# Online Infinite-Dimensional Regression: Learning Linear Operators

**Vinod Raman**[*] VKRAMAN@UMICH.EDU
**Unique Subedi**[*] SUBEDI@UMICH.EDU
**Ambuj Tewari** TEWARIA@UMICH.EDU
*University of Michigan*

**Editors:** Claire Vernade and Daniel Hsu

## Abstract

We consider the problem of learning linear operators under squared loss between two infinite-dimensional Hilbert spaces in the online setting. We show that the class of linear operators with uniformly bounded $p$-Schatten norm is online learnable for any $p \in [1, \infty)$. On the other hand, we prove an impossibility result by showing that the class of uniformly bounded linear operators with respect to the operator norm is *not* online learnable. Moreover, we show a separation between sequential uniform convergence and online learnability by identifying a class of bounded linear operators that is online learnable but uniform convergence does not hold. Finally, we prove that the impossibility result and the separation between uniform convergence and learnability also hold in the batch setting.

**Keywords:** Online Learnability, Linear Operators, Regression

## 1. Introduction

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., 2021; Li et al., 2020). Moreover, many of the traditional learning settings such as multitask 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 applications 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

[*] Equal Contribution

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 (1989, 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, 1989). However, since our loss is real-valued, showing that the class is learnable requires some novel algorithmic ideas, which can be of independent interest.

## 1.1. 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 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., 2021; Li et al., 2020). 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).

## 2. Preliminaries

### 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 $\ell_2$ 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 $\varphi_n$ and $\phi_n$. 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 $\ell_p$ 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 $\ell_\infty$ 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 $\infty$-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).

### 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 1 (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 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 2 (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*

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

*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 3 (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*

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

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

If there exists a $B > 0$ such that $\sup_{f,v,w} \|f(v) - w\|^2 \leq B$, 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 $\text{Rad}_T(\mathcal{F}) = o(T)$. Given this equivalence, in this work, we only rely on the sequential Rademacher complexity of $\mathcal{F}$ to study its sequential uniform convergence property.

## 3. $p$-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 4 match the lowerbounds in the batch settig established in Section 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 4 (Uniformly Bounded Subsets of $S_p(\mathcal{V}, \mathcal{W})$ are Online Learnable)** *Fix* $c > 0$. *Let* $\mathcal{X} = \{v \in \mathcal{V} \mid \|v\| \leq 1\}$ *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,*

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

Theorem 4 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 4 is vacuous. Indeed, in Section 4, we prove that any uniformly bounded subset of $S_\infty(\mathcal{V}, \mathcal{W})$ is not online learnable.

Our proof of Theorem 4 relies on Lemma 5 which shows that the $q$-Schatten norm of Rademacher sums of rank-1 operators concentrates for every $q \geq 1$. The proof of Lemma 5 is in Appendix A.

**Lemma 5 (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 5 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 $A \mapsto \text{tr}(h(F))$ is a convex functional on the set $\{F \in \mathcal{T} \mid \text{spectra}(F) \subseteq [\alpha, \beta]\}$ for any convex function $h$ and the class of finite-rank self-adjoint operators $\mathcal{T}$. Our proof of Lemma 5, on the other hand, only uses standard inequalities.

Equipped with Lemma 5, our proof of Theorem 4 follows by upper bounding the sequential Rademacher complexity of the loss class. Although this proof of online learnability is nonconstructive, we can use Proposition 1 from (Rakhlin et al., 2012) 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 4. 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 4 establishes a stronger guarantee– not only are these classes online learnable but they also enjoy sequential uniform convergence.

### 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. $\|\cdot\|_{\text{op}}$ 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 4 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 $\Omega$. 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 $(\Omega, \mu)$. 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 4 implies that $\mathcal{F}$ is online learnable with regret $6c^2\sqrt{T}$.

## 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 6 (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 6 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 6 matches the upperbound in Theorem 4 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 6) Fix an algorithm $\mathcal{A}$, and consider a labeled stream $\{(e_t, c\, \sigma_t \psi_t)\}_{t=1}^T$ where $\sigma_t \sim$ Unif$(\{-1, 1\})$. Then, the expected loss of $\mathcal{A}$ is

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

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. ■

### 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 7 (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 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 8 (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 8 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. 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 9 (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

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

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 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 sort$(S_i)$. We let sort$(S_i)[j]$ denote the $j^{th}$ element of the ordered list 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 $\sum_{t>t^\star} \left\|E_{t^\star}^{r^\star}(x_t) - y_t\right\|^2 = \sum_{t>t^\star} \|f_{k^\star}(x_t) - y_t\|^2$. Therefore, it suffices to show that

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

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

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

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

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

We prove the inequality above by establishing

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

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

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

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

### 5.1. Batch Learnability without Uniform Convergence

Although we state Theorem 9 in the online setting, an analogous result also holds in the batch setting. To establish the batch analog of Theorem 9, consider $f_k$ defined in (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 9. Recall that in our proof of Theorem 9 (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.

## 6. Discussion and Open Questions

In this work, we study the online learnability of bounded linear operators between two infinite-dimensional Hilbert spaces. In Theorems 4 and 6, 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.

What is $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}_p)$ 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 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.

Is $\mathrm{Rad}_T(\mathcal{F}) = o(T)$ if and only if $\inf_{\mathcal{A}} \mathrm{R}_{\mathcal{A}}(T, \mathcal{F}) = o(T)$ 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 9 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.

## Acknowledgments

We acknowledge the assistance of Judy McDonald in locating a misdelivered package containing Shalev-Shwartz and Ben-David (2014). Without the benefit of ideas in Shalev-Shwartz and Ben-David (2014), this paper would have never been written. AT acknowledges the support of NSF via grant IIS-2007055. VR acknowledges the support of the NSF Graduate Research Fellowship.

## References

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.

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.

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.

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.

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.

John B Conway. A course in functional analysis (1990). *Graduate Texts in Mathematics*, 1990.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 207–232, Budapest, Hungary, 09–11 Jun 2011. PMLR.

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.

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.

Frédéric Ferraty. *Nonparametric functional data analysis*. Springer, 2006.

Steve Hanneke, Shay Moran, Vinod Raman, Unique Subedi, and Ambuj Tewari. Multiclass online learning and uniform convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5682–5696. PMLR, 2023.

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, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021.

Samuel Lanthaler. Operator learning with pca-net: upper and lower complexity bounds. *arXiv preprint arXiv:2303.16317*, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

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.

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.

Balaubramaniam Kausik Natarajan. Some results on learning. Technical Report CMU-RI-TR-89-06, The Robotics Institute, Carnegie Mellon University, 1989.

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.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264. PMLR, 2014.

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, 2012.

Michael Reed and Barry Simon. *II: Fourier analysis, self-adjointness*, volume 2. Elsevier, 1975.

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.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

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.

Puoya Tabaghi, Maarten de Hoop, and Ivan Dokmanić. Learning schatten–von neumann operators. *arXiv preprint arXiv:1901.10076*, 2019.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

Gunther Uhlmann. *Inside out: inverse problems and applications*, volume 47. Cambridge University Press, 2003.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016.

Joachim Weidmann. *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media, 2012.

## Appendix A. Upperbound Proofs for Online Setting

Our proof of Theorem 4 also relies on the following technical Lemma.

**Lemma 10** *Let $v \in \mathcal{V}$, $w \in \mathcal{W}$, and $f \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. Then, we have $\langle f(v), w \rangle = \operatorname{tr}(f \circ (v \otimes w))$.*

**Proof** (of Lemma 10) 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 $\ell_2$ summable sequence $\{\alpha_n\}_{n\in\mathbb{N}}$. Then, by definition of the trace operator, we have

$$\operatorname{tr}(f \circ (v \otimes w)) = \sum_{n=1}^{\infty} \langle f \circ (v \otimes w)(\psi_n), \psi_n \rangle = \sum_{n=1}^{\infty} \langle \alpha_n \, f(v), \psi_n \rangle = \left\langle f(v), \sum_{n=1}^{\infty} \alpha_n \psi_n \right\rangle = \langle f(v), w \rangle \,,$$

which completes our proof. ■

### A.1. Proof of Lemma 5

Let $F = \sum_{t=1}^{T} \sigma_t \, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})$. Since

$$\text{rank}\,(F) \leq \sum_{t=1}^{T} \text{rank}\,(\sigma_t \, v_t(\sigma_{<t}) \otimes w_t(\sigma_{<t})) \leq T,$$

$F$ 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 5, 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 $\|F\|_2 = \sqrt{\text{tr}\,(F^\star F)}$. Using linearity of trace and Jensen's inequality gives $\mathbb{E}\left[\sqrt{\text{tr}\,(F^\star F)}\right] \leq \sqrt{\text{tr}\,(\mathbb{E}\,[F^\star F])}$. 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 $\sigma_r$ 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.

## A.2. Proof of Theorem 4

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

$$\begin{aligned}
\operatorname{Rad}_T(\mathcal{F}_p) &= \sup_{x,y} \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t})) - y_t(\sigma_{<t})\right\|^2\right] \\
&\leq \sup_{x,y} \left( \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t}))\right\|^2\right] + 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} -\sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle\right] \right. \\
&\qquad \left. + \mathbb{E}\left[\sum_{t=1}^{T} \sigma_t \left\|y_t(\sigma_{<t}))\right\|^2\right]\right) \\
&= \sup_{x,y} \left( \mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\|f(x_t(\sigma_{<t}))\right\|^2\right] + 2\,\mathbb{E}\left[\sup_{f \in \mathcal{F}_p} \sum_{t=1}^{T} \sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle\right]\right).
\end{aligned}$$

To handle the second term above, recall that Lemma 10 implies $\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\rangle = \operatorname{tr}(f \circ (x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t})))$. Using the linearity of the trace operator, we obtain

$$\sum_{t=1}^{T} \sigma_t \left\langle f(x_t(\sigma_{<t})), y_t(\sigma_{<t})\right\rangle = \operatorname{tr}\left(f \circ \sum_{t=1}^{T} \sigma_t\, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t})\right) \leq \|f\|_p \left\|\sum_{t=1}^{T} \sigma_t\, x_t(\sigma_{<t}) \otimes y_t(\sigma_{<t})\right\|_q,$$

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

$$\|f(x_t(\sigma_{<t}))\|^2 = \langle f(x_t(\sigma_{<t})), f(x_t(\sigma_{<t}))\rangle = \langle f^\star f(x_t(\sigma_{<t})), x_t(\sigma_{<t})\rangle = \mathrm{tr}(f^\star f\circ(x_t(\sigma_{<t})\otimes x_t(\sigma_{<t}))),$$

where the final equality follows from Lemma 10. 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

$$\mathrm{Rad}_T(\mathcal{F}_p) \leq c^2\,\mathbb{E}\left[\left\|\sum_{t=1}^{T}\sigma_t\, x_t(\sigma_{<t})\otimes x_t(\sigma_{<t})\right\|_q\right] + 2c\,\mathbb{E}\left[\left\|\sum_{t=1}^{T}\sigma_t\, x_t(\sigma_{<t})\otimes y_t(\sigma_{<t})\right\|_q\right] \leq 3c^2\,T^{\max\left\{\frac{1}{2},\frac{1}{q}\right\}},$$

where the final inequality follows from using Lemma 5 twice. Recalling that $\frac{1}{q} = 1 - \frac{1}{p}$ completes our proof of second inequality.

## Appendix B. Proof of Theorem 8

### B.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 $\sigma$ 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

$$\begin{aligned}
&\inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x \sim U^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm 1\}^{mn}}\left[\mathop{\mathbb{E}}_{x\sim U}\left[\left\|\hat{f}_n(x) - h_\sigma(x)\right\|^2\right]\right]\right] \\
&= \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x \sim U^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm 1\}^{mn}}\left[\frac{1}{mn}\sum_{i=1}^{mn}\left\|\hat{f}_n(e_i) - h_\sigma(e_i)\right\|^2\right]\right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N\sim \text{Unif}(\{1,,\ldots,mn\})^n}\left[\mathop{\mathbb{E}}_{\sigma\sim\{\pm 1\}^{mn}}\left[\frac{1}{mn}\sum_{i\notin N}\left\|\hat{f}_n(e_i) - c\,\sigma_i\psi_i\right\|^2\right]\right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N\sim \text{Unif}(\{1,,\ldots,mn\})^n}\left[\frac{1}{mn}\sum_{i\notin N}\left(\mathop{\mathbb{E}}_{\sigma\sim\{\pm 1\}^{mn}}\left[\left\|\hat{f}_n(e_i) - c\,\sigma_i\psi_i\right\|\right]\right)^2\right].
\end{aligned}$$

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma\sim\{\pm 1\}^{mn}}\left[\left\|\hat{f}_n(e_i) - c\,\sigma_i\psi_i\right\|\right] = \mathbb{E}\left[\mathop{\mathbb{E}}_{\sigma_i}\left[\left\|\hat{f}_n(e_i) - c\,\sigma_i\psi_i\right\|\right] \mid \hat{f}_n\right] \\
&= \mathbb{E}\left[\frac{1}{2}\left(\left\|\hat{f}_n(e_i) - c\,\psi_i\right\| + \left\|\hat{f}_n(e_i) + c\,\psi_i\right\|\right) \mid \hat{f}_n\right] \\
&\geq \frac{1}{2}\left\|c\psi_i + c\psi_i\right\| \\
&= c,
\end{aligned}$$

where we use the fact $\hat{f}_n$ is independent of $\sigma_i$ 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 $m = \lceil 2n^{\frac{1}{p-1}} \rceil$. 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. ■

### B.2. Proof of lowerbound of $\frac{c^2}{8}\,n^{-\frac{2}{p}}$.

Our proof here follows similar arguments as the proof in B.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 $\sigma$. 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

$$
\begin{aligned}
&\inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x \sim U^n}\left[\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{mn}}\left[\mathop{\mathbb{E}}_{x \sim U}\left[\left\|\hat{f}_n(x) - f_{\sigma,p}(x)\right\|^2\right]\right]\right] \\
&= \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{S_x \sim U^n}\left[\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{mn}}\left[\frac{1}{mn}\sum_{i=1}^{mn}\left\|\hat{f}_n(e_i) - f_{\sigma,p}(e_i)\right\|^2\right]\right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N \sim \text{Unif}(\{1,,\ldots,mn\})^n}\left[\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{mn}}\left[\frac{1}{mn}\sum_{i \notin N}\left\|\hat{f}_n(e_i) - \frac{c\,\sigma_i}{(mn)^{1/p}}\psi_i\right\|^2\right]\right] \\
&\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N \sim \text{Unif}(\{1,,\ldots,mn\})^n}\left[\frac{1}{mn}\sum_{i \notin N}\left(\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{mn}}\left[\left\|\hat{f}_n(e_i) - \frac{c\,\sigma_i}{(mn)^{1/p}}\psi_i\right\|\right]\right)^2\right].
\end{aligned}
$$

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 $\sigma_i$ as $i \notin N$ and triangle inequality. Thus, combining everything, our lower bound is

$$
\geq \inf_{\hat{f}_n} \mathop{\mathbb{E}}_{N \sim \text{Unif}(\{1,,\ldots,mn\})^n}\left[\frac{1}{mn}\sum_{i \notin N}\frac{c^2}{(mn)^{2/p}}\right] = \frac{c^2}{(mn)^{2/p}}\left(1 - \frac{1}{mn}\right)^n.
$$

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 $\frac{c^2}{8}n^{-\frac{2}{p}}$.

■