valued functions: Metric entropy of smooth function classes. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, 2023.

Uri Peskin and Nimrod Moiseyev. The solution of the time-dependent schrödinger equation by the (t, t') method: Theory, computational algorithm and applications. *The Journal of chemical physics*, 99(6):4590–4596, 1993.

David Pollard. Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR, 1990.

Johann Radon. Mengen konvexer körper, die einen gemeinsamen punkt enthalten. *Mathematische Annalen*, 83(1-2):113–115, 1921.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022.

Jan Rais, Hendrik van Hees, and Carsten Greiner. Bound-state formation in time-dependent potentials. *Physical Review C*, 106(6):064004, 2022.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize : From value to algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015b.

Sasha Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. *Advances in Neural Information Processing Systems*, 25, 2012b.

Vinod Raman, Unique Subedi, and Ambuj Tewari. Online learning with set-valued feedback. In *Proceedings of Thirty Seventh Conference on Learning Theory*, 2024a.

Vinod Raman, Unique Subedi, and Ambuj Tewari. Online infinite-dimensional regression: Learning linear operators. In *International Conference on Algorithmic Learning Theory*, pages 1113–1133. PMLR, 2024b.

Vinod Raman, Unique Subedi, and Ambuj Tewari. A unified theory of supervised online learnability. In *36th International Conference on Algorithmic Learning Theory*, 2025.

C Radhakrishna Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3/4):447–458, 1965.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of pdes. *Advances in Neural Information Processing Systems*, 36:77187–77200, 2023.

Michael Reed and Barry Simon. *II: Fourier analysis, self-adjointness*, volume 2. Elsevier, 1975.

Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pages 8030–8040. PMLR, 2020.

Matthew Reimherr. Functional regression with repeated eigenvalues. *Statistics & Probability Letters*, 107:62–70, 2015.

Niklas Reinhardt, Sven Wang, and Jakob Zech. Statistical learning theory for neural operators. *arXiv:2412.17582*, 2024.

Jack Richter-Powell, Yaron Lipman, and Ricky TQ Chen. Neural conservation laws: A divergence-free perspective. *Advances in Neural Information Processing Systems*, 35: 38075–38088, 2022.

Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. *arXiv preprint arXiv:2510.06646*, 2025.

Jun John Sakurai and Jim Napolitano. *Modern quantum mechanics*. Cambridge University Press, 2020.

Venkataraman Santhanam, Vlad I Morariu, and Larry S Davis. Generalized deep image to image regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5609–5619, 2017.

Burr Settles. Active learning literature survey. 2009.

Freya Shah, Taylor L Patti, Julius Berner, Bahareh Tolooshams, Jean Kossaifi, and Anima Anandkumar. Fourier neural operators for learning dynamics in quantum spin systems. *arXiv preprint arXiv:2409.03302*, 2024.

Karan Shah, Patrick Stiller, Nico Hoffmann, and Attila Cangi. Physics-informed neural networks as solvers for the time-dependent schrodinger equation. *arXiv preprint arXiv:2210.12522*, 2022.

Ava Shahrokhi and Alireza Jahangirian. Airfoil shape parameterization for optimum navier–stokes design with genetic algorithm. *Aerospace science and technology*, 11(6):443–450, 2007.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

Saumya Sinha, Brandon Benton, and Patrick Emami. On the effectiveness of neural operators at zero-shot weather downscaling. *Environmental Data Science*, 4:e21, 2025.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. *Advances in neural information processing systems*, 24, 2011.

Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In *COLT*, pages 1–13. Citeseer, 2010.

J Michael Steele. Empirical discrepancies and subadditive processes. *The Annals of Probability*, pages 118–127, 1978.

George Stepaniants. Learning partial differential equations in reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 24(86):1–72, 2023.

Unique Subedi and Ambuj Tewari. On the benefits of active data collection in operator learning. *International Conference on Machine Learning (ICML)*, 2025a.

Unique Subedi and Ambuj Tewari. Controlling statistical, discretization, and truncation errors in learning fourier linear operators. *Transactions of Machine Learning Research (TMLR)*, 2025b.

Unique Subedi and Ambuj Tewari. Operator learning: A statistical perspective. *Annual Review of Statistics and Its Application*, 13, 2026.

Puoya Tabaghi, Maarten de Hoop, and Ivan Dokmanić. Learning schatten–von neumann operators. *arXiv preprint arXiv:1901.10076*, 2019.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

Yang Tang, Jürgen Kurths, Wei Lin, Edward Ott, and Ljupco Kocarev. Introduction to focus issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6), 2020.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

Michael E Taylor. *Partial Differential Equations I Basic Theory*. Springer, 2011.

The Nobel Committee. The Nobel Prize in Chemistry 2024, 2024. URL https://www.nobelprize.org/prizes/chemistry/2024/summary/. Accessed: 2025-02-09.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023.

Gunther Uhlmann. *Inside out: inverse problems and applications*, volume 47. Cambridge University Press, 2003.

Nobuyuki Umetani and Bernd Bickel. Learning three-dimensional flow for interactive aerodynamic design. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018.

Leslie G. Valiant. A theory of the learnable. In *Symposium on the Theory of Computing*, 1984.

Dirk van der Hoeven, Federico Fusco, and Nicolò Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. *Advances in Neural Information Processing Systems*, 34: 13280–13291, 2021.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer, 1996.

W. van Dijk and F. M. Toyama. Accurate numerical solutions of the time-dependent schrödinger equation. *Phys. Rev. E*, 75:036707, Mar 2007. doi: 10.1103/PhysRevE.75.036707. URL https://link.aps.org/doi/10.1103/PhysRevE.75.036707.

W van Dijk, J Brown, and K Spyksma. Efficiency and accuracy of numerical solutions to the time-dependent schrödinger equation. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(5):056703, 2011.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.

Vladimir N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. *Teoriya Raspoznavaniya Obrazov: Statisticheskie Problemy Obucheniya*. Nauka, Moscow, 1974. In Russian. English title: Theory of Pattern Recognition: Statistical Problems of Learning.

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

Vladimir Vovk, Harris Papadopoulos, and Alexander Gammerman. Measures of complexity. *Festschrift for Alexey*, 2015.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Haixin Wang, Yadi Cao, Zijie Huang, Yuxuan Liu, Peiyan Hu, Xiao Luo, Zezheng Song, Wanjia Zhao, Jilin Liu, Jinan Sun, et al. Recent advances on machine learning for computational fluid dynamics: A survey. *arXiv preprint arXiv:2408.12171*, 2024.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016.

J André C Weideman and Ben M Herbst. Split-step methods for the solution of the nonlinear schrödinger equation. *SIAM Journal on Numerical Analysis*, 23(3):485–507, 1986.

Joachim Weidmann. *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media, 2012.

CA Weidner, Hoon Yu, Ronnie Kosloff, and Dana Z Anderson. Atom interferometry using a shaken optical lattice. *Physical Review A*, 95(4):043624, 2017.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Ning Wu, Arun Nanduri, and Herschel Rabitz. Rabi oscillations, decoherence, and disentanglement in a qubit–spin-bath system. *Physical Review A*, 89(6):062105, 2014.

Shuo Xu, Xin An, Xiaodong Qiao, Lijun Zhu, and Lin Li. Multi-output least-squares support vector regression machines. *Pattern Recognition Letters*, 34(9):1078–1084, 2013.

Liang Yang, Xi-Zhu Wu, Yuan Jiang, and Zhi-Hua Zhou. Multi-label learning with deep forest. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 1634–1641. IOS Press, 2020.

Qidong Yang, Alex Hernandez-Garcia, Paula Harder, Venkatesh Ramesh, Prasanna Sattigeri, Daniela Szwarcman, Campbell D Watson, and David Rolnick. Fourier neural operators for arbitrary resolution climate data downscaling. *Journal of Machine Learning Research*, 25(420):1–30, 2024.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

Yuki Yasuda and Ryo Onishi. Zero-shot super-resolution from unstructured data using a transformer-based neural operator for urban micrometeorology. *arXiv preprint arXiv:2504.21361*, 2025.

Huaiqian You, Quinn Zhang, Colton J Ross, Chung-Hao Lee, Ming-Chen Hsu, and Yue Yu. A physics-guided neural operator learning approach to model biological tissues from digital image correlation measurements. *Journal of Biomechanical Engineering*, 144(12): 121012, 2022a.

Huaiqian You, Quinn Zhang, Colton J Ross, Chung-Hao Lee, and Yue Yu. Learning deep implicit fourier neural operators (ifnos) with applications to heterogeneous material modeling. *Computer Methods in Applied Mechanics and Engineering*, 398:115296, 2022b.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(38):1–47, 2018.

Chuanwei Zhang, SL Rolston, and S Das Sarma. Manipulation of single neutral atoms in optical lattices. *Physical Review A—Atomic, Molecular, and Optical Physics*, 74(4): 042316, 2006.

Jiaji Zhang, Carlos L Benavides-Riveros, and Lipeng Chen. Artificial-intelligence-based surrogate solution of dissipative quantum dynamics: physics-informed reconstruction of the universal propagator. *The Journal of Physical Chemistry Letters*, 15(13):3603–3610, 2024.

Jiaji Zhang, Carlos L Benavides-Riveros, and Lipeng Chen. Neural quantum propagators for driven-dissipative quantum dynamics. *Physical Review Research*, 7(1):L012013, 2025a.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, Keir Adams, Maurice Weiler, Xiner Li, Tianfan Fu, Yucheng Wang, Alex Strasser, Haiyang Yu, YuQing Xie, Xiang Fu, Shenglong Xu, Yi Liu, Yuanqi Du, Alexandra Saxton, Hongyi Ling, Hannah Lawrence, Hannes Stärk, Shurui Gui, Carl Edwards, Nicholas Gao, Adriana Ladera, Tailin Wu, Elyssa F. Hofgard, Aria Mansouri Tehrani, Rui Wang, Ameya Daigavane, Montgomery Bohde, Jerry Kurtin, Qian Huang, Tuong Phung, Minkai Xu, Chaitanya K. Joshi, Simon V. Mathis, Kamyar Azizzadenesheli, Ada Fang, Alán Aspuru-Guzik, Erik Bekkers, Michael Bronstein, Marinka Zitnik, Anima Anandkumar, Stefano Ermon, Pietro Liò, Rose Yu, Stephan Günnemann, Jure Leskovec, Heng Ji, Jimeng Sun, Regina Barzilay, Tommi Jaakkola, Connor W. Coley, Xiaoning Qian, Xiaofeng Qian, Tess Smidt, and Shuiwang Ji. Artificial intelligence

for science in quantum, atomistic, and continuum systems. *Foundations and Trends in Machine Learning*, 18(4):385–912, 2025b.

Wei Zheng and Hui Zhai. Floquet topological states in shaking optical lattices. *Physical Review A*, 89(6):061603, 2014.