Conformally Robust Decision Making

by

Yash Patel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2025

Doctoral Committee:

Professor Ambuj Tewari, Chair
Associate Professor Yang Chen
Assistant Professor Jeffrey Regier
Assistant Professor Alexander Rodriguez

Yash Patel
yppatel@umich.edu
ORCID iD: 0000-0002-3221-1908

DEDICATION

This thesis is dedicated to the crazy ones. The misfits. The rebels. The troublemakers. The ones who think they can change the world.

ACKNOWLEDGMENTS

This thesis is the culmination of my last four years at the University of Michigan. Leaving industry to return to school after working for three years is not at all how I expected my life to play out all those years back when I walked through FitzRandolph Gate, undergraduate degree in hand. Yet, there I was, more excited to be back than I can possibly articulate. At the time, there was a question of whether this decision was the "right" one. Having reached the end, I can clearly look back and say that it was, and I would like to take some space to highlight the people who made it so.

I would first like to thank my advisor, Ambuj Tewari, for making the PhD experience such a truly exciting and engaging one. I remember first joining the group to work on sampling algorithms for intrinsically disordered proteins. Since then, my research trajectory has taken every left and right turn imaginable from that starting point, landing in the spaces of cryo-electron microscopy, robust decision-making, control theory, and operator learning. All throughout this meandering, meeting with Ambuj to talk about progress or lack thereof was an infectiously energizing experience: I am immensely grateful that I had an advisor that not only tolerated but was excited to navigate all these new areas with me. I really appreciate the atmosphere of excited curiosity he cultured both in me and the research group more broadly.

I would also like to the other members of my committee: Jeff Regier, Yang Chen, and Alex Rodriguez. I would like Jeff for having given me the opportunity to learn about simulation-based inference and exposing me to an area of research early on in my PhD when I was still deciding in which direction to head. I would also like to thank Yang for having taught STATS 605: this class opened my eyes to the broader world of methods of uncertainty estimation and robustness just as I was gravitating to that area as my research focus. Finally, I would like to thank Alex for having, however briefly, collaborated on ideas around decision making: I am sure a counterfactual world in which there exists more time would have seen this lead to many interesting projects.

I would next like to thank other members of the wonderful statistics department staff. I would like to thank Becca Usoff for being such an awesome department coordinator: no matter how bizarre a request I would send, whether it was about insurance, coordinating visit day, or scheduling my thesis, Becca would always have an answer ready. I would also like to thank Judy McDonald for making our department such a warm and welcoming place: organizing large events like Statsgiving and even smaller ones like the department picnic or ice cream social made the department feel like a home in a way that will feel special to me.

In addition, I would like to acknowledge all of the great friends I have made over my time at Michigan: you all made the PhD so much more memorable than it would have been with research alone. I would like to thank the whole CASI group for having made for very stimulating conversations about research and otherwise. I would especially like to thank Saptarshi, Unique, Vinod, and Chinmaya for spending countless hours with me, whether it was talking about the future of "AI for Science" or singing "I'll Make a Man Out of You." In addition to my academic pursuits, the gym has always been a special place to me, and I would like to acknowledge the eclectic group of friends I made either directly or indirectly from there: Pan, Andrew, Syed, Cooper, John Kenny, Abbas, and many others. The countless calories we consumed at Blue Nile and interesting conversations we shared over spicy peas will be permanently etched into my memory. I would also like to think my somehow-still-in-touch PARE group from all those years back, Max Jerdee and Ariana Bueno: getting to mooch off of Max's cooking and see other grad students outside of statistics was always a breath of fresh air. I would next like to thank my other friends from the department, Jake, Jaylin, Aaron, and everyone else: all the time we got to spend together, whether at potlucks or board game nights or just in the office, are memories I will always cherish. Finally, I would like to thank Sahana Rayan: from spending near endless time with me to being an ear for my constant blabbering about neural operators to being a genuinely excited cheerleader of my crazy dreams, I am eternally thankful for having you in my life.

Lastly, I would like to thank my family. To my parents, I appreciate you always having supported me, from all the whack job ideas projects I was doing in high school up through this PhD. My curiosity and desire to pursue this PhD ultimately stems from the environment you both nurtured for me growing up, and I am permanently indebted to you for that. I would also like to thank my younger sister, Aditi. Even though we are such different people, I always look forward to spending time together and have cherished all the trips we have managed to squeeze in over the last few years.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Black-box machine learning models are seeing increasing deployment in safety-critical settings, such as in autonomous vehicles and healthcare settings. This coupling increases the need to have reliable, post-hoc, distribution-free methods of uncertainty quantification. Among these is "conformal prediction," which replaces point predictions with "prediction regions," subsets of the output space with probabilistic guarantees of covering the truth. Despite such guarantees, these implicitly defined predictions regions do not immediately lend themselves to practical use. In this thesis, we propose and develop one such use: model-based decision-making.

We develop this conformal decision-making framework over three works. In the first, we focus on the development of conformal prediction in the space of scientific inquiry. Increasingly common in certain domains, such as astrophysics and neuroscience, is the use approximate variational inference to do posterior estimation for parameters. Unfortunately, there are few guarantees about the quality of these approximate posteriors. We propose Conformalized Amortized Neural Variational Inference (CANVI), a procedure that is scalable, provides guaranteed marginal coverage, and seeks maximal predictive efficiency.

In the next work, we generalize the scope to "predict-then-optimize" problems, where the decision-maker is forced to estimate the unknown parameters of a task and optimize their decision against this surrogate objective. In the nominal approach, the parameters predicted are assumed to precisely coincide with the true, unknown parameters, which can result in suboptimal decision-making. Towards this end, we develop a robust analog of this nominal problem formulation, called "Conformal Predict-Then-Optimize" (CPO), and demonstrate how such a robust formulation can be efficiently solved. In the third work, we demonstrate the generality of CPO, demonstrating its extensibility to recent works that proposed vector conformal scores and to model-based linear control problems.

CHAPTER 1

Introduction

Wonder is the beginning of wisdom.

Socrates

Machine learning algorithms are seeing increasingly widespread adoption in safety-critical settings, including in autonomous vehicles, medical applications, and general policy enforcement [Jenn et al., 2020, Gharib and Bondavalli, 2019, Tambon et al., 2022]. For this reason, the need for uncertainty quantification has become ever more apparent. Such uncertainty estimates can be subsequently used in several manners, from simply forcing the model to abstain from making a prediction [Hamid et al., 2017, Schuster, 2025] to having a human expert verify the model predictions in-the-loop [Mozannar and Sontag, 2020, Keswani et al., 2021] to making downstream decisions that design for worst-case scenarios from the upstream model [Gabrel et al., 2014].

Classically, such uncertainty estimates would be derived from either Bayesian statistics or large-sample asymptotic analyses [Bolstad and Curran, 2016, Bucher, 2009]. These methods, however, are becoming increasingly at odds with those models that are achieving state-of-the-art predictive performance. In particular, parametric models have been largely replaced by black-box models across many safety-critical settings, most prominently in the form of large language models (LLMs) and agentic workflows [Zhang et al., 2024, Liu et al., 2024]. For this reason, there is increasing interest in the accompanied development of methods of uncertainty quantification that only require "black-box" access to such models [Lee and Chen, 2009, 2007]. One particular method in this vein is "conformal prediction" [Shafer and Vovk, 2008, Angelopoulos et al., 2023]. While much work has been directed towards the improvement of the uncertainty estimation produced by conformalized predictors, there has been comparatively little emphasis placed on how such uncertainty estimates can be practically leveraged.

The central focus of this thesis, therefore, is to answer the following question:

How can we design principled uncertainty estimates for black-box models and use such uncertainty optimally for decision-making?

We answer this question over the course of three chapters, with each making novel contributions to the broader space of robust decision-making. In Chapter 2, we focus on scientific decision-making: scientific hypotheses are being increasingly informed by simulation-based parameter inference. Investigations across certain domains, such as astrophysics and neuroscience, however, have reached a scale where classical Bayesian inference techniques are rendered intractable and call for the use of ML-based approximate inference techniques. Unlike classical approaches, these approximations lack any guarantees, rendering subsequent claims on scientific conclusions dubious. We, therefore, here develop a novel approach to leverage conformal prediction to correct for these potential deficiencies of coverage of amortized variational inference and do so in a manner that targets maximal predictive efficiency.

This use of conformal prediction for calibrated scientific decision making, however, is merely a microcosm of the more general phenomenon of decision-making under uncalibrated predictors. We, therefore, generalize from the scientific context to a broader "predict-then-optimize" problem setting in Chapter 3, in which an upstream prediction informs the setup of a downstream decision-making problem. In place of making a decision simply by trusting this upstream predictor, we develop a novel framework that considers a robust formulation informed by conformalizing the upstream predictor, by which formal guarantees of the robust decision under the true, *unseen* parameter can be established. This investigation is amongst the first in the broader conformal prediction community to concretely address the question of:

What can prediction regions with formal coverage guarantees, especially over continuous spaces, be used for?

Since this initial contribution, a number of follow-up works in the broader conformal prediction community have generalized both the framework and applications of conformal decision-making. Chapter 4 presents two such extensions. In particular, we first highlight how this framework can be extended for decision-making under predictors conformalized using a recently proposed extension to the conformal prediction framework in which the standard scalar score function is replaced with a vector score formulation. We then extend this framework to settings of robust linear quadratic regulator (LQR) control, developing the first linear control framework and algorithm to have distribution-free, worst-case regret guarantees under dynamics misspecification.

We conclude this report by finally motivating some future-looking extensions in Chapter 5, where we highlight further applications that would be of interest in the intersection of decision-making with scientific discovery and engineering.

1.1 Preliminaries

We now present the background content necessary to understand the remainder of this thesis. The presentation herein is a compressed version of what would be found in more comprehensive treatments of each background topic. We first present a background on conformal prediction, the framework that underpins the collection of methods developed throughout this thesis. We then present the remaining topics, making reference to the chapter in which such material is respectively leveraged.

1.1.1 Conformal Prediction

Given a dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(N)}, y^{(N)})\}$ of i.i.d. observations from a distribution $\mathcal{P}(X, Y)$, conformal prediction [Angelopoulos and Bates, 2021, Shafer and Vovk, 2008] produces prediction regions with distribution-free theoretical guarantees. A prediction region maps from observations of X to sets of possible values for Y and is said to be marginally valid at the $1 - \alpha$ level if $\mathcal{P}_{X,Y}(Y \notin \mathcal{C}(X)) \leq \alpha$.

Split conformal is one popular version of conformal prediction. In this approach, marginally calibrated regions $\mathcal{C}$ are designed using a "score function" $s(x, y)$. Intuitively, the score function should have the quality that $s(x, y)$ is smaller when it is more reasonable to guess that $Y = y$ given the observation $X = x$. For example, if one has access to a function $\hat{f}(x)$ which attempts to predict Y from X, one might take $s(x, y) = \|\hat{f}(x) - y\|$. The score function is evaluated on each point of a dataset $\mathcal{D}_{\mathcal{C}}$ called the "calibration dataset," yielding $\mathcal{S} = \{s(x^{(j)}, y^{(j)})\}_{j=1}^{N_{\mathcal{C}}}$, where $N_{\mathcal{C}} := |\mathcal{D}_{\mathcal{C}}|$. Note that the calibration dataset cannot be used to pick the score function; if data is used to design the score function, it must independent of $\mathcal{D}_{\mathcal{C}}$. We then define $\widehat{q}(\alpha)$ as the $\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil / N_{\mathcal{C}}$ quantile of $\mathcal{S}$. For any future x, the set $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ satisfies $1 - \alpha \leq \mathcal{P}(Y \in \mathcal{C}(X))$. This inequality is known as the coverage guarantee, and it arises from the exchangeability of the score of a test point $s(x', y')$ with $\mathcal{S}$. The coverage guarantee possesses finite-sample properties.

As noted in Vovk's tutorial [Shafer and Vovk, 2008], while the coverage guarantee holds for any score function, different score functions may lead to more or less informative prediction regions. For example, the score $s(x, y) = 1$ leads to the highly uninformative prediction region of all possible values of Y. Predictive efficiency is one way to quantify informativeness, defined as the inverse of the expected Lebesgue measure of the prediction region, i.e. $(\mathbb{E}[|\mathcal{C}(X)|])^{-1}$ [Yang and Kuchibhotla, 2021, Sesia and Candès, 2020]. Methods employing conformal prediction often seek to identify prediction regions that are efficient and calibrated.

1.1.2 Variational Inference

We now discuss variational inference, the central focus of Chapter 2. Bayesian methods aim to sample the posterior distribution $\mathcal{P}(\Theta \mid X)$, typically using either MCMC or VI. VI has risen in popularity recently due to how well it lends itself to amortization. Given an observation X, variational inference transforms the problem of posterior inference into an optimization problem by seeking

$$\varphi^*(X) = \arg\min_{\varphi} D(q_\varphi(\Theta)||\mathcal{P}(\Theta \mid X)), \tag{1.1}$$

where D is a divergence and q_φ is a member of a variational family of distributions $\mathcal{Q}$ indexed by the free parameter φ. Normalizing flows have emerged as a particularly apt choice for $\mathcal{Q}$, as they are highly flexible and perform well empirically [Rezende and Mohamed, 2015, Agrawal et al., 2020]. Amortized variational inference expands on this approach by training a neural network to approximate $\varphi^*(X)$. This leads to a variational posterior approximator $q(\Theta \mid X) = q_{\varphi^*(X)}(\Theta)$ that can be rapidly computed for any value X. The characteristics of φ^* depend in part on the variational objective, D. For instance, using a reverse-KL objective, i.e. $D_{KL}(q_\varphi(\Theta)||\mathcal{P}(\Theta \mid X))$, is known to produce mode-seeking posterior approximations, whereas using a forward-KL objective, i.e. $D_{KL}(\mathcal{P}(\Theta \mid X)||q_\varphi(\Theta))$, encourages mode-covering behavior [Murphy, 2023]. Popular variational objectives include the Forward-Amortized Variational Inference (FAVI) objective [Ambrogioni et al., 2019, Bornschein and Bengio, 2014], the Evidence Lower Bound (ELBO), and the Importance Weighted ELBO (IWBO) [Burda et al., 2015].

1.1.3 Predict-Then-Optimize

We now discuss predict-then-optimize decision making problems, the central focuses of Chapter 3 and Chapter 4. Predict-then-optimize problems are formulated as

$$w^*(x) := \min_{w \in \mathcal{W}} \quad \mathbb{E}[C^T w \mid x], \tag{1.2}$$

where w are decision variables, C an *unknown* cost parameter, x observed contextual variables, and $\mathcal{W}$ a compact feasible region. The predict-then-optimize framework is so called as the nominal approach first predicts $\widehat{c} := f(x)$ and subsequently solves $\min_w \widehat{c}^T w$. Alternatively, a predictive contextual distribution $\mathcal{P}(C \mid x)$ is assumed, with respect to which the optimization formulation is solved. A full review is presented in [Elmachtoub and Grigas, 2022].

This formulation, however, is inappropriate in risk-sensitive downstream tasks. For this reason, recent works have begun investigating a risk-sensitive variant or "robust" alternative to this traditional formulation, namely by replacing $\mathbb{E}[C^T w \mid x]$ with $\max_{\widehat{c} \in \mathcal{U}(x)} \widehat{c}^T w$ [Ohmori, 2021, Chenreddy et al., 2022, Sun et al., 2023], where $\mathcal{U}(x)$ is constructed to guarantee coverage of c.

1.1.4 Representative Points

We now discuss representative points, an aspect studied in Chapter 3. The problem of summarizing the distribution of a random vector with points $\Xi := \{\xi^{(i)}\}_{i=1}^N$ arises in many contexts, such as in optimal stratification [Dalenius, 1950, Dalenius and Gurney, 1951], density estimation [Flury and Tarpey, 1993], and signal quantization [Max, 1960]. Such points are known as representative points (RPs). Denoting the space of all sets $\widehat{\Xi}$ such that $|\widehat{\Xi}| \leq n$ as ζ, the RPs of a random variable X are

$$\Xi := \underset{\widehat{\Xi} \in \zeta}{\arg\min}\, \mathbb{E}_X \left[\min_{\xi^{(i)} \in \widehat{\Xi}} ||X - \xi^{(i)}||_2^2 \right]. \tag{1.3}$$

For a comprehensive review, see [Fang and Pan, 2023]. Despite extensive study, no general algorithm exists for the efficient construction of representative points for arbitrary distributions. Typical implementations use clustering algorithms, such as Lloyd's algorithm, on $\{x^{(i)}\}_{i=1}^M \sim \mathcal{P}(X)$.

1.1.5 LQR & Control Co-Design

We now present a review of linear controls, a central focus of Chapter 4. The field of controls has a long history in engineering physics and robotics [Zabczyk, 2020]. In the linear quadratic regulator (LQR) setup, the state dynamics have a linear form $\dot{x} = Ax + Bu + w$, where x is the state, u the control inputs, and $w \sim \mathcal{D}_w$ the noise. Optimal control is then posed as an optimization problem, with the objective $J(u)$ weighing both the deviation from a target state and the necessary control input. LQR optimal controllers take a linear feedback form, namely $u^*(x) = -K^*x$ where K^* is known as the "optimal gain matrix" and solves

$$\begin{aligned} K^*(A, B) &:= \underset{K \in \mathcal{K}(A,B)}{\arg\min}\, \mathbb{E}[J(K, A, B)] \\ \text{where } J(K, A, B) &:= \int_0^\infty (x^\top Q x + (Kx)^\top R(Kx))dt \end{aligned} \tag{1.4}$$

where $\mathcal{K}(A, B) := \{K : \mathrm{Re}(\lambda_i(A - BK)) < 0\ \forall i\}$ is known as the set of "stabilizing controllers" for the A, B system dynamics and $\dot{x} = (A - BK)x + w$. Variations, where the integral is replaced by a discretized sum or considered to some finite T, are also of interest. Solving this problem is often done either by solving the algebraic Ricatti equation (ARE) [Willems, 1971] or via policy gradient [Sun and Fazel, 2021].

We now briefly summarize the relevant pieces of uncertain co-control design to highlight the robust control problem subproblem therein; for a full survey, refer to [Azad and Herber, 2022]. Engineering designs can often be specified by parameters θ, which could capture, for instance, the dimensions of an airfoil or material properties of a DC battery grid. The dynamics

are highly dependent on the design; for example, an airfoil with a shape θ_1 will fly differently from one given by θ_2. Worst-case robust UCCD with dynamics misspecification, thus, solves $\min_{K,\theta} \max_{\widehat{A}\in\mathcal{A}(\theta),\widehat{B}\in\mathcal{B}(\theta)} \mathbb{E}[o(K,\widehat{A},\widehat{B},\theta)]$, where $\dot{x} = \widehat{A}x + \widehat{B}u + w$ and $(\mathcal{A}(\theta),\mathcal{B}(\theta))$ are uncertainty sets of the dynamics for such a design and o is the objective.

Often, the objective takes a decomposable form, namely with one term relating to system control and the other depending on the design parameter, i.e. $o(K,A,B,\theta) := \ell(\theta)+J(K,A(\theta),B(\theta))$ [Chanekar et al., 2018, Ahmadi et al., 2023]. One commonly applied solution technique in this setting is via bilevel optimization, in which an outer optimization loop is performed over design parameters and an inner one over controllers for the current design iterate [Herber and Allison, 2019, Kamadan et al., 2017]. For this reason, the specification of the robust control subproblem can be studied independently of the outer design optimization loop, as done herein.

1.1.6 Quantile Envelopes

We finally present a review of quantile envelopes, which form the backbone of the ensembling approach discussed in Chapter 4. Generalizations of quantiles have a long history in statistics [Rousseeuw and Struyf, 1998, Serfling, 2002]. Unlike univariate data, multivariate data do not lend itself to an unambiguous definition of a quantile, as there is no canonical ordering in higher dimensional spaces. The notion of a "directional quantile" for a random variable $X \in \mathbb{R}^n$ can, however, be directly defined given some direction $u \in \mathcal{S}^{n-1}$, namely as $Q(X,\alpha,u) = \inf\{q \in \mathbb{R} : \mathcal{P}(u^\top X \leq q) \geq \alpha\}$ [Kong and Mizera, 2012, Paindaveine and Šiman, 2011, Hallin et al., 2010]. When there is no ambiguity, we just denote it as $Q(\alpha,u)$. For any given u, notice the choice of quantile defines a corresponding halfplane $H(u,Q(\alpha,u)) = \{x \in \mathcal{X} : u^\top x \leq Q(\alpha,u)\}$. The quantile envelope is then the intersection thereof:

$$D(\alpha) = \bigcap_{u\in\mathcal{S}^{n-1}} H(u,Q(\alpha,u)). \tag{1.5}$$

Notably, while each individual $H(u,Q(\alpha,u))$ captures $1-\alpha$ of the points, $D(\alpha)$ does *not*, as it is the intersection thereof and hence captures $< 1-\alpha$ of the mass. If $1-\alpha$ combined coverage is sought, a correction, such as Bonferroni adjustment, is used for the individual planes.

CHAPTER 2

Amortized Variational Inference with Coverage Guarantees

> If you thought that science was certain –
> well, that is just an error on your part.
> Richard Feynman

In this chapter, we discuss deficiencies of coverage under amortized variational inference and remedies thereof using conformal prediction. Amortized variational inference is an often employed framework in simulation-based inference that produces a posterior approximation that can be rapidly computed given any new observation. Unfortunately, there are few guarantees about the quality of these approximate posteriors. We propose Conformalized Amortized Neural Variational Inference (CANVI), a procedure that is scalable, easily implemented, and provides guaranteed marginal coverage. Given a collection of candidate amortized posterior approximators, CANVI constructs conformalized predictors based on each candidate, compares the predictors using a metric known as predictive efficiency, and returns the most efficient predictor. CANVI ensures that the resulting predictor constructs regions that contain the truth with a user-specified level of probability. CANVI is agnostic to design decisions in formulating the candidate approximators and only requires access to samples from the forward model, permitting its use in likelihood-free settings. We prove lower bounds on the predictive efficiency of the regions produced by CANVI and explore how the quality of a posterior approximation relates to the predictive efficiency of prediction regions based on that approximation. Finally, we demonstrate the accurate calibration and high predictive efficiency of CANVI on a suite of simulation-based inference benchmark tasks and an important scientific task: analyzing galaxy emission spectra.

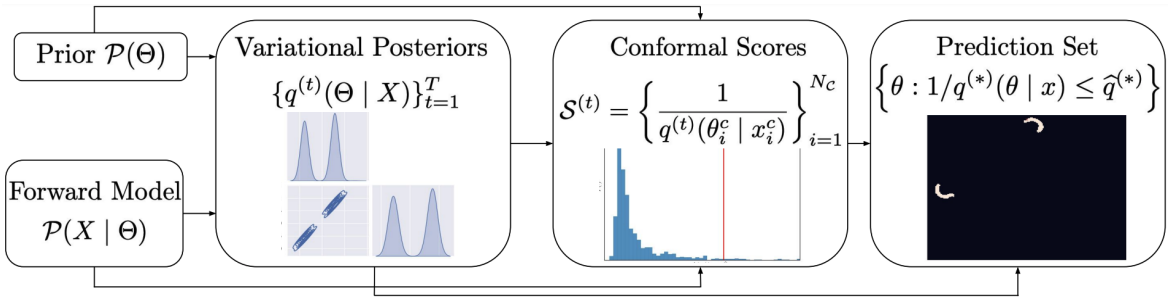

Figure 2.1: CANVI is a wrapper around variational inference requiring minimal implementation and computational overhead that produces prediction regions with guaranteed marginal calibration. Among a family of candidate amortized posterior approximators, CANVI can identify the approximator leading to the most efficient prediction regions. CANVI can be used in any setting where the forward model $\mathcal{P}(X \mid \Theta)$ can be sampled.

2.1 Introduction

In many scientific applications, such as in astrophysics, neuroscience, and particle physics [Papamakarios and Murray, 2016, Lueckmann et al., 2017, Greenberg et al., 2019, Deistler et al., 2022, Papamakarios et al., 2019, Boelts et al., 2022], posterior distributions $\mathcal{P}(\Theta \mid x)$ are sought over a large collection of x, typically on the order of 10,000 or more. In such scientific settings, (θ, x) pairs are assumed to come from nature, which has led to the growth of "simulation-based inference," in which a likelihood $\mathcal{P}(X \mid \theta)$ and prior $\mathcal{P}(\Theta)$ are posited and simulated to fit posteriors [Cranmer et al., 2020, Lueckmann et al., 2021]. Even when the likelihood and prior are well specified, exact sampling from the posteriors is intractable, requiring running 10,000 separate MCMC chains.

In these settings, amortized variational inference is frequently employed. Variational inference (VI) has become a staple in Bayesian inference; however, it has been repeatedly noted that a major shortcoming of VI is its lack of any theoretical guarantees and tendency to produce biased posterior estimates [Blei et al., 2017, Murphy, 2022, Zhang et al., 2018, Yao et al., 2018]. The most common metric for assessing the calibration of such inference algorithms is the "expected coverage." Having calibrated expected coverage is a necessary but not sufficient condition for conditional coverage, yet amortized variational approximations fail to even achieve this minimal requirement, despite significant work to remedy this shortcoming [Deistler et al., 2022, Delaunoy et al., 2022, Lemos et al., 2023, Delaunoy et al., 2023]. A lack of such calibration limits the capacity to reach downstream scientific conclusions, highlighted in a recent meta-study of likelihood-free inference algorithms [Hermans et al., 2021b].

In applications where many posteriors need to be estimated, credible regions with *marginally* calibrated coverage can be sufficient for downstream scientific inquiries. For instance, in astro-

physics, there is great interest in constraining the ΛCDM model, the current concordance model in cosmology [Gilman et al., 2021, Hezaveh et al., 2016, Vegetti et al., 2010, Vegetti and Koopmans, 2009, Hogg and Blandford, 1994]. A recent work from this community, [Hermans et al., 2021a], leveraged Bayesian inference towards this end, obtaining approximate posteriors for the parameters of interest on each of 10,000 observations. Crucially, these posteriors were then used to produce credible intervals with *marginally* valid frequentist coverage, from which they made claims on the ΛCDM model.

Our insight is that conformal prediction can be leveraged to provide variational approximators with marginal coverage guarantees and provide users new ways to measure the quality of variational approximators. In this manuscript, we present CANVI (Conformalized Amortized Neural Variational Inference), a novel, general framework for producing marginally calibrated, informative prediction regions from a collection of variational approximators. Such regions can be produced with minimal implementation and computational overhead, requiring only samples from the prior and $\mathcal{P}(X \mid \Theta)$, as shown in Figure 2.1. In Section 2.3.3, we provide theoretical analysis of the informativeness of the prediction regions produced by CANVI using a measure known as "predictive efficiency." High predictive efficiency is necessary to draw conclusions in downstream scientific inquiries and relates to asymptotic conditional coverage with the appropriate choice of score function. Finally, in Section 2.4, we show calibration and predictive efficiency across simulation-based inference benchmark tasks and an important scientific task: analyzing galaxy emission spectra.

2.2 Related Literature

Efforts to correct for miscalibration have been of great interest recently in light of its apparent omnipresence highlighted in [Hermans et al., 2021b]. The notion of calibration studied therein, and the one we concentrate on here, centers on the expected coverage probability of the highest predictive density (HPD) of a posterior estimate $q(\Theta \mid X)$, defined for such a q, some fixed x, and pre-specified coverage level α to be the set $\mathrm{HPD}(q(\Theta \mid x), 1-\alpha)$ with smallest Lebesgue measure such that $\mathcal{P}_{\Theta \sim q(\Theta \mid x)}(\Theta \in \mathrm{HPD}(q(\Theta \mid x), 1-\alpha)) \geq 1-\alpha$. The expected coverage is then $\mathbb{E}_{X,\Theta \sim \mathcal{P}(X,\Theta)}[\mathbb{1}[\Theta \in \mathrm{HPD}(q(\Theta \mid X), 1-\alpha)]]$.

Such calibration has been studied across a number of posterior estimation techniques in the likelihood-free inference community, which generally fall into one of three categories. Neural posterior approximation (NPE) methods directly approximate the posterior density $\mathcal{P}(\Theta \mid X)$ [Papamakarios and Murray, 2016, Lueckmann et al., 2017, Greenberg et al., 2019, Deistler et al., 2022], neural likelihood estimation (NLE) methods estimate the assumed intractable likelihood $\mathcal{P}(X \mid \Theta)$ [Papamakarios et al., 2019, Boelts et al., 2022], and neural ratio estimation (NRE)

methods estimate the $\mathcal{P}(X \mid \Theta)/\mathcal{P}(X)$ ratio [Hermans et al., 2020, Durkan et al., 2020b, Miller et al., 2022]. Both NLE and NRE rely on MCMC for posterior sampling after estimation.

While these approaches differ in their estimation strategies, they all suffer miscalibration if naively employed. One suggestion advocated by [Hermans et al., 2021b] was ensembling, compatible with all the aforementioned posterior approximation strategies. Despite its improvement in empirical calibration, ensembling affords no guarantees on calibration and also dramatically increases the computational cost of training and inference alike. As an effort to address this lack of guarantees and computational cost, recent calibration efforts have focused on modifying the loss function to appropriately encourage conservatism in the learned posterior estimates, since the downstream scientific use cases can afford such conservatism, unlike underdispersed posteriors. In particular, [Delaunoy et al., 2022] took a first step in this direction by proposing a modification over the vanilla NRE formulation with a regularization term that results in more conservative posterior estimates. Such a method, however, has no guarantees and further relies on the selection of a tunable λ parameter, whose optimal selection is unknown without awareness of the true posterior. This approach was then extended to NPE and NLE in [Delaunoy et al., 2023] and [Falkiewicz et al., 2023], which both again suffer from the deficiencies of producing overly conservative posteriors and relying on the careful selection of λ.

2.3 Method

In response to the shortcomings highlighted in the previous section, we were interested in procuring a method that has (1) guarantees on calibration without tuning parameters, (2) minimal computational overhead, and (3) informative prediction regions. We demonstrate in Section 2.3.1 that leveraging conformal prediction on an amortized VI approximator $q(\Theta \mid X)$ immediately addresses points (1) and (2). We then study in the following sections how this naive application can be extended to the full CANVI algorithm by considering a collection of amortized VI approximators $\{q^{(1)}(\Theta \mid X), \ldots, q^{(T)}(\Theta \mid X)\}$ to address point (3). CANVI can be applied whenever $\mathcal{P}(X, \Theta)$ can be sampled. The coverage validity of CANVI is proven in Section 2.3.2, and analyses of its predictive efficiency in Section 2.3.5.

2.3.1 CANVI: Score Function

In the simplest case, CANVI takes as input a single amortized posterior approximator $q(\Theta \mid X)$. In traditional applications of split conformal, much concern is given to the loss of accuracy of the predictor q in having to reserve a subset of the training data for calibration. Here we have no such issues; we sample $\mathcal{D}_{\mathcal{C}} = \{(x_i^c, \theta_i^c)\}_{i=1}^{N_{\mathcal{C}}} \overset{\text{iid}}{\sim} \mathcal{P}(\Theta)\mathcal{P}(X \mid \Theta)$ from the joint distribution to produce a

calibration dataset that can be arbitrarily large. Given $q(\Theta \mid X)$, we employ the following score, as used in [Angelopoulos and Bates, 2021]:

$$s(x_i, \theta_i) = (q(\theta_i \mid x_i))^{-1}. \tag{2.1}$$

Denoting the $\lceil (N_{\mathcal{C}} + 1)(1 - \alpha) \rceil / N_{\mathcal{C}}$ quantile of the score distribution over $\mathcal{D}_C$ as $\widehat{q}_{\mathcal{C}}(\alpha)$, $\mathcal{C}(x) = \{\theta : 1/q(\theta \mid x) \leq \widehat{q}_{\mathcal{C}}(\alpha)\}$ is then marginally calibrated. It may be disjoint if the posterior is multimodal.

While other choices of score functions also result in regions $\mathcal{C}(x)$ that address both the lack of guarantees and computational cost of methods highlighted in Section 2.2, the particular choice of Equation (2.1) has the desirable property that, if we recover the true posterior, that is $q(\Theta \mid X) = \mathcal{P}(\Theta \mid X)$, we recover the HPDs, namely $\mathcal{C}(x) = \mathrm{HPD}(\mathcal{P}(\Theta \mid x), 1 - \alpha)$, achieving conditional coverage. Note that the procedure described in this section and those that follow can be easily extended to the group conditional setting if such coverage is of interest, whose discussion we defer to Section A.1.

2.3.2 CANVI: Approximator Selection

To mitigate the risk of producing uninformative prediction regions, it is natural to explore multiple posterior approximations, $\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}$, since a poorly chosen approximator may lead to poor predictive efficiency. These posterior approximations could, for instance, differ in their choice of training objective, variational family, or hyperparameters. CANVI seeks to identify the variational approximator $q^{(t^*)}$ for which the efficiency is greatest, defined for a threshold τ to be

$$\ell(q, \tau) := \mathbb{E}_X \left[\mathcal{L} \left(\{\theta : 1/q(\theta \mid X) \leq \tau\} \right) \right], \tag{2.2}$$

We defer the discussion of estimating $\ell(q, \tau)$ to Section 2.3.4. Naively, one would expect by taking $(q^{(t^*)}, \widehat{q}_{\mathcal{C}}^{(t^*)}(\alpha))$ where $t^* := \arg\min_t \ell(q^{(t)}, \widehat{q}_{\mathcal{C}}^{(t)}(\alpha))$, we achieve maximal efficiency and retain coverage guarantees. However, defining $\mathcal{C}(x)$ with $\widehat{q}_{\mathcal{C}}^{(t^*)}(\alpha)$ fails to retain coverage guarantees, as the exchangeability of scores of future test points $s(x', \theta')$ with $\mathcal{S}$ is lost in conditioning on $\mathcal{D}_{\mathcal{C}}$ for selecting t^*.

We must, therefore, perform an *additional* recalibration step after selecting t^* to retain coverage guarantees. CANVI performs such recalibration using an additional dataset $\mathcal{D}_{\mathcal{R}}$ again constructed with i.i.d. draws from $\mathcal{P}(X, \Theta)$. We take $\mathcal{D}_{\mathcal{R}}$ to be the same size as $\mathcal{D}_{\mathcal{C}}$, i.e. $|\mathcal{D}_{\mathcal{R}}| = N_{\mathcal{C}}$. CANVI then computes the quantile $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) := \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)$, which is used to define prediction regions $\mathcal{C}(x) := \{\theta : 1/q^{(t^*)}(\theta \mid x) \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)\}$. However, such regions require analysis to guarantee high efficiency, as we explore in Section 2.3.3.

The full CANVI framework is provided in Algorithm 1. The validity of the CANVI procedure follows directly from that of split conformal prediction, formally stated below and explicitly proven in Appendix A.2.

Lemma 2.3.1. *Let* $\alpha \in (0, 1)$ *and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X, \Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

Let $(x', \theta') \sim \mathcal{P}(X, \Theta) \perp\!\!\!\perp \mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{R}} \cup \mathcal{D}_{\mathcal{T}}$*, with* $\mathcal{D}$ *being the data used to train* $\{q^{(t)}(\Theta|X)\}_{t=1}^{T}$*. Then* $1-\alpha \leq \mathcal{P}(1/q^{(*)}(\theta'|x') \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha))$.

Algorithm 1 CANVI: Note that VOLUMEEST is a volume estimator subroutine detailed in Section 2.3.4.

procedure CPQUANTILE
Inputs: Posterior approximation $q(\Theta \mid X)$, Calibration set $\mathcal{D}_{\mathcal{C}}$, Desired coverage $1-\alpha$
  $\mathcal{S} \leftarrow \{\frac{1}{q(\theta_i|x_i)}\}_{i=1}^{N_{\mathcal{C}}}$
  Return $\frac{\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil}{N_{\mathcal{C}}}$ quantile of $\mathcal{S}$
end procedure
procedure CANVI
Inputs: Posterior approximators $\{q^{(t)}(\Theta|X)\}_{t=1}^{T}$, Prior $\mathcal{P}(\Theta)$, Forward model $\mathcal{P}(X \mid \Theta)$, Desired coverage $1-\alpha$, Calibration size $N_{\mathcal{C}}$, Test size $N_{\mathcal{T}}$
  $\mathcal{D}_{\mathcal{C}}, \mathcal{D}_{\mathcal{R}} \sim \mathcal{P}(X, \Theta), \mathcal{D}_{\mathcal{T}} \sim \mathcal{P}(X)$
  for $t \in \{1, \dots T\}$ **do**
    $\widehat{q}_{\mathcal{C}}^{(t)} \leftarrow$ CPQUANTILE$(q^{(t)}, \mathcal{D}_{\mathcal{C}}, 1-\alpha)$
    $\widehat{\ell}^{(t)} \leftarrow$ VOLUMEEST$(q^{(t)}, \mathcal{P}(\Theta), \widehat{q}_{\mathcal{C}}^{(t)}, \mathcal{D}_{\mathcal{T}})$
  end for
  $t^* \leftarrow \arg\min_t \widehat{\ell}^{(t)}$
  $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) \leftarrow$ CPQUANTILE$(q^{(t^*)}, \mathcal{D}_{\mathcal{R}}, 1-\alpha)$
  Return $q^{(*)}(\Theta \mid X)$, $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$
end procedure

2.3.3 CANVI: Efficiency Analysis Assumptions

We now show that, with high probability, the pair CANVI produces $(q^{(*)}(\Theta|X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha))$ is the most efficient amongst the candidate posteriors considered. The concern is that the post-recalibrated quantile may result in significant degradation of the efficiency, i.e. $\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \gg \ell(q^{(*)}, \widehat{q}_{\mathcal{C}}^{(*)}(\alpha))$. This tradeoff between coverage and efficiency was studied in [Yang and Kuchibhotla, 2021].

Recall $1-\alpha \leq \mathcal{P}(\Theta \in \mathcal{C}(X)) \leq 1-\alpha+1/(N_{\mathcal{C}}+1)$. Denote the CDF of the score function under the joint distribution $\mathcal{P}(X,\Theta)$ as $\mathcal{F}(s) := \mathcal{P}_{\Theta,X}(1/q(\Theta \mid X))$. The coverage guarantee, thus, implies $\widehat{q}(\alpha) \in [\mathcal{F}^{-1}(1-\alpha), \mathcal{F}^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1))]$ for $\widehat{q}(\alpha)$ from *any* calibration set. In particular, $\widehat{q}_{\mathcal{C}}^{(*)}(\alpha)$ and $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$ both lie in this range.

We bound the efficiency suboptimality by proceeding in two steps. We first demonstrate that the quantiles of $q^{(*)}(\Theta \mid X)$ under $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ are close by demonstrating the quantile range of $\mathcal{F}^{-1}$ is small. We then demonstrate the efficiency varies smoothly as a function of the quantile, allowing us to bound the resulting efficiency change. Formally, we state these assumptions respectively as follows, per [Yang and Kuchibhotla, 2021], which we then demonstrate follow from properties of the chosen variational families.

Assumption 1. *For each t, the $\ell(q^{(t)}, \tau)$ is Lipschitz continuous in τ with constant L_W.*

Assumption 2. *For each t, $\exists\ r, \gamma \in (0,1]$ such that $\mathcal{F}_t^{-1}(s)$ (inverse score CDF under $q^{(t)}$), is γ-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ with continuity constant L_t.*

2.3.3.1 Lipschitz Continuity of Efficiency

We now demonstrate Assumption 1 can be guaranteed with the appropriate selection of variational family by the end user. It suffices to demonstrate $\ell_x(q,\tau) := \mathcal{L}(\{\theta : 1/q(\theta \mid x) \leq \tau\})$ is L-Lipschitz continuous in τ for any $x \in \mathcal{X}$, proven in Appendix A.4.1.

For any fixed x, $\ell_x(q,\tau)$ is the intrinsic volume of the sublevel set of $1/q(\theta \mid x)$. We summarize and subsequently use relevant results from [Jubin, 2019] below. "Intrinsic volume" defines the notion of volume for a lower dimensional manifold embedded in a higher dimensional space. For a brief review of Riemannian manifolds, see Appendix A.3; a more complete presentation is available in [Lee and Lee, 2012]. The $n-k$ degree intrinsic volume of a flat compact n-dimensional manifold N is

$$\mathcal{L}_{n-k}(N) = b_k \int_{\partial N} \operatorname{tr}\left(\bigwedge^{k-1} S\right) \operatorname{vol}_{\partial N}, \tag{2.3}$$

where $0 \leq k \leq n$, $b_k \in \mathbb{R}$, and S is the second fundamental form of ∂N in N. Lipschitz continuity of $\mathcal{L}_{n-k}(M_f^\tau)$ was established as follows. The level sets $M_f^\tau := f^{-1}((-\infty, \tau])$ are restricted to "regular values" of f, namely τ such that $f(x) = \tau \implies df(x) \neq 0$.

Theorem 2.3.2. *Let (M,g) be an n-dimensional Riemannian manifold, $f \in \mathcal{C}^3(M,\mathbb{R})$ bounded below, and τ a regular value of f equipped with the standard uniform $\mathcal{C}^3$ topology. Then, if $0 \leq k \leq n$, $\tau \to \mathcal{L}_{n-k}(M_f^\tau)$ is Lipschitz continuous [Jubin, 2019].*

The Lipschitz continuity of $\ell_x(q,\tau)$ then follows as a corollary for any variational families for which the density is sufficiently smooth, namely $q(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$. The proof follows by taking

$\ell_x(q,\tau) := \mathcal{L}_n((\mathbb{R}^n)^\tau_s)$ for $s(\theta) = 1/q(\theta|x)$, with $\tau = s(\theta)$ for $\{\theta : \nabla_\theta q(\theta|x) \neq 0\}$ being the set of regular values. This domain restriction is discussed more in Section 2.3.5. Notice $s(\theta) \in \mathcal{C}^3(\mathbb{R}^n)$ as both $f(x) := 1/x$ and $q(\theta \mid x)$ are $\in \mathcal{C}^3(\mathbb{R}^n)$ and $\mathcal{C}^3$ is closed under function composition. Formally,

Corollary 2.3.3. *Suppose for any $x \in \mathcal{X}$, $q(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ is bounded above and τ is a regular value of $q(\theta \mid x)$. Then, $\ell(q,\tau)$ is Lipschitz continuous in τ.*

Notably, this assumption on smoothness holds for most variational families used in practice, including highly expressive flow-based variational families [Köhler et al., 2021].

2.3.3.2 Continuity of Conformal Quantiles

We now discuss the validity of Assumption 2. Comparable assumptions are commonly used in the quantile estimation literature, as discussed in [Lei et al., 2018]. The Hölder constant cannot be characterized in general, as it is intimately tied to specific details of the score distribution under $\mathcal{P}(X,\Theta)$. We, therefore, provide an explicit characterization for a particular family of distributions in Theorem 2.3.4 and defer extensions to a broader set of families to future work. Details for this proof are given in Appendix A.5.

Theorem 2.3.4. *Let Θ and X be zero-mean unit-variance Gaussian random variables with correlation ρ. Let $q^{(t)}(\theta|x) = \mathcal{N}(\theta; tx, 1-\rho^2)$. Let $\kappa := t^2 - 2t\rho + 1$ and $r > 0$. Then $F_t^{-1}(z)$, is 1-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ with Hölder constant*

$$\frac{\kappa\Phi^{-1}(\frac{1-\alpha}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{1-\alpha}{2})^2 - \frac{(1-\alpha)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}} \tag{2.4}$$

Notably, the Hölder constant is minimized in recovering the true posterior, as Equation (2.4) is minimized at $\varphi = \rho$.

2.3.4 CANVI: Volume Estimation

We now provide an estimation procedure for $\ell(q,\tau)$. Naively, we might expect taking the sample average of $\ell_{x_i}(q,\tau)$ over $\mathcal{D}_\mathcal{T} := \{x_i\}_{i=1}^{N_\mathcal{T}} \sim \mathcal{P}(X)$ would suffice. However, exact calculation of $\ell_{x_i}(q,\tau)$ requires a grid-discretization over Supp$(\Theta|x_i)$, which is only feasible when the support has a known, small extent.

As a result, $\ell_x(q,\tau)$ is estimated using an importance-weighted Monte Carlo estimate over S samples from q. Such an estimator, however, suffers from high variance if q is underdispersed,

as $\mathcal{C}(x)$ will cover regions of low variational density. To combat this issue, we use the well-known fact that $\mathcal{P}(\Theta \mid X)$ is narrower than $\mathcal{P}(\Theta)$ to construct a "mixed sampler," specifically with $z_j \sim \text{Bern}(\lambda)$ and $\theta_j \sim q(\Theta|x_i)^{z_j}\mathcal{P}(\Theta)^{1-z_j}$, where the mixed density is now $\widetilde{q}(\theta_j) = \lambda q(\theta_j|x_i) + (1-\lambda)\mathcal{P}(\theta_j)$. The necessity of such mixing is dependent on the nature of the variational posterior with respect to the true posterior, which is unknown in practice. We, thus, average several estimates over $\{\lambda_k\} \in [0,1]$. Denoting the mixed density with λ_k as $\widetilde{q}_k$, for each x_i and λ_k, we make S draws $\{\theta_{jk}\}_{j=1}^{S} \sim \widetilde{q}_k(\Theta \mid x_i)$. We empirically demonstrate the necessity of such mixed sampling in Section 2.4.2.2. This procedure is summarized in Algorithm 2, with the final estimate $\widehat{\ell}(q,\tau)$ being

$$\frac{1}{K\mathcal{N}_{\mathcal{T}}} \sum_{i,k=1}^{\mathcal{N}_{\mathcal{T}},K} \frac{1}{S} \sum_{j=1}^{S} \frac{1}{\widetilde{q}_k(\theta_{jk}|x_i)} \mathbb{1}\left[\frac{1}{q(\theta_{jk}|x_i)} \leq \tau\right] \tag{2.5}$$

Algorithm 2 VOLUMEEST

procedure VOLUMEEST
  Inputs: Posterior approximation $q(\Theta \mid X)$, Prior $\mathcal{P}(\Theta)$, CP quantile $\widehat{q}$, Test set $\mathcal{D}_{\mathcal{T}}$
  for $i \in \{1, \ldots \mathcal{N}_{\mathcal{T}}\}, k \in \{1, \ldots K\}$ **do**
    for $j \in \{1, \ldots S\}$ **do**
      $z_j \sim \text{Bern}(k/K)$
      $\theta_j \sim q(\Theta \mid x_i)^{z_j}\mathcal{P}(\Theta)^{1-z_j}$
      $\widetilde{q}_j \leftarrow (k/K)q(\theta_j \mid x_i) + (1-k/K)\mathcal{P}(\theta_j)$
    end for
    $V_{i,k} \leftarrow \frac{1}{S}\sum_{j=1}^{S} \frac{1}{\widetilde{q}_j} \mathbb{1}[1/q(\theta_j|x_i) \leq \widehat{q}]$
  end for
  Return $\frac{1}{K\mathcal{N}_{\mathcal{T}}} \sum_{i,k} V_{i,k}$
end procedure

2.3.5 CANVI: Efficiency Proof

We now state the result of recovery of the optimal recalibrated approximator. To do so, we require the Monte Carlo estimate to be sufficiently well-behaved to recover the optimal pre-recalibration approximator.

Assumption 3. *If* $t^* := \arg\min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha))$ *and* $\widehat{t}^* := \arg\min_{1 \leq t \leq T} \widehat{\ell}(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha))$ *for* $\alpha \in (0,1)$*, then* $\exists \Delta, \epsilon > 0$*, such that with probability at least* $1-\epsilon$*,* $|\ell(q^{(\widehat{t}^*)}, \widehat{q}_{\mathcal{R}}^{(\widehat{t}^*)}(\alpha)) - \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha))| < \Delta$.

Important to note is that $\widehat{\ell}$ is only used to select t^*, after which we make claims on ℓ (i.e. *not* the estimate) for the recalibrated quantiles in Theorem 2.3.5. As with Assumption 2, this assumption is intimately tied to specific details of the score distribution under $\mathcal{P}(X, \Theta)$, making its restatement

in more natural distributional properties of q impossible. We demonstrate its validity empirically across several posteriors in Section 2.4.2.

The proof of Theorem 2.3.5 now follows as an extension of Theorem 3 from [Yang and Kuchibhotla, 2021] and is explicitly provided in Appendix A.4.2. Notably, we use Corollary 2.3.3 to replace Assumption 1 with a more natural set of conditions for this context. This requires ensuring that, for any x, $\widehat{q}_{\mathcal{C}}^{(*)}(\alpha)$ and $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha)$ are regular values of $s(\theta)$. We, thus, assume for any x and $\theta \neq 0$, $\mathcal{L}(\{\theta : \nabla_\theta q(\theta|x)\}) = 0$, which naturally holds for variational families used in practice, i.e. any non-piecewise constant density estimator.

Theorem 2.3.5. *Suppose for any $x \in \mathcal{X}$ and $t = 1, ..., T$, $q^{(t)}(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ is bounded above and for $\theta \neq 0$, $\mathcal{L}(\{\theta : \nabla_\theta q^{(t)}(\theta|x)\}) = 0$. Further assume $P(X, \Theta)$ is bounded above. Let $\alpha \in (0, 1)$ and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \\ \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X, \Theta), 1 - \alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ and $\delta \in [0, 1]$, Assumption 2 holds and for $\Delta, \epsilon > 0$ Assumption 3 holds, then with probability at least $(1 - \epsilon)(1 - \delta)$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}} \right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}} \right)^{\gamma} \right], \quad (2.6)$$

where γ, L_W, and $L_{[T]} = \max_{1 \leq t \leq T} L_t$ are constants defined in Assumptions 2 and 1.

Again, in any setting where it is possible to sample from $\mathcal{P}(X, \Theta)$, $N_{\mathcal{C}}$ can be made arbitrarily large, tightening the bound in Equation 2.6. Practitioners can, thus, focus on obtaining efficient predictors knowing that CANVI will make the optimal selection with high probability.

2.4 Experiments

As discussed, the main advantages of CANVI over alternative strategies are its guaranteed calibration without the need for tuning parameters, minimal computational overhead, and informative prediction regions. For this reason, we present three experiments herein. The first (Section 2.4.1) seeks to validate the first two claims by comparing BNRE, BNRE C, NPE, NRE, NRE C, and Ratio BNPE to the vanilla version of CANVI i.e. when it is applied to a single $q(\Theta \mid X)$, where no recalibration is necessary.

Notably, the informativeness of prediction regions, the focus of the third claim, is only of interest once a predictor is calibrated, meaning the comparison with alternatives is only meaningful

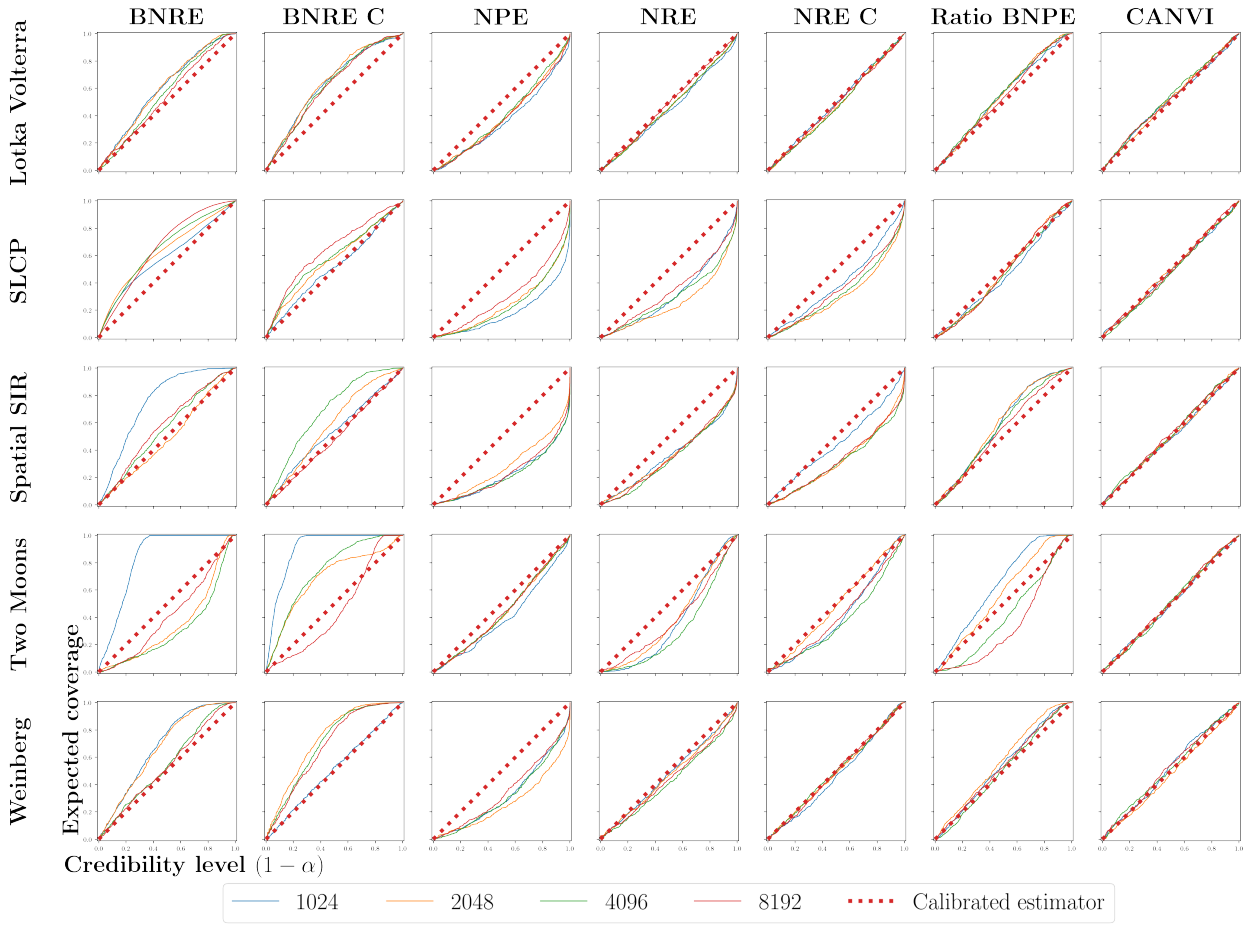

Figure 2.2: Calibration on the SBI benchmarks across different calibration strategies. Perfect calibration corresponds to the highlighted $y = x$ curve. Conservative prediction regions lie above this calibrated line and overconfident ones below. Conformalized lines (CANVI) are difficult to distinguish, as they all lie along the desired $y = x$ curve.

when they are nearly calibrated. As demonstrated in Section 2.4.1, however, the alternatives fail to consistently demonstrate calibration, rendering such a comparison moot. For this reason, the following experiment (Section 2.4.2) focuses on demonstrating that applying the full CANVI procedure to a collection $\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}$ both retains coverage under recalibration and ultimately recovers the most efficient predictor. The latter hinges upon the validity of Assumption 3 for the Monte Carlo efficiency estimator in Equation (2.5). We, therefore, demonstrate in the subsequent experiments that this estimator exhibits the desired estimation consistency when applied in settings where posterior estimators are trained to different epochs (Section 2.4.2.1) or with different training objectives (Section 2.4.2.2). Finally, we demonstrate CANVI can computationally scale up to scientific problems of interest in Section 2.4.3.

In all experiments, coverage for variational posteriors is assessed using Monte Carlo estimation, namely by constructing the highest density credible region per x_i. That is, for a given x_i, the ζ such that $\{\theta_j \mid q(\theta_j \mid x_i) \geq \zeta\}$ captures $1 - \alpha$ of the probability mass is estimated by drawing $\{\theta_j\}_{j=1}^{N} \sim q(\Theta|x_i)$ and finding the $1-\alpha$ quantile of $\{q(\theta_j|x_i)\}_{j=1}^{N}$. Coverage of the true parameter θ can be assessed by checking if $q(\theta \mid x_i) \geq \zeta$. Details are provided in Section A.7, and code is available at https://github.com/yashpatel5400/canvi.git.

2.4.1 Coverage Calibration

We evaluate on the standard SBI benchmark tasks, highlighted in [Delaunoy et al., 2023]. For full descriptions of the tasks, refer to Section A.6. Again following the precedent from previous SBI works, we present the calibration of the models trained over several simulation budgets ($|\mathcal{D}|$). CANVI was applied to an NPE, in which $\mathcal{D}_C$ was taken to be 10% of the simulation budgets and the remainder used for training. Calibration was completed in **under one second** for each task. Coverage was assessed 1,000 i.i.d. samples. Figure 2.2 demonstrates the miscalibration of the alternative approaches and the correction afforded with CANVI.

2.4.2 Predictive Efficiency

2.4.2.1 Training Epochs

We now study the application of CANVI to a collection of posteriors. In this experiment, $q^{(t)}$ is taken to be the t-th iterate of training a Neural Spline Flow family against the $\mathcal{L}_{\text{FAVI}}$ objective [Durkan et al., 2019]. Generally, we expect efficiency to improve with training iterates, as $q(\Theta \mid x)$ better approximates $\mathcal{P}(\Theta \mid x)$; however, the most efficient iterate may not occur at $t = T$. Selecting an intermediate t is comparable to the practice of retaining the training iterate with the best validation performance in prediction tasks.

Table 2.1: Coverage rates and standard errors for θ before (rows 1-4) and after conformalization (rows 5-8) by CANVI, for ARCH (left table) and SED (right table), assessed by checking for inclusion of θ in the $1-\alpha$ highest density region. Non-conformalized regions were estimated empirically from batches of 1000 i.i.d. samples per point. $\ell(q, \widehat{q}_{\mathcal{C}}(0.05))$ were computed with explicit gridding, discretizing each dimension into 200 bins; such estimation was intractable for $\Theta \in \mathbb{R}^{11}$ for SEDs. $\widehat{\ell}$ was estimated using Algorithm 2, with $K = 1$ and $K = 10$ mixing discretizations.

$1-\alpha$	ELBO	IWBO	FAVI
0.50	0.0007 (0.0008)	0.1031 (0.0110)	0.5514 (0.0127)
0.75	0.0044 (0.0018)	0.1994 (0.0115)	0.7534 (0.0127)
0.90	0.0195 (0.0044)	0.3263 (0.0122)	0.8797 (0.0065)
0.95	0.0396 (0.0061)	0.4074 (0.0186)	0.9260 (0.0083)
0.50	0.4970 (0.0124)	0.5019 (0.0167)	0.5086 (0.0144)
0.75	0.7488 (0.0139)	0.7559 (0.0126)	0.7565 (0.0092)
0.90	0.8978 (0.0080)	0.9036 (0.0111)	0.9005 (0.0110)
0.95	0.9496 (0.0071)	0.9548 (0.0052)	0.9487 (0.0081)
$\ell(q, \widehat{q}_{\mathcal{C}}(0.05))$	1.3184 (0.2178)	1.4211 (0.1128)	0.7451 (0.0641)
$\widehat{\ell}_{K=1}(q, \widehat{q}_{\mathcal{C}}(0.05))$	50.6595 (8.1313)	69.6484 (2.2495)	19.3635 (0.8868)
$\widehat{\ell}_{K=10}(q, \widehat{q}_{\mathcal{C}}(0.05))$	1.4151 (0.2181)	1.5206 (0.1114)	0.8402 (0.0656)

$1-\alpha$	ELBO	IWBO	FAVI
0.50	0.1683 (0.0088)	0.6185 (0.0111)	0.4986 (0.0180)
0.75	0.4556 (0.0146)	0.7978 (0.0079)	0.7550 (0.0105)
0.90	0.5803 (0.0106)	0.8881 (0.0064)	0.9028 (0.0092)
0.95	0.6824 (0.0144)	0.9248 (0.0083)	0.9532 (0.0085)
0.50	0.4901 (0.0156)	0.4937 (0.0101)	0.5024 (0.0203)
0.75	0.7485 (0.0163)	0.7513 (0.0139)	0.7542 (0.0163)
0.90	0.8985 (0.0098)	0.9013 (0.0096)	0.9022 (0.0067)
0.95	0.9499 (0.0084)	0.9510 (0.0061)	0.9461 (0.0059)
$\widehat{\ell}(q, \widehat{q}_{\mathcal{C}}(0.05))$	∞	1.3849 (1.2357) $\times 10^9$	5.3732 (3.3302) $\times 10^6$

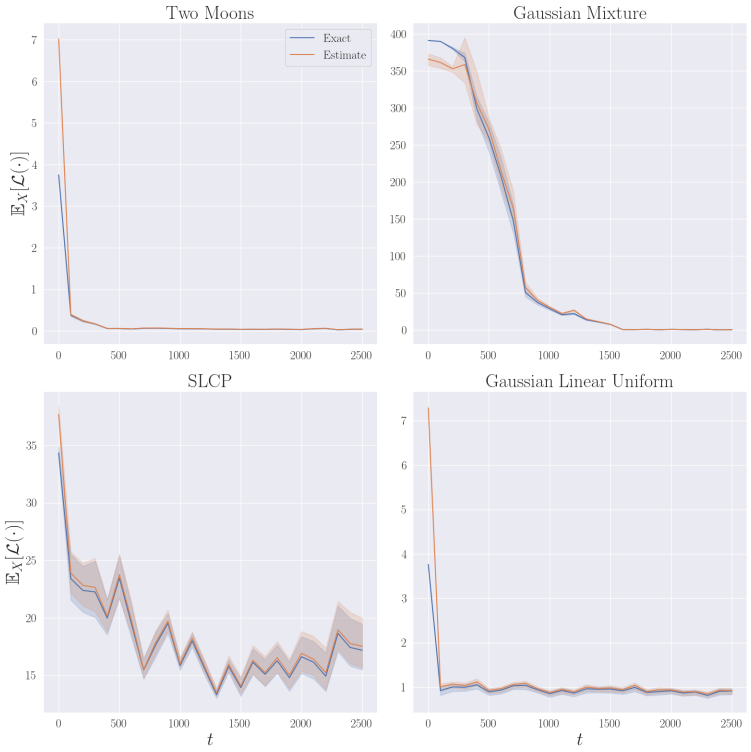

Figure 2.3: $\ell(q, \tau)$ and $\widehat{\ell}(q, \tau)$ for $\tau = \widehat{q}_{\mathcal{C}}^{(t)}(\alpha)$. Error bars across test batches are plotted.

We first study the validity of Assumption 3 by comparing $\widehat{\ell}(q, \tau)$ to $\ell(q, \tau)$. To compare to $\ell(q, \tau)$, we restrict experiments to projected posteriors of $\tilde{\Theta} = (\theta_1, \theta_2)$ for the Two Moons, Gaussian Mixture, SLCP, and Gaussian Linear Uniform tasks, for which explicit gridding was tractable.

$\widehat{\ell}(q^{(t)}, \widehat{q}_{\mathcal{C}}^{(t)}(\alpha))$ was estimated for a fixed $\alpha = 0.05$ and t taken every 100 training steps with 5 batches of 100 test points ($|\mathcal{D}_{\mathcal{T}}| = 100$) using $S = 10,000$ importance-weighted i.i.d. samples for each of $K = 10$ mixed samplers. From Figure 2.3, we see that $\widehat{\ell}(q, \tau)$ tracks closely to $\ell(q, \tau)$ across all tasks, giving credence to Assumption 3. We visualize the credible regions in Section A.8.

2.4.2.2 Training Objectives

We now similarly study $q^{(t)}$ across training objectives, taking one iterate of q trained against each of $\mathcal{L}_{\text{FAVI}}$, $\mathcal{L}_{\text{ELBO}}$, and $\mathcal{L}_{\text{IWBO}}$ ($K = 10$), giving us three amortized posteriors as input for CANVI, for a lag-one ARCH model:

$$y^{(m)} = \theta_1 y^{(m-1)} + e^{(m)},\ e^{(m)} = \xi^{(m)} \sqrt{0.2 + \theta_2 \left(e^{(m-1)}\right)^2},$$

where $y^{(0)} = 0$, $e^{(0)} = 0$, $M = 100$, and the $\xi^{(m)}$ are independent standard normal random variables [Thomas et al., 2022], detailed in Section A.6.9.

Table 2.1 shows that prior to conformalization, the variational posteriors are generally miscalibrated: training by the ELBO or IWBO results in significant under-coverage, as targeting either is known to find solutions that are mode-seeking. While the variational posterior obtained by FAVI is better calibrated, it still needs correction. As multiple posterior approximators were considered, CANVI had to be applied with recalibration. Table 2.1 shows that the recalibrated $1 - \alpha$ prediction regions are nearly perfectly calibrated. Importantly, correction by CANVI can result in either larger or smaller $1 - \alpha$ regions, depending on the direction of miscalibration. In settings where the variational posterior is overdispersed, applying CANVI results in smaller $1 - \alpha$ density regions, explicitly shown in Section A.9.7. Table 2.1 also demonstrates using the better calibrated $\mathcal{L}_{\text{FAVI}}$-trained approximation results in higher efficiency compared to the $\mathcal{L}_{\text{ELBO}}$ and $\mathcal{L}_{\text{IWBO}}$ counterparts.

We also demonstrate the necessity of using mixed sampling for $\widehat{\ell}$. In particular, we compare the estimates when using only the variational posterior as a sampler ($\widehat{\ell}_{K=1}$) and when averaging 10 mixed samplers ($\widehat{\ell}_{K=10}$), from which we observe the estimator accuracy greatly improves with mixing, especially in the underdispersed IWBO and ELBO cases.

2.4.3 Galaxy Spectral Energy Distributions

We now present the application of CANVI to an important scientific problem. The spectrum of an astronomical object is measured via a spectrograph, which records the flux across a large grid of wavelength values [York et al., 2000, Abareshi et al., 2022], which we simulate with the Probabilistic Value-Added Bright Galaxy Survey simulator (PROVABGS). PROVABGS maps $\theta \in \mathbb{R}^{11}$ to galaxy spectra, detailed further and visualized in Appendix A.10.

A mixture of 20 Gaussian distributions was used as the variational posterior and trained against the $\mathcal{L}_{\text{FAVI}}, \mathcal{L}_{\text{ELBO}},$ and $\mathcal{L}_{\text{IWBO}}$ objectives, as in Section 2.4.2.2. Table 2.1 shows that the ELBO and IWBO tend to be overly concentrated, failing to contain the entire parameter vector θ in the $1 - \alpha$ highest-density region often. FAVI, on the other hand, is reasonably well-calibrated. After applying CANVI, all three methods achieve nearly perfect calibration across a range of desired confidence levels. Of course, the utility of these corrected regions depends on the level of information contained in the original model. For the ELBO or IWBO cases, the corrected regions achieve statistical validity, but are too large to be informative. Notably, the underdispersion of the ELBO approximator led to $\widehat{q}_{\mathcal{C}}(0.05) = 0$ (up to machine precision), resulting in a volume estimate of ∞. For FAVI, on the other hand, application of CANVI results in statistical guarantees with minimal alterations to the high-density regions.

CHAPTER 3

Conformal Contextual Robust Optimization

> If a machine is expected to be infallible, it cannot also be intelligent.
>
> Alan Turing

In this chapter, we extend our discussion on leveraging conformal prediction for decision-making from the scientific context to the more general predict-then-optimize setting. In doing so, we additionally demonstrate how the implicitly defined, non-convex prediction regions of conformal prediction can be practically leveraged. Data-driven approaches to predict-then-optimize decision-making problems seek to mitigate the risk of uncertainty region misspecification in safety-critical settings. Current approaches, however, suffer from considering overly conservative uncertainty regions, often resulting in suboptimal decision-making. To this end, we propose Conformal-Predict-Then-Optimize (CPO), a framework for leveraging highly informative, nonconvex conformal prediction regions over high-dimensional spaces based on conditional generative models, which have the desired distribution-free coverage guarantees. Despite guaranteeing robustness, such black-box optimization procedures alone inspire little confidence owing to the lack of explanation of why a particular decision was found to be optimal. We, therefore, augment CPO to additionally provide semantically meaningful visual summaries of the uncertainty regions to give qualitative intuition for the optimal decision.

3.1 Introduction

Predict-then-optimize or contextual robust optimization problems are of long-standing interest in safety-critical settings where decision-making happens under uncertainty [Sun et al., 2023, Elmachtoub and Grigas, 2022, Elmachtoub et al., 2020, Peršak and Anjos, 2023]. In traditional robust optimization, results are made to be robust to distributions anticipated to be present upon deployment [Ben-Tal et al., 2009, Beyer and Sendhoff, 2007]. Since such decisions are sensitive to proper model specification, recent efforts have sought to supplant this with data-driven uncertainty regions [Cheramin et al., 2021, Bertsimas et al., 2018, Shang and You, 2019, Johnstone and Cox, 2021].

Model misspecification is ever more present in *contextual* robust optimization, spurring efforts to define similar data-driven uncertainty regions [Ohmori, 2021, Chenreddy et al., 2022, Sun et al., 2023]. Such methods, however, focus on box- and ellipsoid-based uncertainty regions, both of which are necessarily convex and often overly conservative, resulting in suboptimal decision-making.

Conformal prediction provides a principled framework for producing distribution-free prediction regions with marginal frequentist coverage guarantees [Angelopoulos and Bates, 2021, Shafer and Vovk, 2008]. By using conformal prediction on a user-defined score function $s(x, y)$ and obtaining an empirical $1 - \alpha$ quantile $\widehat{q}(\alpha)$ of $s(x, y)$ over a calibration set $\mathcal{D}_{\mathcal{C}}$, prediction regions $\mathcal{C}(x) = \{y \mid s(x, y) \leq \widehat{q}(\alpha)\}$ attain marginal coverage guarantees. Such prediction regions, however, are notably defined *implicitly*. For simple scores, such as residuals, an explicit expression of such regions can be written, making these the most common approaches used in practice [Tumu et al., 2023, Horwitz and Hoshen, 2022, Angelopoulos et al., 2022, Hu et al., 2022, Mao et al., 2022].

The disadvantage is that such score functions ignore the structure that is often present in high-dimensional data, such as images. Choices of simplistic scores, thus, tend to be overly conservative and often produce convex prediction regions even when $\mathcal{P}(Y|X)$ is non-convex. Recent work has demonstrated that defining scores using conditional generative models produces sharper and, hence, more informative prediction regions [Feldman et al., 2023a, Wang et al., 2022, Patel et al., 2023]. We, thus, extend the line of data-driven predict-then-optimize work by considering such generative model-based prediction regions.

In addition to contributing to the predict-then-optimize line of inquiry, we view this work as addressing a concern of the conformal prediction community: how to use implicitly defined non-convex, high-dimensional prediction regions. Works producing such regions have themselves noted the difficulty in their use [Sesia and Romano, 2021, Izbicki et al., 2022]. Initial works on coverage for images have framed the utility of their results in highlighting regions of the image with

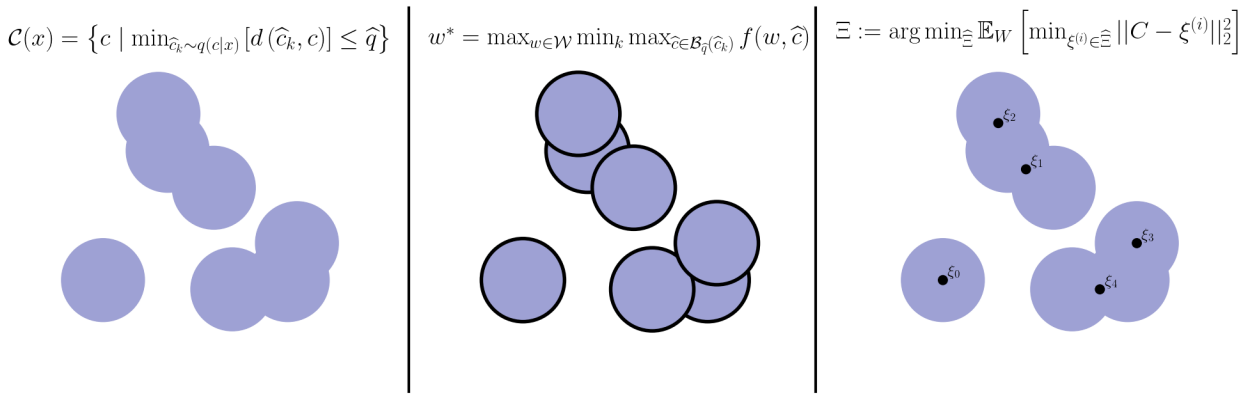

Figure 3.1: CPO leverages informative, non-convex conformal prediction regions for robust predict-then-optimize decision making. CPO uses a score function such that the resulting prediction regions can be decomposed into convex subregions over which optimization can be carried out efficiently. Visual summaries $\{\xi^{(i)}\}$ of the prediction region can similarly be efficiently sampled to gain intuition on the optimal decision w^*.

the greatest variability and, hence, uncertainty [Angelopoulos et al., 2022, Horwitz and Hoshen, 2022, Belhasin et al., 2023].

Extending such visualization gives invaluable intuition to the end user. For instance, a black-box optimization procedure for producing drug candidates to robustly bind to a predicted protein structure offers little insight into the decision-making process; however, semantic summaries of the uncertainty region would reveal regions of flexibility of the protein, clarifying why particular structures were deemed optimal in the candidate drug. Such interest in explainable robust decision-making was highlighted in a recent survey [Sadana et al., 2023], especially given the "right to explanation" mandated by the EU's "General Data Protection Regulation" [Doshi-Velez and Kim, 2017, Kaminski, 2019].

Our main contributions, thus, are:

- Proposing Conformal-Predict-Then-Optimize (CPO) to leverage informative, non-convex prediction regions for decision-making.
- Providing interpretable visual summaries of uncertainty regions using representative points.
- Demonstrating the generality of CPO across a suite of benchmark tasks and a traffic routing task based on probabilistic weather prediction.

3.2 Method

We now propose CPO, a way to perform robust predict-then-optimize decision-making over informative, non-convex prediction regions based on generative models. We then discuss how to construct visual summaries of the contents of the conformal prediction regions using a collection of N representative points.

3.2.1 CPO: Problem Formulation

Let $c \in \mathcal{C}$, where $(\mathcal{C}, d)$ is a general metric space, and $\mathcal{F}$ be the σ-field of $\mathcal{C}$. While the standard predict-then-optimize framework assumes a linear objective function $c^T w$, we consider general convex-concave objective functions $f(w, c)$ that are L-Lipschitz in c under the metric d for any fixed w, as follows:

$$\begin{aligned} w^*(x) := \min_{w, \mathcal{U}} \max_{\widehat{c} \in \mathcal{U}(x)} \quad & f(w, \widehat{c}) \\ \text{s.t.} \quad & \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{3.1}$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{F}$ is a uncertainty region predictor. Exact solution of this problem is intractable, as no practical methods exist to optimize over the predictor function space $\mathcal{U}$. Practical solution of this optimization problem, thus, involves optimizing over several prespecified uncertainty region predictors $\{\mathcal{U}_i\}_{i=1}^N$. For any fixed $\mathcal{U}$, this robust counterpart to the nominal predict-then-optimize problem produces a valid upper bound if $c \in \mathcal{U}(x)$. Denoting the pessimism gap as $\Delta(x, c) := \min_w \max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - \min_w f(w, c)$, we clearly see $\Delta(x, c) \geq 0$ if $c \in \mathcal{U}(x)$, formalized below.

Lemma 3.2.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$. Then,*

$$\mathcal{P}_{X,C}\left(0 \leq \Delta(X, C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \tag{3.2}$$

The proof is deferred to Section B.1. Thus, $1 - \alpha$ validity of $\mathcal{U}$ ensures the RO procedure produces a valid bound with probability $1 - \alpha$, with more efficient prediction regions resulting in tighter upper bounds.

3.2.2 CPO: Score Function

We assume a conditional generative model $q(C \mid X)$ is learned for this prediction task. For most score functions, the min-max optimization problem of Equation (3.1) is computationally intractable. Crucially, however, we can consider an extension to the score proposed in [Wang et al., 2022], which lends itself to a decomposition under which such optimization becomes tractable. For

a fixed K and $\{\widehat{c}_k\}_{k=1}^K \sim q(C \mid x)$, let

$$s(x, c) = \min_k \left[d\left(\widehat{c}_k, c\right) \right]. \tag{3.3}$$

We refer to this score as "Generalized Probabilistic Conformal Prediction," (GPCP) whose validity follows from that of the original PCP framework [Wang et al., 2022]. We discuss the selection of K in Section 3.2.4.

3.2.3 CPO: Optimization Algorithm

We fix $\alpha \in [0, 1]$ and take $\mathcal{U}(x)$ to be the $1 - \alpha$ prediction region $\mathcal{C}(x)$. Let $\phi(w) := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. It follows that $\phi(w)$ is convex by Danskin's Theorem by assumption of the convexity of f in w. Exact solution of the min-max problem, thus, follows using standard gradient-based optimization techniques on $\phi(w)$. By Danskin's Theorem, $\nabla_w \phi(w) = \nabla_w f(w, c^*)$, where $c^* := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. We follow the standard projected gradient descent optimization scheme, projecting into $\mathcal{W}$ at each iterate, denoted by $\Pi_{\mathcal{W}}$.

Efficient solution of this RO problem, therefore, reduces to being able to efficiently solve the maximization problem over $\mathcal{C}(x)$. While challenging over general nonconvex regions, the GPCP score formulation lends itself to a highly structured prediction region, namely of the form $\mathcal{C}(x) = \bigcup_{k=1}^K \mathcal{B}_{\widehat{q}}(\widehat{c}_k)$ with $\mathcal{B}_{\widehat{q}}$ being a ball of radius $\widehat{q}$, the conformal quantile, under the d metric. This decomposition of $\mathcal{C}(x)$ means the maximum can be efficiently computed by aggregating the maxima over the individual balls:

$$\max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c}) = \max_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}), \tag{3.4}$$

where the maximum over a ball can be efficiently computed with traditional convex optimization techniques. This procedure is summarized in Algorithm 3. The convergence of this procedure proceeds as follows, whose proof is deferred to Section B.2.

Lemma 3.2.2. *Let $\phi(w) := \max_{\widehat{c} \in \bigcup_{k=1}^K \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$ for $\{\widehat{c}_k\}_{k=1}^K \subset \mathcal{C}$, $\widehat{q} \in \mathbb{R}^+$, and $f(w, c)$ convex-concave and L-Lipschitz in c for any fixed w. Let $w^* \in \mathcal{W}$ be a minimizer of ϕ. For any $\epsilon > 0$, define $T := \frac{L^2 ||w_0 - w^*||}{\epsilon^2}$ and $\eta := \frac{||w_0 - w^*||}{L\sqrt{T}}$. Then the iterates $\{w_t\}_{t=0}^T$ returned by Algorithm 3 satisfy*

$$\phi\left(\frac{1}{T+1} \sum_{t=0}^{T} w_t\right) - \phi(w^*) \leq \epsilon. \tag{3.5}$$

Algorithm 3 CPO-OPT

procedure CPO-OPT
Inputs: Context x, CGM $q(C \mid X)$, Optimization steps T, Score samples K, Conformal quantile $\widehat{q}$
  $w \sim U(\mathcal{W}), \{\widehat{c}_k\}_{k=1}^{K} \sim q(C \mid x)$
  for $t \in \{1, \dots T\}$ **do**
    $\left\{c_k^* \leftarrow \arg\max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})\right\}_{k=1}^{K}$
    $c^* \leftarrow \arg\max_{c_k^*} f(w, c_k^*)$
    $w \leftarrow \Pi_{\mathcal{W}}(w - \eta \nabla_w f(w, c^*))$
  end for
  Return w
end procedure

3.2.4 CPO: K Selection

Crucially, the convergence highlighted in Theorem 3.2.2 reveals that the number of "outer" iterations (i.e. T) has no dependence on K. This is apparent from the proof, in which the iterate count T hinges upon the Lipschitz constant of $\phi(w) = \max_k \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}) := \max_k \phi_k(w)$, which critically is L-Lipschitz *regardless* of what K is selected, as each $\phi_k(w)$ is L-Lipschitz.

We can, thus, solely focus attention on the impact the choice of K has on the "inner" optimization computational cost, namely $\max_k \phi_k(w)$. This linearly increasing cost with K, however, must be juxtaposed with the improved *statistical* efficiency of such prediction regions. In particular, [Wang et al., 2022] empirically demonstrated region size generally decreased nonlinearly up to a saturation point as a function of K.

Critically, this inflection point can be determined *prior* to performing the optimization, since doing so only requires access to $q(C \mid X)$ and test samples to estimate the prediction region size. As pointed out in [Wang et al., 2022] and proven in [Chan, 2008], estimation of the volume of a union of hyperspheres is complicated by the need to account for overlapped regions. K is, thus, chosen based on Monte Carlo estimates of the prediction region volume using Voronoi cells of the hypersphere centers given by [Edelsbrunner, 1995]:

$$\widehat{\ell}(\{\mathcal{B}_{\widehat{q}}(\widehat{c}_k)\}) := |\mathcal{B}_{\widehat{q}}| \sum_{k=1}^{K} \mathcal{P}_{C \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))}(C \in V(\widehat{c}_k)), \tag{3.6}$$

where $C \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$ denotes a random variable defined uniformly over the region associated with $\widehat{c}_k$, $|\mathcal{B}_{\widehat{q}}|$ the volume of a hypersphere of radius $\widehat{q}$, and $V(\widehat{c}_k)$ the Voronoi cell of $\widehat{c}_k$, defined as $\{z \in \mathbb{R}^d \mid d(\widehat{c}_k, z) \leq d(\widehat{c}_{k'}, z), k' \neq k\}$. Muller's method enables efficient sampling of $U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$ [Muller, 1959, Fishman, 2013].

We then choose K^* to be the inflection point, namely the $\arg\min_K |\widehat{\ell}_K - \widehat{\ell}_{K+1}| \leq \epsilon$ for some user-specified ϵ volume tolerance. Critically, these volume estimates must be performed on a distinct subset of the data from $\mathcal{D}_{\mathcal{C}}$ as exchangeability with future test points is otherwise lost in conditioning on $\mathcal{D}_{\mathcal{C}}$ for selecting K^* [Yang and Kuchibhotla, 2021]. We, thus, partition $\mathcal{D}_{\mathcal{C}} := \mathcal{D}_{\mathcal{C}_1} \cup \mathcal{D}_{\mathcal{C}_2}$, using $\mathcal{D}_{\mathcal{C}_1}$ for calibration and $\mathcal{D}_{\mathcal{C}_2}$ for volume estimation, detailed in Algorithm 4.

Algorithm 4 CPO

procedure VOLUMEEST
Inputs: Context x, CGM $q(C \mid X)$, Conformal quantile $\widehat{q}$
  $\{\widehat{c}_k\}_{k=1}^K \sim q(C_{1:K} \mid x)$
  $\left\{\{c_{k,m}\}_{m=1}^M \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))\right\}_{k=1}^K$
  Return $|B_{\widehat{q}}| \sum_{k=1}^K \frac{1}{M} \sum_{m=1}^M \mathbb{1}\left[c_{k,m} \in V(\widehat{c}_k)\right]$
end procedure

procedure CPO
Inputs: Context x, CGM $q(C \mid X)$, Optimization steps T, Desired coverage $1-\alpha$, Max samples $K_{\max}$, Volume Tolerance ϵ, Calibration sets $\mathcal{D}_{\mathcal{C}_1}, \mathcal{D}_{\mathcal{C}_2}$
  for $K \in \{1, \dots K_{\max}\}$ **do**
    $s_K(x, c) \leftarrow \min_{\widehat{c}_k \in \{\widehat{c}_i\} \sim q(C_{1:K}|x)} \left[d\left(\widehat{c}_k, c\right)\right]$
    $\mathcal{S}_K \leftarrow \left\{s_K(x^{(i)}, c^{(i)}) \mid (x^{(i)}, c^{(i)}) \in \mathcal{D}_{\mathcal{C}_1}\right\}$
    $\widehat{q}_K \leftarrow \frac{\lceil (|\mathcal{D}_{\mathcal{C}_1}|+1)(1-\alpha) \rceil}{|\mathcal{D}_{\mathcal{C}_1}|}$ quantile of $\mathcal{S}_K$
    $\widehat{\ell}_K \leftarrow \frac{1}{|\mathcal{D}_{\mathcal{C}_2}|} \sum_{i=1}^{|\mathcal{D}_{\mathcal{C}_2}|}$ VOLUMEEST$(x^{(i)}, q, \widehat{q}_K)$
  end for
  $K^* \leftarrow \arg\min_K \left|\widehat{\ell}_K - \widehat{\ell}_{K+1}\right| \leq \epsilon$
  Return CPO-OPT$(x, q, T, K^*, \widehat{q}_{K^*})$
end procedure

3.2.5 CPO: Representative Points

We now frame the problem of summarizing the prediction region $\mathcal{C}(x)$. We critically note that this issue of interpretability is non-existent in traditional approaches to robust predict-then-optimize, where uncertainty regions are interpretable by construction, being balls around nominal estimates $\mathcal{B}_\epsilon(\widehat{c})$. In other words, there is a fundamental tension in qualitative interpretability and the expressiveness of uncertainty regions, requiring a bespoke method for recovering intuition when leveraging conformal prediction regions. Formally, for a user-specified number of summary points N and query x, we seek

$$\Xi(x) := \arg\min_{\widehat{\Xi} \in \zeta} \mathbb{E}_{C \sim U(\mathcal{C}(x))} \left[\min_{\widehat{\xi}^{(i)} \in \widehat{\Xi}} d(C, \xi^{(i)})\right]. \tag{3.7}$$

We use the shorthand $d(C, \Xi) := \min_{\xi^{(i)} \in \Xi} d(C, \xi^{(i)})$. In other words, we wish to construct representative points for a uniform sampling of the prediction region. A naive approach would simply involve explicitly gridding the output space $\mathcal{C}$, filtering such points with the rejection criterion of $\mathcal{C}(x)$, and clustering the remaining points per the d metric. This, however, is intractable in high-dimensional cases. Thus, a sampling method is employed to circumvent gridding, paralleling the technique leveraged for volume estimation.

M samples are initially drawn $\{c_{k,m}\}_{m=1}^{M} \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))$ for each k. Importantly, such uniform sampling of the balls leads to *non*-uniform sampling over $\mathcal{C}(x)$ if naively aggregated across k, as overlapped regions will be more densely sampled. For this reason, we subsample by discarding those samples $c_{k,m}$ for which $c_{k,m} \in V(\widehat{c}_{k'})$ for $k \neq k'$. This results in samples $C := \{c_i\}$ drawn from the desired $U(\mathcal{C}(x))$.

RPs must be aggregated separately for each connected subregion of $\Omega_\ell \subset \mathcal{C}(x)$ to ensure each $\xi^{(i)} \in \mathcal{C}(x)$. That is, we must identify $C_\ell := C \cap \Omega_\ell$. To do so, we determine if two points (c_i, c_j) belong to the same Ω_ℓ by considering the corresponding connected components problem defined on the graph induced by the edge criterion $e_{i,j} = \mathbb{1}[d(c_i, c_j) < \widehat{q}]$. For each C_ℓ, we find a subset $N_\ell := N(|C_\ell|/|C|)$ of the total N representative points, for which we use K-MEANS++ with the d metric. The full procedure is in Algorithm 5.

Algorithm 5 CPO-RPS: QUERYBALL$(\mathcal{T}, x, r)$ is an assumed subroutine that returns all points in the kd tree $\mathcal{T}$ that are within a radius r of x.

procedure CPO-RPS
Inputs: Context x, CGM $q(C \mid X)$, RP count N, Conformal quantile $\widehat{q}$
  $\{\widehat{c}_k\}_{k=1}^{K} \sim q(C_{1:K} \mid x)$
  $\left\{\{c_{k,m}\}_{m=1}^{M} \sim U(\mathcal{B}_{\widehat{q}}(\widehat{c}_k))\right\}_{k=1}^{K}$
  $C \leftarrow \{c_{k,m} \mid c_{k,m} \in V(\widehat{c}_k)\}_{k=1,m=1}^{K,M}$
  $\mathcal{T} \leftarrow$ KD-TREE(C)
  $\mathcal{E} \leftarrow \bigcup_i \{c_i \times$ QUERYBALL$(\mathcal{T}, c_i, \widehat{q}) \mid c_i \in \mathcal{T}\}$
  $\{C_\ell\} \leftarrow$ CONNECTEDCOMPONENTS$(\mathcal{G}(C, \mathcal{E}))$
  $\Xi \leftarrow \bigcup_{\ell=1}^{L} \{$K-MEANS++$(C_\ell, N\left(\frac{|C_\ell|}{|C|}\right), d)\}$
  Return Ξ
end procedure

3.2.6 CPO: Projection

After obtaining Ξ, further insight can be gleaned by exploring the local projection around each $\xi^{(i)}$. An example of this is visualizing the road-level variability in traffic predictions from uncertainty in upstream weather predictions, shown in Figure 3.5. To do this, we visualize the extent of the Voronoi cell $V^{(i)} \subset \mathcal{C}(x)$ associated with $\xi^{(i)}$ along the $\mathcal{C}$ space dimensions. That is, for each

Voronoi cell, we visualize the Frechet variance along the projections $\{\pi_j\}_{j=1}^J$, where $J = \dim(\mathcal{C})$. Such projections preserve the structure of the objects being modeled, making them visually interpretable. For instance, π_j in the traffic example corresponds to the projection of $V^{(i)}$ to a *single* road j. Similarly, π_j would project to a single atom for a molecular reconstruction task. Formally,

$$\left|V_j^{(i)}\right| := \sum_{c \in V^{(i)}} d^2(\pi_j(c), \pi_j(\xi^{(i)})). \tag{3.8}$$

3.3 Experiment

We now demonstrate the utility of the CPO framework. Code is available at https://github.com/yashpatel5400/csi.

3.3.1 SBI: Fractional Knapsack

We first study the fractional knapsack problem under various complex contextual mappings, namely

$$\begin{aligned} w^*(x) := \min_{w, \mathcal{U}} \max_{\widehat{c} \in \mathcal{U}(x)} \quad & -\widehat{c}^T w \\ \text{s.t.} w \in [0,1]^n, p^T w \leq B, & \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{3.9}$$

where $p \in \mathbb{R}^n$ and $B > 0$. The distributions $\mathcal{P}(C)$ and $\mathcal{P}(X \mid C)$ are taken to be those from various simulation-based inference (SBI) benchmark tasks provided by [Hermans et al., 2021b], chosen as they have $\mathcal{P}(C \mid X)$ with complex structure. We specifically study Two Moons, Lotka-Volterra, Gaussian Linear Uniform, Bernoulli GLM, Susceptible-Infected-Recovered (SIR), and Gaussian Mixture, fully described in Section A.6. We note that, while these particular distributions have little semantic meaning in the traditional context of fractional knapsack, this experiment highlights the capacity for CPO to succeed even for complex distributions, which we leverage in a more semantically meaningful case in Section 3.3.2.

3.3.1.1 SBI: Quantitative Assessment

We first demonstrate the quantitative improvement in decision-making from leveraging CPO over the box- (PTC-B) and ellipsoid-based (PTC-E) regions proposed in [Sun et al., 2023], as well as box- and ellipsoid-based sets constructed based solely on observations of $\mathcal{P}(C)$, i.e. where we ignore x, referred to as Box and Ellipsoid. For CPO, we use

$$s(x, c) = \min_k ||\widehat{c}_k - c||_2^2. \tag{3.10}$$

Table 3.1: Coverages across tasks for $\alpha = 0.05$ are shown in the left table, where coverage was assessed over a batch of 1,000 i.i.d. test samples. Objective optima are shown in the right table, averaged over a batch of 10 i.i.d. test samples with standard deviations in parentheses. The nominal optima are included as reference points.

	Box	PTC-B	Ellipsoid	PTC-E	CPO
Gaussian Uniform	0.94	0.96	0.95	0.95	0.95
Gaussian Mixture	0.95	0.93	0.94	0.93	0.94
Bernoulli GLM	0.96	0.95	0.95	0.94	0.94
Lotka Volterra	0.95	0.96	0.94	0.94	0.95
SIR	0.94	0.95	0.93	0.95	0.93
Two Moons	0.93	0.94	0.94	0.94	0.96

Box	PTC-B	Ellipsoid	PTC-E	CPO	Nominal
0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	**-0.27 (0.35)**	**-0.43 (0.4)**	*-4.48 (0.56)*
0.0 (0.0)	**-6.6 (1.67)**	0.0 (0.0)	**-7.38 (1.78)**	**-7.77 (1.87)**	*-11.66 (1.23)*
0.0 (0.0)	**-0.18 (0.49)**	0.0 (0.0)	**-0.06 (0.25)**	**-0.18 (0.37)**	*-3.53 (0.27)*
-0.52 (0.02)	-0.05 (0.24)	-0.02 (0.0)	-0.22 (0.18)	**-0.68 (0.26)**	*-1.88 (0.01)*
-0.16 (0.02)	-0.22 (0.09)	-0.08 (0.01)	-0.22 (0.06)	**-0.38 (0.05)**	*-0.52 (0.02)*
0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	**-0.15 (0.11)**	*-0.38 (0.01)*

$q(\widehat{c} \mid x)$ was taken to be a neural spline normalizing flow [Durkan et al., 2019] trained with FAVI [Ambrogioni et al., 2019]. Visualizations of the exact and variational posteriors are provided in Section A.9. Ks were chosen by studying the inflections of the prediction region volume estimate under each distributional setup, with $|\mathcal{D}_{\mathcal{C}_{1,2}}| = 1000$, seen in Figure 3.2. Inflection points were around $K = 10$ for most setups.

For assessing coverage and the robust objective value, we sampled $|\mathcal{D}_{\mathcal{T}}| = 1000$ test points i.i.d. from $\mathcal{P}(X, C)$. Coverage was assessed across all 1000 samples by measuring the proportion of samples for which $s(x^{(i)}, c^{(i)}) \leq \widehat{q}$. For assessing the objective, optimization was performed across 10 samples, with $p \sim U([0, 1000]^n)$, $u \sim U(0, 1)$, and $B \sim U(\max_i p_i, \sum_i p_i - u \max_i p_i)$ sampled per run.

The results are seen in Table 3.1. We include the nominal optima as a reference, i.e. $\min_w -c^T w$ for the *true* c. Recall that, by Theorem 3.2.1, with proper $\mathcal{U}(x)$, the robust objective values should be valid upper bounds on the nominal optima, with more conservative regions resulting in more vacuous bounds. We see this as, although all approaches result in valid coverage guarantees and hence produce valid upper bounds, the overly conservative nature of alternate regions results in their consistent looseness compared to CPO. Notably, these differences are more accentuated in cases where $\mathcal{P}(C|X)$ has complex structure; level sets under the Gaussian Linear, Gaussian Mixture, and Bernoulli GLM cases are roughly ellipsoidal, seen in Section A.9, resulting in comparable performance between CPO and PTC-E. Thus, as discussed and highlighted in Section 3.3.2, the benefits of CPO primarily manifest under difficult-to-model contextual distributions, where sets for simple geometries become overly large.

3.3.1.2 SBI: Representative Point Recovery

We next demonstrate that Algorithm 5 can approximately recover RPs for such uncertainty regions, leveraged to glean insights in the modeling task of Section 3.3.2. Notably, RPs are not unique; for instance, any rigid rotation of Ξ for a uniform distribution over a 2D ball results in a distinct yet

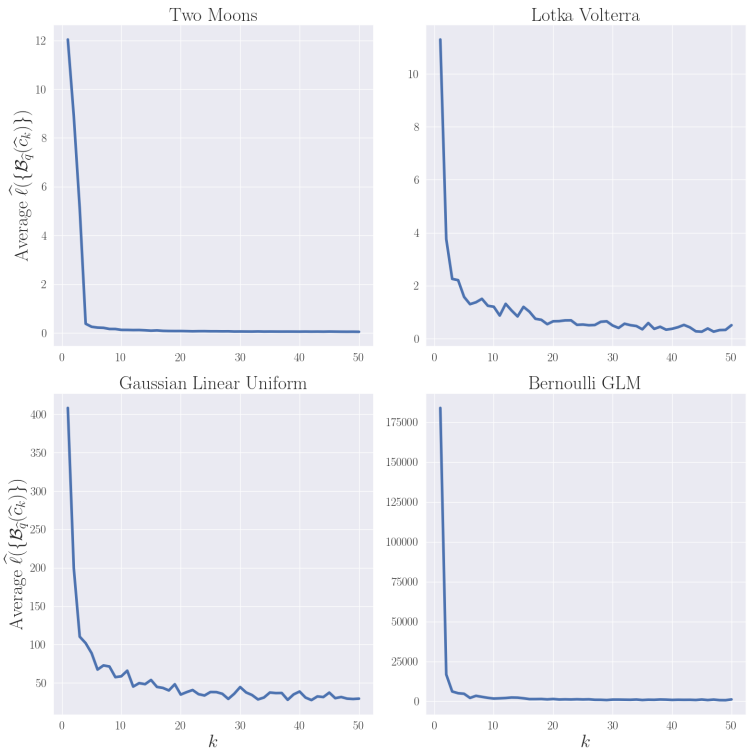

Figure 3.2: Average volume estimates $\widehat{\ell}(\{\mathcal{B}_{\widehat{q}}(\widehat{c}_k^{(i)})\})$ over $x^{(i)} \in \mathcal{D}_{\mathcal{C}_2}$ across SBI benchmarks.

optimal set $\widehat{\Xi}$ of RPs. The RP objective minimum, however, is unique, meaning suboptimality can be assessed by measuring

$$\Delta(\Xi, \widehat{\Xi}) := \mathbb{E}_{C \sim U(\mathcal{C}(x))} \left[d(C, \widehat{\Xi}) - d(C, \Xi) \right]. \tag{3.11}$$

$N = 5$ representative points were produced per setup. To compute Ξ, a grid discretization over the space was performed followed by a clustering for each connected component of this discretization. That is, the support $\mathcal{C}$ was discretized into 60 bins per dimension. Each discretized point c_k was assessed for membership in $\mathcal{C}(x)$, resulting in a collection of points C, from which we could recover Ξ in the manner described in Section 3.2.5. Visualizations of the exact and approximate RPs are provided for tasks where $\mathcal{C} \subset \mathbb{R}^2$ in Section B.3.

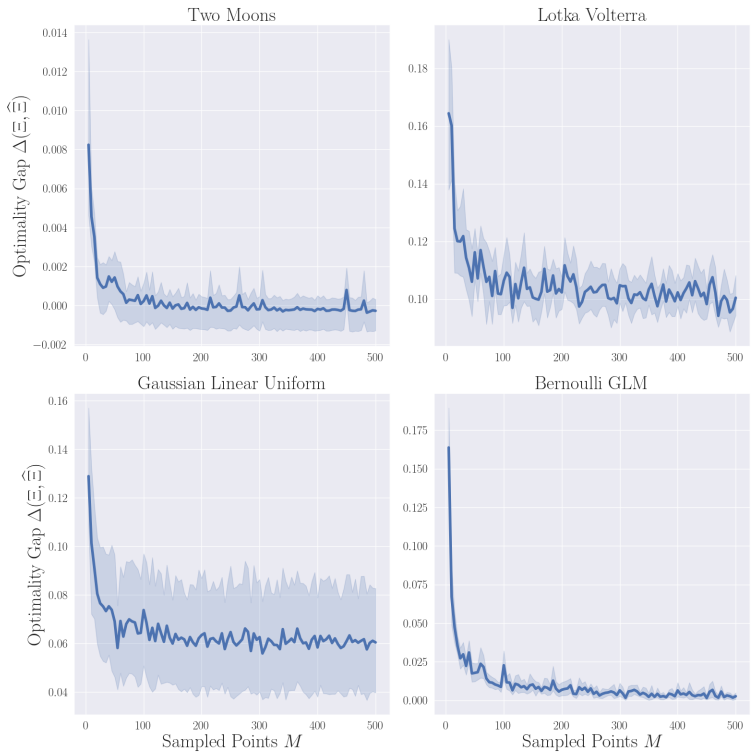

Figure 3.3: Suboptimality of the approximate representative points $\Delta(\Xi, \widehat{\Xi})$ decreases over increased sampling from the conformal prediction region.

To make explicit discretization possible, problems were projected into lower-dimensional versions, namely $\mathcal{C} \subset \mathbb{R}^4$. Figure 3.3 demonstrates the suboptimality of $\widehat{\Xi}$ decreases with increasing samples. Of note is that this convergence is slower in higher dimensional problems: for low dimensional cases, recovery of optimal RPs happens for small M, meaning any fluctuations thereafter are noise, as seen in the Two Moons case.

3.3.2 Robust Vehicle Routing

Optimal routing is a long-standing point of interest in the operations research community, with widespread applications such as in resource distribution and urban traffic flow management [Mor and Speranza, 2022, Saberi and Verbas, 2012, Okulewicz and Mańdziuk, 2019, Kořenář, 2003]. We study the traffic flow problem from [Angelelli et al., 2021].

Recent work has demonstrated the utility of generative models in quantifying uncertainty for weather predictions over traditional physics-based approaches [Agrawal et al., 2019, Ayzel et al., 2020, Franch et al., 2020, Shi et al., 2017]. We specifically leverage a latent diffusion model for such forecasting from [Leinonen et al., 2023]. Formally, a forecaster $\mathcal{P}(\widetilde{Y} \mid x)$ maps precipitation readings from radar networks $x \in \mathbb{R}^{T \times W \times H}$, specifically over T time steps with resolutions $W \times H$,

Table 3.2: Coverage was assessed over 128 i.i.d. test samples and average objective optima over 10 i.i.d. test samples with standard deviations in parentheses.

	Box	PTC-B	Ellipsoid	PTC-E	CPO	*Nominal*
Coverage	0.94	0.93	0.94	0.92	0.94	—
Objective	7863.45 (0.0)	34559.03 (171.3)	7038.77 (0.0)	8807.68 (4.22)	**4171.22 (321.34)**	*299.50 (0.0)*

to $\widetilde{Y} \in \mathbb{R}^{W \times H}$, the precipitation for some fixed ΔT point beyond x.

We consider the robust traffic flow problem (RTFP) for a source-target pair (s, t) over the network graph of Manhattan, where $|\mathcal{V}| = 4584$ and $|\mathcal{E}| = 9867$. The precipitation $\widetilde{Y}$ was combined with the nominal speed limits to obtain the final travel costs c along edges, fully described in Section B.4. Formally, we seek

$$w^*(x) := \min_{w} \max_{\widehat{c} \in \mathcal{U}(x)} \widehat{c}^T w \tag{3.12}$$
$$\text{s.t.} w \in [0,1]^{\mathcal{E}}, Aw = b, \mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$$

where w_e represents the proportion of traffic routed along edge e, $C \in \mathbb{R}^{|\mathcal{E}|}$ is the edge weight vector, $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}|}$ is the node-arc incidence matrix, and $b \in \mathbb{R}^{|\mathcal{V}|}$ has entries $b_s = 1, b_t = -1$, and $b_k = 0$ for $k \notin \{s, t\}$.

We again demonstrate the quantitative improvement in decision-making resulting from using the more informative CPO prediction regions. Experiments were conducted with s and t chosen uniformly at random from $\mathcal{V}$. We take the score as defined in Equation (3.10) on the edge weight space rather than the initial precipitation map space. Results are shown in Table 3.2. Again, although all approaches achieve coverage guarantees, bounds resulting from alternate regions are significantly looser compared to those from CPO. This is especially prominent in this task compared to those of Section 3.3.1 due to the high dimension of the prediction space($\mathbb{R}^{|\mathcal{E}|}$) and complex nature of $\mathcal{P}(C|X)$.

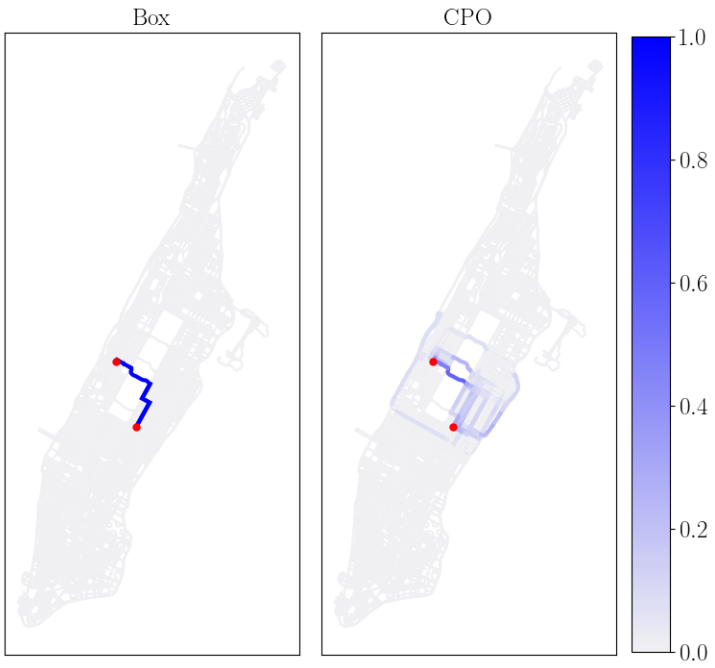

Figure 3.4: Solutions for the RTFP under the Box (left) and CPO (right) uncertainty regions.

Notably, the formulation in Equation (3.12) is a relaxation of the standard LP formulation of the robust shortest paths problem (RSPP), in which $\mathcal{W} = \{0, 1\}^{\mathcal{E}}$. Given that A is a totally unimodular matrix, the solutions of the *box*-constrained RTFP and RSPP are equivalent, i.e. for both Box and PTC-B; they, however, are *not* equivalent under more general constraint sets [Chaerani et al., 2005], i.e. Ellipsoid, PTC-E, and CPO, resulting in the observed suboptimality of box constraints. This is highlighted in Figure 3.4, where the Box constraint results in a fully concentrated allocation of traffic along a single path.

Despite apparent quantitative improvements resulting from the CPO optimal solution, it is difficult to directly understand *why* such allocations were deemed optimal without a qualitative impression of $\mathcal{U}(x)$, as framed in Section 3.2.5.

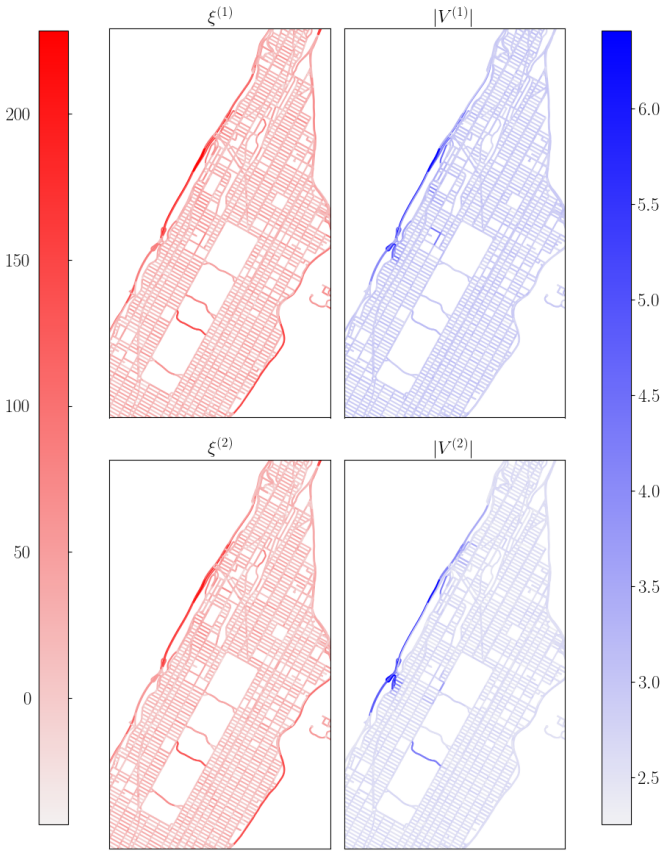

Figure 3.5: Two RPs for $\mathcal{C}(x)$ for travel time prediction (left) and the extents of their Voronoi cells (right).

We, therefore, now construct $N = 5$ representative points and their corresponding projections, two of which are visualized in Figure 3.5. The RPs highlight the multimodal nature of the edge weights distribution, where $\xi^{(1)}$ exhibits a case of precipitation more heavily concentrating along the northeast corridor across Manhattan and $\xi^{(2)}$ one where it concentrates on the west. In addition, the projection around $\xi^{(2)}$ reveals especially high uncertainty on the path through Central Park with less on surrounding roads. CPO, thus, hedges its allocation in Figure 3.4 more evenly across paths, unlike the concentrated allocation under the Box region.

CHAPTER 4

Applications of Conformal Decision Making

> A good decision is based on knowledge
> and not on numbers.
>
> Plato

We now present two applications of the conformal decision-making framework developed over the previous chapter. In the first application, we demonstrate how this framework can be generalized to cases where the predictor is an ensemble of models and is conformalized via an extension to the standard conformal prediction pipeline known as "conformal score aggregation (CSA)." In CSA, the standard scalar score function is replaced by a vector-valued function and the quantile threshold, consequently, replaced by a "quantile envelope." We empirically demonstrate that this interfacing with CSA improves the conservatism of the resulting conformal decision-making.

We then explore the application of this decision-making to settings of robust linear control. End-to-end engineering design pipelines, in which designs are evaluated using concurrently defined optimal controllers, are becoming increasingly common in practice. To discover designs that perform well even under the misspecification of system dynamics, such end-to-end pipelines have now begun evaluating designs with a robust control objective. Current approaches of specifying such robust control subproblems, however, rely on hand specification of perturbations anticipated to be present upon deployment or margin methods that ignore problem structure, resulting in a lack of theoretical guarantees and overly conservative empirical performance. We, instead, propose a novel methodology for LQR systems that leverages conformal prediction to specify such uncertainty regions in a data-driven fashion. Such regions have distribution-free coverage guarantees on the true system dynamics, in turn allowing for a probabilistic characterization of the regret of the resulting robust controller. We then demonstrate that such a controller can be efficiently produced via a novel policy gradient method that has convergence guarantees. We then demonstrate the superior empirical performance of our method over alternate robust control specifications, such as $\mathcal{H}_\infty$ and LQR with multiplicative noise, across a collection of engineering tasks.

4.1 Conformal Decision Making for Ensembles

4.1.1 Introduction

Ensemble methods are an oft-used class of statistical modeling techniques due to their ability to reduce variance or improve predictive accuracy [Schapire et al., 1999, Zhang and Ma, 2012, Dietterich, 2000]. Such methods are increasingly being coupled with complex, black-box models, such as in multi-modal language models [Zhang et al., 2023, Radford et al., 2021, Sun, 2013, Zhao et al., 2017, Yan et al., 2021]. Couplings of this sort are seeing ever-widening deployment in safety-critical settings, such as medicine [Yuan et al., 2018, Li et al., 2018, Yuan et al., 2017] and robotics [Brena et al., 2020, Blasch et al., 2021, Alatise and Hancke, 2020].

Increasing interest is, therefore, now being placed on quantifying uncertainty for such models [Subedar et al., 2019, Tian et al., 2020, Denker and LeCun, 1990, Havasi et al., 2020, Malinin and Gales, 2018]. Towards this end, methods of uncertainty quantification have arisen, such as deep ensembles and committee estimation [Rahaman et al., 2021, Abdar et al., 2021, Carrete et al., 2023]. Such methods, however, sacrifice generality with the imposition of distributional assumptions, motivating the need for distribution-free uncertainty quantification for ensemble methods.

One method for performing distribution-free uncertainty quantification is conformal prediction, which provides a principled framework for producing distribution-free prediction regions with marginal frequentist coverage guarantees [Angelopoulos and Bates, 2021, Shafer and Vovk, 2008]. By using conformal prediction on a user-defined score function, prediction regions attain marginal coverage guarantees. While calibration is guaranteed from this procedure, predictive efficiency, i.e. the size of the resulting prediction regions, can be unboundedly large for poorly chosen score functions.

As a result, methods have arisen to perform conformal model aggregation, which both provide uncertainty estimates of the ensembled predictions and do so in ways as to minimize the prediction region size [Gasparin and Ramdas, 2024a, Trunov and V'yugin, 2023, Yang and Kuchibhotla, 2024, V'yugin and Trunov, 2023, Gasparin and Ramdas, 2024b]. While such approaches succeed in reducing the prediction region size over naive aggregation, they all aggregate the *separately conformalized* prediction regions of the predictors in the ensemble. In doing so, they forgo the possibility of automatically leveraging shared structure amongst the scores of the individual predictors, resulting in conservative prediction regions.

In its place, therefore, a recent work [Rivera et al., 2024] proposed to perform aggregation in *score space* by extending traditional conformal prediction to consider a multivariate score function and defining prediction regions using "quantile envelopes" in place of scalar quantiles. Doing so enables efficient, data-driven, automated conformal model aggregation. They demonstrated that such aggregation improved predictive efficiency across classification tasks. We here demonstrate

that such an approach can similarly be leveraged in regression settings, both for direct coverage and decision-making applications.

4.1.2 Related Works

Ensemble methods consist of K predictors $f_k : \mathcal{X}_k \to \mathcal{Y}$; notably, such predictors need not map from the same set of covariates. A naive approach for uncertainty quantification would then be to conformalize the ensembled predictor. That is, for an ensembling algorithm $\mathcal{F} : \mathcal{Y}^K \to \mathcal{Y}$, a score function $s(\mathcal{F}(f_1(x), ..., f_K(x)), y)$ would be defined. Denoting the $\lceil (N_{\mathcal{C}}+1)(1-\alpha) \rceil / N_{\mathcal{C}}$ quantile of the score distribution over $\mathcal{D}_C$ as $\widehat{q}(\alpha)$, $\mathcal{C}(x) = \{y : s(x, y) \leq \widehat{q}(\alpha)\}$ would then be calibrated.

Such an approach, however, lacks some desirable properties. In particular, prediction regions $\mathcal{C}(x)$ should have the quality that, if a particular predictor has less uncertainty in its predictions, as is frequently true of ensemble settings where the predictors span multiple input data modalities, upon routing to that predictor, the corresponding size of the prediction region should be smaller than if it had been routed to a different predictor. While the naive approach does, in principle, support this property, it ultimately relies on defining an *uncertainty-aware* ensembling algorithm $\mathcal{F}$. In its typical form, however, $\mathcal{F}$ simply takes *point predictions* $f_1(x), ..., f_K(x)$ in as input, meaning any uncertainty-awareness would need to be baked in a priori into the definition of $\mathcal{F}$ through domain knowledge of the uncertainties of the predictors $f_1, ..., f_K$, which can seldom be specified precisely, sacrificing the predictive efficiency of $\mathcal{C}(x)$.

Conformal model aggregation, thus, seeks to mitigate these deficiencies by aggregating the prediction regions $\mathcal{C}_1(x), ..., \mathcal{C}_K(x)$ rather than the individual point predictions [Gasparin and Ramdas, 2024a, Yang and Kuchibhotla, 2024, V'yugin and Trunov, 2023, Gasparin and Ramdas, 2024b]. While there are several methods in this vein, they can be categorized into one of two general approaches. The first line of work seeks to perform model *selection*, in which a single conformal predictor is selected $\mathcal{C}_{k^*}$, typically based on the criterion of minimizing region size $k^* := \arg\min_k \mathbb{E}[\mathcal{L}(\mathcal{C}_k(X))]$ [Yang and Kuchibhotla, 2024, V'yugin and Trunov, 2023].

Generally, however, methods leveraging the full collection of predictors produce less conservative regions [Gasparin and Ramdas, 2024a,b]. Such works aggregate the individual prediction regions into a final region by defining $\mathcal{C}(x) := \{y \mid \sum_{k=1}^{K} w_k \mathbb{1}[y \in \mathcal{C}_k(x)] \geq \widehat{a}\}$ for weights $\{w_k\} \in [0, 1]$ such that $\sum_{k=1}^{K} w_k = 1$ and a threshold $\widehat{a}$. Methods then differ in the procedure by which $\{w_k\}$ and $\widehat{a}$ are prescribed, several of which were prescribed by [Gasparin and Ramdas, 2024b], whose detailed presentation is deferred to Section C.7 for space reasons. We note that the methods of [Gasparin and Ramdas, 2024a] are designed for a different setting than that considered herein, namely that in which conformal coverage is sought adaptively over data streams.

In this vein, [Luo and Zhou, 2024] have recently proposed a vector-score extension as that

discussed herein, in which candidate weight vectors $\{w_m\} \in \mathbb{R}^K$ are searched over for score aggregation. That is, a vector $s(x) := (s_1(x,y), ..., s_K(x,y)) \in \mathbb{R}^K$ of scores $s_k(x,y)$ corresponding to each predictor $f_k(x)$ is predicted and its aggregate prediction region defined on the projection $\langle w_{m^*}, s \rangle$ for w_{m^*} the weight resulting in the smallest prediction region. This method, however, has two shortcomings addressed in [Rivera et al., 2024]. The first is that their method can only be applied in classification settings and the second that their approach only uses a *single* weighted projection, resulting in suboptimal aggregation and, therefore, conservative prediction regions.

We now present a review of the score aggregation strategy proposed in [Rivera et al., 2024]. As mentioned, this approach proposed considering a vector $s(x) := (s_1(x,y), ..., s_K(x,y)) \in \mathbb{R}^K$ of scores components $s_k(x,y)$ similar to [Luo and Zhou, 2024]. From here, a collection of projection directions $\{u_m\}_{m=1}^M$, with each $u_m \in \mathcal{S}^{K-1} \cap \mathbb{R}_+^K$, and corresponding quantiles $\{\widehat{q}_m\}$ were defined such that the prediction region defined by simultaneous coverage under each projection direction, i.e. $\mathcal{C}(x) := \{y \mid u_m^\top s(x,y) \leq \widehat{q}_m \ \ \forall m = 1, ..., M\}$, satisfies the typical marginal conformal coverage guarantees.

In particular, such $\{\widehat{q}_m\}$ are defined over a two-step procedure: by splitting the calibration set $\mathcal{S}_C$ into two sets $\mathcal{S}_C^{(1)}$ and $\mathcal{S}_C^{(2)}$, an initial set of quantiles $\{\widetilde{q}_m\}$ are defined that achieve optimally tight coverage of $1-\alpha$ of the points in $\mathcal{S}_C^{(1)}$. This procedure, however, requires conditioning on $\mathcal{S}_C^{(1)}$ to define the optimal $\{\widetilde{q}_m\}$, in turn violating the required exchangeability assumptions. For this reason, the held-out $\mathcal{S}_C^{(2)}$ is then used for final calibration, in which an adjustment factor $\widehat{t}$ is fit such that $\widehat{q}_m := \widehat{t}\widetilde{q}_m$ satisfies $1-\alpha$ coverage of $\mathcal{S}_C^{(2)}$. This two stage procedure parallels the recalibration step of CANVI (Section 2.3.2).

4.1.3 Ensemble Predict-Then-Optimize

With this generalization of the score function, a natural question is how to leverage the resulting prediction regions $\mathcal{C}(x)$. For *classification*, where $|\mathcal{Y}| \in \mathbb{N}$, explicit construction of $\mathcal{C}(x)$ is straightforward: for any x, $\mathcal{C}(x)$ can be constructed by iterating through $y \in \mathcal{Y}$ and checking if $s(x,y) \in \widehat{\mathcal{Q}}$ by comparing $s(x,y)$ against each one of the thresholds $\widehat{q}_m$ after projection.

In the case of regression, however, the prediction region cannot be explicitly constructed in the general case, since $\mathcal{Y}$ contains uncountably many elements. In fact, explicit construction is generally not of interest for downstream regression applications. We, therefore, focus on the predict-then-optimize application discussed in the previous chapter, and demonstrate the CSA prediction regions can be leveraged in their framework. Such an extension has natural applications to the settings discussed in this original study. For instance, the robust traffic routing setting considered in Section 3.3.2 naturally lends itself to an ensembling approach in considering multiple predictive models, such as a $q_2(C \mid X)$ predicting traffic based instead on historical trends.

We note that the below described algorithm is better understood with a visual accompaniment, which we provide in Section C.2. In Section 3.2.3, we demonstrated that solving the robust problem variant $w^*(x) := \min_w \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$ in a computationally efficient manner is feasible by performing gradient-based optimization on w, where the gradient $\nabla_w \phi(w)$ of $\phi(w) := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$ can be computed by leveraging Danskin's Theorem so long as $\max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$ is efficiently computable for any fixed w. We focus on demonstrating that this remains the case for CSA, specifically considering the case where individual view score functions take the form of the "GPCP" score considered therein. In this setup, each constituent predictor is a generative model $q_k(C \mid X)$ from which $\{\widehat{c}_{kj}\}_{j=1}^{J_k} \sim q_k(C \mid X)$ samples are drawn. Note that J_k need not be constant across k. The GPCP score, used to define the score components, is

$$s_k(x, c) = \min_{j \in 1, ..., J_k} \left[||\widehat{c}_{kj} - c||_2 \right]. \tag{4.1}$$

Notably, this framework subsumes many standard regression settings, e.g., for a deterministic predictor, one can take $q_k(C \mid X) = \delta(f_k(X))$. To compute $\max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$, we first let $\vec{j} \in \mathcal{J} = \{j_1, ..., j_K\}$ be an indexing tuple, where each $j_k \in \{1, ..., J_k\}$. That is, each $\vec{j}$ is a vector that "selects" one sample per predictor. Notably then, the projection $u_m^\top s(\widehat{c}_{\vec{j}}, c)$ is convex in c, since the projection directions are all restricted to $\mathcal{S}_+^{K-1}$. Thus,

$$c_{\vec{j}}^* := \arg\max_c f(w, c) \qquad \text{s.t.} \qquad u_m^\top s(\widehat{c}_{\vec{j}}, c) \leq \widehat{q}_m \quad \forall m \in \{1, ..., M\} \tag{4.2}$$

remains a standard convex optimization problem. The final maximum can then be found by aggregation, namely $c^* = \arg\max_{\vec{j} \in \mathcal{J}} f(w, c_{\vec{j}}^*)$. While $|\mathcal{J}| = \prod_{k=1}^K J_k$, we note that certain cases of ensemble prediction, such as multi-view prediction, tend to have a limited number of predictors in practice, most typically $K = 2$ or $K = 3$. This coupled with the fact that computing over these indices is trivially parallelizable means this approach is still computationally tractable. The full procedure is outlined in Algorithm 8 in Section C.2.1 due to space limitations.

4.1.4 Experiments

We now study CSA empirically across several tasks, demonstrating its coverage guarantees with reduced conservatism. We first discuss how this method directly lends itself to improvement in regression settings in Section 4.1.4.1 by directly analyzing the sizes of the resulting prediction regions. In particular, we compare the score aggregation technique to those methods presented in Section 4.2.3, viz. the model selection of [Yang and Kuchibhotla, 2024], the aggregation methods of [Gasparin and Ramdas, 2024a], and the single weighted score projection (VFCP) of [Luo and Zhou, 2024]. We additionally include the naive strategy in which the ensemble predictor is directly

conformalized, using a natural aggregate "ensemble" score, explicitly described in the following sections. From the work of [Gasparin and Ramdas, 2024a], we consider the following methods: the standard majority-vote $\mathcal{C}^M$, partially randomized thresholding $\mathcal{C}^R$, and fully randomized thresholding $\mathcal{C}^U$ approaches (see Section C.7).

From these experiments, we find that the score aggregation approach significantly outperforms all alternate ensembling strategies in terms of predictive efficiency, in turn motivating its use in decision-making contexts. We, therefore, next apply the methods developed over Section 4.1.3 to an extension of the traffic prediction task studied in the previous chapter. Notably, the methods against which we perform comparisons in the direct regression experiments do not lend themselves for use in the predict-then-optimize setting, so we eliminate them from consideration therein.

4.1.4.1 Regression Tasks

We now study the predictive efficiency of CSA across a suite of regression tasks from [Fischer et al., 2023]. The data for each task were split with 50/45/5% for training, calibration, and testing for coverage and interval lengths, with five trials conducted over randomized selections of such sets. A $5/95\%$ split was used for $\mathcal{D}_{\mathcal{C}}{}^{(1)}$-$\mathcal{D}_{\mathcal{C}}{}^{(2)}$. The problem setup was replicated from [Gasparin and Ramdas, 2024b], in which four prediction methods were ensembled, namely an OLS model, a LASSO linear model, a random forest (RF), and an XGBoost model. A residual function was used as the score across all methods, namely $s(x, y) = |\widehat{f}(x) - y|$. Here, the "Ensemble" score was the standard $s(x, y) := \frac{|\mu(x)-y|}{\sigma(x)}$, where $(\mu(x), \sigma(x))$ are the ensemble mean and standard deviation. Prediction intervals could be analytically constructed for the $\mathcal{C}^M$, $\mathcal{C}^R$, and $\mathcal{C}^U$ methods. To assess CSA, however, a discretized grid $\mathcal{G}_Y \subset \mathcal{Y}$ of coarseness Δy was considered, and an interval length estimate given by $\mathcal{L}(\mathcal{C}(x)) \approx \Delta y \cdot |\{y : y \in \mathcal{G}_Y, s(x, y) \in \widehat{\mathcal{Q}}\}|$. We also present an ablation, labeled "Single-Stage," to demonstrate the two-stage calibration of the calibration procedure presented in [Rivera et al., 2024] is necessary to retain coverage; this single-stage approach does not split $\mathcal{S}_C$ and instead directly computes $\{\widehat{q}_m\}$ on $\mathcal{S}_C$ per Section 4.1.2.

We provide the results for $\alpha = 0.05$ and $\alpha = 0.025$ to demonstrate the consistency of the method performance. A subset of the results is given in Table 4.1; the full set of results is deferred to Section C.4. We see that CSA retains the coverage guarantees typical of conformal prediction yet produces significantly smaller prediction intervals than both the individual models and the alternate aggregation strategies. We additionally see that the "Single-Stage" approach fails to retain coverage, demonstrating the necessity of the two-stage calibration. We provide a visual comparison of the prediction regions resulting from these methods in Section C.5.

We additionally assessed the robustness of our method to imbalanced ensembles. The experiments of [Gasparin and Ramdas, 2024b] were conducted on a UCI benchmark task [Asuncion et al., 2007] with an ensemble of an OLS model, a LASSO linear model, a random forest, and an

Table 4.1: The results for five distinct tasks are shown below for $\alpha = 0.05$ (top five rows) and $\alpha = 0.025$ (bottom five rows). For each, the average coverages (grey rows) and prediction set lengths (white rows) with standard deviations are given, both assessed over 5 randomized draws of the training, calibration, and test sets. In cases where the method failed to achieve sufficient coverage (i.e. $< .93$ for $\alpha = 0.05$ and < 0.96 for $\alpha = 0.025$), we do not include it in comparison for set length.

Dataset/α	OLS	LASSO	RF	XGBoost	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	*Single-Stage*	CSA
361234	0.97 (0.011)	0.966 (0.011)	0.939 (0.002)	0.954 (0.006)	0.956 (0.011)	0.948 (0.01)	0.96 (0.013)	0.95 (0.006)	*0.955 (0.013)*	0.957 (0.01)
($\alpha = 0.05$)	9.673 (0.160)	9.645 (0.154)	10.080 (0.160)	9.157 (0.052)	9.196 (0.123)	8.703 (0.086)	9.524 (0.056)	17.759 (0.275)	*7.646 (0.073)*	**7.688 (0.181)**
361235	0.947 (0.0)	0.945 (0.005)	0.968 (0.016)	0.95 (0.005)	0.955 (0.016)	0.897 (0.005)	0.953 (0.011)	0.932 (0.021)	*0.745 (0.011)*	0.984 (0.005)
($\alpha = 0.05$)	20.961 (0.651)	24.241 (0.246)	**10.096 (0.587)**	11.387 (0.452)	11.782 (0.057)	—	16.088 (0.118)	15.823 (1.272)	*6.162 (0.458)*	11.695 (0.266)
361236	0.975 (0.008)	0.975 (0.008)	0.961 (0.0)	0.948 (0.012)	0.948 (0.012)	0.938 (0.012)	0.965 (0.008)	0.934 (0.004)	*0.94 (0.004)*	0.963 (0.004)
($\alpha = 0.05$)	4.44e4 (1.17e3)	4.45e4 (1.23e3)	5.08e4 (3.86e2)	4.10e4 (1.22e3)	4.32e4 (1.00e3)	4.09e4 (1.09e3)	4.44e4 (8.52e2)	6.05e4 (2.41e3)	*3.10e4 (2.48e3)*	**3.34e4 (1.28e3)**
361237	0.969 (0.023)	0.969 (0.023)	0.981 (0.0)	0.923 (0.0)	0.954 (0.015)	0.9 (0.008)	0.969 (0.023)	0.885 (0.038)	*0.8 (0.015)*	0.977 (0.008)
($\alpha = 0.05$)	44.019 (0.990)	44.069 (1.115)	27.035 (1.014)	—	26.524 (1.244)	—	31.967 (1.118)	—	*14.473 (0.503)*	**23.145 (0.199)**
361241	0.954 (0.001)	0.956 (0.001)	0.944 (0.005)	0.957 (0.002)	0.954 (0.002)	0.923 (0.0)	0.952 (0.0)	0.949 (0.001)	*0.917 (0.006)*	0.951 (0.001)
($\alpha = 0.05$)	19.133 (0.062)	20.245 (0.095)	18.102 (0.055)	18.482 (0.062)	17.958 (0.062)	—	18.932 (0.034)	29.548 (0.191)	*15.199 (0.427)*	**17.328 (0.097)**
361234	0.987 (0.008)	0.987 (0.008)	0.974 (0.004)	0.977 (0.008)	0.982 (0.008)	0.971 (0.01)	0.981 (0.01)	0.97 (0.008)	*0.976 (0.01)*	0.973 (0.006)
($\alpha = 0.025$)	11.939 (0.137)	11.871 (0.084)	12.484 (0.168)	11.972 (0.009)	11.587 (0.110)	11.157 (0.086)	11.965 (0.050)	25.598 (0.974)	*9.306 (0.259)*	**8.855 (0.059)**
361235	0.987 (0.0)	0.982 (0.011)	0.979 (0.011)	0.984 (0.005)	0.989 (0.005)	0.966 (0.016)	0.976 (0.005)	0.958 (0.021)	*0.889 (0.011)*	0.989 (0.005)
($\alpha = 0.025$)	24.595 (0.825)	28.841 (1.129)	**11.811 (0.992)**	14.237 (0.786)	14.472 (0.172)	**12.278 (0.026)**	19.231 (0.356)	—	*7.719 (0.467)*	**12.563 (0.766)**
361236	0.992 (0.004)	0.992 (0.004)	0.981 (0.0)	0.965 (0.008)	0.975 (0.008)	0.965 (0.008)	0.977 (0.012)	0.955 (0.008)	*0.955 (0.012)*	0.973 (0.004)
($\alpha = 0.025$)	4.86e4 (8.74e2)	4.86e4 (8.68e2)	5.61e4 (3.69e2)	4.66e4 (1.57e3)	4.76e4 (8.10e2)	4.57e4 (1.06e3)	4.92e4 (7.53e2)	—	*3.29e4 (3.11e3)*	**3.58e4 (1.95e3)**
361237	0.981 (0.0)	0.981 (0.0)	0.981 (0.0)	0.977 (0.008)	0.962 (0.0)	0.962 (0.0)	0.977 (0.008)	0.965 (0.008)	*0.927 (0.031)*	0.981 (0.0)
($\alpha = 0.025$)	47.738 (0.542)	47.440 (0.959)	30.785 (0.037)	**26.208 (0.897)**	30.554 (0.561)	**27.182 (0.803)**	35.982 (0.619)	67.660 (6.380)	*18.214 (0.436)*	**26.897 (0.515)**
361241	0.979 (0.001)	0.978 (0.001)	0.976 (0.001)	0.978 (0.001)	0.978 (0.0)	0.964 (0.002)	0.977 (0.0)	0.972 (0.002)	*0.958 (0.003)*	0.979 (0.0)
($\alpha = 0.025$)	21.772 (0.085)	23.089 (0.106)	21.543 (0.009)	21.454 (0.109)	20.862 (0.088)	**19.291 (0.060)**	21.905 (0.041)	40.082 (0.045)	*17.765 (0.329)*	19.897 (0.062)

MLP, and they found the conformalized random forest to outperform all the proposed aggregation strategies, due to the lack of orthogonal information in considering the other predictors. We find that, in these degenerate cases, where the best decision is to simply choose a single predictor, our method outperforms other aggregation methods and nearly matches the performance of the best conformalized predictor in hindset; the results are presented in Section C.6 across a number of UCI benchmarks.

4.1.4.2 CSA Predict-Then-Optimize

We now revisit the robust traffic routing task of Section 3.3.2. As a recap, in this task, a time series of T preceding precipitations is used to predict future precipitations and, in turn, future traffic, as fully described in Section B.4. We consider the traffic routing problem for a fixed source-target pair (s, t) over the graph of Manhattan, where $|\mathcal{V}| = 4584$ and $|\mathcal{E}| = 9867$. Formally,

$$w^*(x) := \min_{w} \max_{\widehat{c} \in \mathcal{C}(x)} \widehat{c}^T w \quad \text{s.t.} \quad w \in [0,1]^{\mathcal{E}}, Aw = b, \mathcal{P}_{X,C}(C \in \mathcal{C}(X)) \geq 1 - \alpha$$

where $x \in \mathbb{R}^{T \times H \times W}$ are the previous precipitation readings, $w_e \in \mathbb{R}^{|\mathcal{E}|}$ the traffic proportion routed along road e, $c \in \mathbb{R}^{|\mathcal{E}|}$ the transit times anticipated across roads, $A \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}|}$ the graph incidence matrix, and $b \in \mathbb{R}^{|\mathcal{V}|}$ the vector that specifies the routing problem, in which $b_s = 1, b_t = -1$, and $b_k = 0$ for s the travel source node, t the terminal node, and $k \notin \{s, t\}$ all other nodes.

We then consider two probabilistic models for traffic prediction, namely one based on the classical probabilistic Lagrangian integro-difference approach (STEPS) of [Pulkkinen et al., 2019] and one on the modern latent diffusion model (LDM) approach of [Gao et al., 2024]. As a result of the higher inference cost of the latter, we consider the setup where $J_1 > J_2$, specifically with $J_1 = 4$ and $J_2 = 1$, highlighting the flexibility of non-uniform sampling from predictors discussed in Section 4.1.3. As discussed in Section 4.1.4, the alternate aggregation strategies do not lend themselves for use in this setting. We, therefore, only compare CSA to the separate conformalizations of the two predictors, with the score from Equation (4.1). We here evaluate the methods using the expected suboptimality gap proportion, $\Delta_{\%} = \mathbb{E}_X[\Delta(X, C(X))/\min_w f(w, C(X))]$, where Δ is defined as discussed in Section 1.1.3. This measures the conservatism of the robust optimal value and is bounded in $[0, 1]$.

Experiments were conducted with $|\mathcal{D}_C| = 200$, with a $20/80\%$ split used for $\mathcal{D_C}^{(1)}$-$\mathcal{D_C}^{(2)}$. The suboptimality was then computed across 100 i.i.d. test samples. To assess the improvement, we conducted two paired t-tests, where $H_0 : \Delta_{\%}^{(\mathrm{CSA})} = \Delta_{\%}^{(\mathrm{STEPS})}$ and $H_1 : \Delta_{\%}^{(\mathrm{CSA})} < \Delta_{\%}^{(\mathrm{STEPS})}$ and similarly for $\Delta_{\%}^{(\mathrm{CSA})}$ and $\Delta_{\%}^{(\mathrm{LDM})}$. The results are provided in Table 4.2, from which we find that CSA significantly reduces the suboptimality after accounting for Bonferroni multiple testing. We see that, while conformalization of either of the two views individually already produces the desired coverage, CSA produces more informative prediction regions, and hence less conservative robust upper bounds.

Table 4.2: Coverages for $\alpha = 0.05$ for the individually conformalized and CSA approach and p-values of the paired t-tests comparing $\Delta_{\%}$ are shown, both computed over 100 i.i.d. test samples.

Coverage			**P-values for H_1**
STEPS	LDM	CSA	$\Delta_{\%}^{(\mathrm{CSA})} < \Delta_{\%}^{(\mathrm{STEPS})}$: 3.61×10^{-4}
0.981	0.962	0.968	$\Delta_{\%}^{(\mathrm{CSA})} < \Delta_{\%}^{(\mathrm{LDM})}$: 9.50×10^{-4}

4.2 Conformal Robust Control of Linear Systems

We now discuss another application of the developed conformal decision-making framework. This extension focuses on the setting of linear controls problems, leveraging conformal prediction for robust model-based linear control.

4.2.1 Introduction

Seeking control over a family of dynamical systems is a problem often encountered in engineering [Killian et al., 2016, Wu et al., 2018, Aksland et al., 2023]. One prevalent application of this is in

cases where engineering designs and their respective controllers are being concurrently developed, known as control co-design (CCD) [Garcia-Sanz, 2019]. Traditional engineering design loops operated sequentially, first proposing a design and then developing a controller [Reyer et al., 2001, Friedland, 1995]. Such workflows, however, sacrificed the improved optimality possible in their coupling, hence the increasing interest in leveraging end-to-end co-control design pipelines [Fathy et al., 2001, Falck et al., 2021].

Initial works in CCD studied optimal design assuming perfectly specified, deterministic system dynamics [Allison et al., 2014, Azad et al., 2018, 2019, Behtash and Alexander-Ramos, 2018]. Such assumptions have, however, become overly restrictive, resulting in interest in robust extensions of the CCD formulation, referred to as uncertain CCD (UCCD) [Azad and Herber, 2022, 2023b, Bird et al., 2023]. Such uncertainty can arise from many sources in the design process, such as noise in the controllers, uncertainties in the design parameters, or unmodeled dynamics. The UCCD specification also differs depending on the risk tolerance in the downstream application. For instance, in risk-neutral settings, stochastic specifications are appropriate [Azad and Alexander-Ramos, 2020b, Cui et al., 2022, Behtash and Alexander-Ramos, 2024], whereas in risk-averse settings, probabilistic [Cui et al., 2020a,b, Nguyen et al., 2022] or worst-case forms [Azad and Alexander-Ramos, 2021, Nash et al., 2021, Azad and Alexander-Ramos, 2020a] are used.

We focus on the worst-case robust UCCD formulation (WCR-UCCD), specifically on dynamics misspecification. WCR-UCCD requires specifying a dynamics uncertainty region. Existing methods of specification, however, tend to be ad-hoc and, thus, fail to provide any guarantees of the robust solution as it relates to the selection of this uncertainty set, rendering its choice often difficult and resulting in suboptimal controller synthesis [Azad and Alexander-Ramos, 2020b].

We, thus, focus herein on providing a principled distribution-free specification of the *robust control subproblem* in WCR-UCCD and an associated solution method with convergence guarantees. One special case of interest in UCCD is in the setting of linear quadratic regulators (LQRs), where the underlying system dynamics take on a linear structure [Ahmadi et al., 2023, Fathy et al., 2003, Jiang et al., 2016]. LQR systems are of broad interest both due to their analytic tractability and widespread applicability to practical engineering systems [Zhao et al., 2024, Mamakoukas et al., 2019, Bevanda et al., 2022]. We, therefore, propose a method for specifying the LQR WCR-UCCD control subproblem that lends itself to efficient solution by leveraging conformal prediction on observed design information. A related use of conformal prediction for predict-then-optimize problems was the focus of the previous chapter. Unlike that setting, however, the application of conformal prediction to control has complications related to the stability of controllers under model uncertainty. Our contributions are as follows:

- Providing a framework to define robust LQR control problems with distribution-free probabilistic regret guarantees, across deterministic or stochastic and discrete- or continuous-

time dynamics, and demonstrating empirical improvements over alternative robust control schemes.

- Extending conformalized predict-then-optimize to cases where calibration data is observed with noise and where the domains of both the maximization *and* minimization components of the robust formulation depend on the conformalized predictor.
- Providing a novel policy subgradient method for robust controller synthesis with convergence guarantees proven via subgradient dominance.

4.2.2 Methodology

We now discuss conformally robust LQR, providing the formulation in Section 4.2.2.1, regret guarantees in Section 4.2.2.3 and Section 4.2.2.4, and a controller synthesis algorithm with convergence guarantees in Section 4.2.2.5.

4.2.2.1 Problem Formulation

For the presentation below, let $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times m}$. Let C denote the full dynamics matrix $C := [A, B] \in \mathbb{R}^{n\times(n+m)}$. We additionally assume a linear control scheme, namely $u_t = -Kx_t$ for some gain matrix K. Additionally, denote $W := [I_{n\times n} \; -K^\top]^\top \in \mathbb{R}^{(n+m)\times n}$, such that the closed-loop dynamics are given by $CW = A - BK$. As discussed in Section 1.1.5, we assume a dataset of designs and associated trajectories is observed. We assume such a dataset $\mathcal{D}$ consists of N samples $(\theta^{(i)}, C^{(i)}) \sim \mathcal{P}(\Theta, C)$ and $K^{(i)} \sim \mathcal{P}(K)$, where $\mathcal{P}(\Theta, C)$ is an unknown joint distribution over designs and dynamics and $\mathcal{P}(K)$ an unknown distribution on gain matrices. We make no assumptions on such distributions other than that each gain matrix $K^{(i)}$ is a stabilizing controller for the respective $C^{(i)}$ dynamics. Note that these underlying true dynamics $C^{(i)}$ are never observed directly by the learning algorithm; only the resulting trajectories are observed. Such trajectories are generated by evolving the state via $x_{t+1}^{(i)} = (C^{(i)}W^{(i)})x_t^{(i)}$ over a time horizon T. The final dataset, therefore, takes the form $\mathcal{D} = \{\theta^{(i)}, \{(x_t^{(i)}, u_t^{(i)})\}_{t=1}^T\}_{i=1}^N$. We are interested in studying a risk-sensitive formulation of LQR:

$$K^*_{\text{rob}}(\mathcal{U}(\theta)) := \underset{K\in\mathcal{K}(\mathcal{U}(\theta))}{\arg\min} \max_{[\widehat{A},\widehat{B}]:=\widehat{C}\in\mathcal{U}(\theta)} \mathbb{E}[J(K, \widehat{A}, \widehat{B})]$$
$$\text{s.t.} \quad \dot{x} = \widehat{A}x + \widehat{B}u + w \qquad \mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq 1 - \alpha,$$

where J is the objective function particular to the setting of interest, differing between infinite and finite time horizons and continuous and discrete time dynamics, and $\mathcal{U}(\theta)$ is an uncertainty set over dynamics. Notably, the notion of stabilizing controllers must be generalized in this robust

formulation, since the nominal formulation is for a specific C. We, thus, consider those controllers that stabilize the entire uncertainty set, which we refer to as the "universal stabilizing set," formally $\mathcal{K}(\mathcal{U}(\theta)) := \bigcap_{\widehat{C} \in \mathcal{U}(\theta)} \mathcal{K}(\widehat{C})$, where $\mathcal{K}(\widehat{C})$ is Equation (1.4) evaluated for a particular $\widehat{A}, \widehat{B}$.

4.2.2.2 Score Function

From the trajectories in $\mathcal{D}$, we can perform system identification using least squares estimation to recover estimates of the system dynamics, $(\widetilde{A}^{(i)}, \widetilde{B}^{(i)})$ [Ljung et al., 1987]. With this, we obtain a final dynamics dataset $\widetilde{\mathcal{D}} = \{\theta^{(i)}, \widetilde{C}^{(i)}\}_{i=1}^{N}$, which we then leverage in the standard manner of split conformal prediction. That is, we split $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}}_{\mathcal{T}} \cup \widetilde{\mathcal{D}}_{\mathcal{C}}$, the former of which we use to train a system parameters predictor $\widehat{C} := f(\theta)$. Notably, leveraging split conformal in this setting has the complication that the ground truth used, namely in $\widetilde{\mathcal{D}}_{\mathcal{C}}$, is itself an estimate $\widetilde{C}$ even though coverage is sought on C. We assume for this initial discussion that for a fixed coverage level α, we can obtain prediction regions with the desired coverage, satisfying $\mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq 1 - \alpha$, using $\widetilde{\mathcal{D}}_{\mathcal{C}}$. The treatment of this gap between $\widetilde{C}$ and C is discussed in Section 4.2.2.4.

We take the score to be $s(\theta, C) = ||f(\theta) - C||_{\text{op}}$, where $|| \cdot ||_{\text{op}}$ is the matrix *operator* norm, i.e. $||A||_{\text{op}} = \sigma_{\max}(A)$, from which the resulting prediction regions take on the form of $\mathcal{B}_{\widehat{q}}(f(\theta))$, namely a ball of radius $\widehat{q}$, the conformal quantile, under the $|| \cdot ||_{\text{op}}$ metric.

4.2.2.3 Coverage Guarantee Consequences

We now characterize the regret induced by the robustness across LQR setups, that is $\mathcal{R}(\theta, C) := \mathbb{E}[J(K^*_{\text{rob}}(\mathcal{U}(\theta)), C) - J(K^*(C), C)]$, where the randomness is over stochastics in the *true* system dynamics $C := [A, B]$ and in the $\mathcal{P}(C \mid \theta)$ map. We explicitly note C in the regret notation to emphasize that, while the controller K is defined using *estimated* system dynamics, the final evaluation over the *true* C dynamics.

We provide the regret statements for the continuous, infinite time horizon cases below and defer the discrete-time and finite time horizon cases to the Appendix. Both settings require a mild assumption that the problem parameters have bounded norms, formalized in Assumption 4; this will hold for any realistic problem setup. Notably, however, the two settings differ in that the stochastic dynamics requires $Q(t)$ and $R(t)$ be discounted over t, while the deterministic case is fully compatible with non-discounted rewards. Intuitively, this discounting is necessary, as stability alone in the stochastic setting does not ensure a bounded objective; the state can continue to oscillate and result in an unbounded accumulation of error if the terms tied to the state covariance matrix do not decay. Other works frame this assumption as "mean-square stability," (see e.g. [Gravell et al., 2020a]). We formally pose this as Assumption 5.

Assumption 4. *For any θ, $K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, and $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$, $D(K) := \sqrt{n}||Q + K^\top RK||_\infty ||x_0||_\infty^2 ||W||_{\text{op}} < \infty$*

Assumption 5. *For any θ, $\exists$ constants $\alpha_1, \beta_1 > 0$ such that for all $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$, $K \in \mathcal{K}(\widehat{C})$, and $t \geq 0$, $||Q(t) + K^\top R(t)K|| \leq \beta_1 e^{-\alpha_1 t}$ and $\min_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))}(2\alpha_2(\widehat{C}) + \alpha_1) > 0$ where $\alpha_2(\widehat{C}) := \max_{K \in \mathcal{K}(\widehat{C})}(-\max_i \text{Re}(\lambda_i(\widehat{C}W))) > 0$.*

The regret bound below decomposes into two terms. The first captures the suboptimality in designing a controller with the conformal dynamics set instead of against the true dynamics; this coincides with the suboptimality characterized in previous works described in Section 1.1.3. The other is a novel aspect that arises in this controls setting: since the robust control problem optimizes over a *restricted* set of controllers, namely those that universally stabilize the full conformal dynamics set instead of those that only stabilize the true dynamics, there is an additional "domain gap" suboptimality. Intuitively, if the true optimal controller falls in $\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, this latter term should vanish. Towards this end, we introduce the following notion.

Definition 1. *Let $M(C, K^*(C)) := A - BK^*(C)$ be diagonalizable. Define $r(C, K^*(C)) := \frac{\min_i \left(-\Re(\lambda_i(M))\right)}{\kappa(U)\,\|W\|_{\text{op}}}$, where $M = U\Lambda U^{-1}$, $\kappa(U)$ is the condition number of U, and $W = [I \;\; -K^*(C)^\top]^\top$.*

Across the theorems stated below, therefore, if the conformal radius is smaller than this $r(C, K^*(C))$ margin term, the "domain gap" term vanishes. Intuitively, this property follows as the stability of the system can be characterized by the closed-loop eigenvalues, whose values change by a bounded amount in considering the perturbations captured in the conformal region. We additionally see that, as $\widehat{q} \to 0$, the suboptimality vanishes, as we would expect in recovering the true dynamics. Thus, users should seek to produce prediction regions with coverage that are as small as possible to produce informative upper bounds on the nominal optimal value. Notably, the statements below are given in terms of the objective Lipschitz constants L: explicit expressions of L along with the discrete-time and finite time horizons theorems and proofs are given in Sections C.8.2 to C.8.5.

Theorem 4.2.1 (Deterministic, continuous-time)**.** *Let $J(K, C) := \int_0^\infty (x(t)^\top (Q + K^\top RK)x(t))dt$ for $w = 0$. Assume that $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$. Then, under Assumption 4,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L is the Lipschitz constant of $J(K, \widehat{C})$ in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm. Further, if $\widehat{q} < r(C, K^*(C))$, (see Definition 1) $\Delta_{\text{dom}}(\Theta, C) = 0$.

Theorem 4.2.2 (Stochastic, continuous-time)**.** *Let* $J(K, C) := \mathbb{E}\left[\int_0^\infty (x(t)^\top (Q(t) + K^\top R(t)K)x(t))dt\right]$ *with* $w(t)$ *a white noise process with spectral density* Σ *such that* $D_2(K) := ||\Sigma||_{\text{op}}||W||_{\text{op}} < \infty$*. Assume further that* $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$*. Then, under Assumption 4 and Assumption 5,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L *is the Lipschitz constant of* $J(K, \widehat{C})$ *in* $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ *under the operator norm. Further, if* $\widehat{q} < r(C, K^*(C))$*, (see Definition 1)* $\Delta_{\text{dom}}(\Theta, C) = 0$*.*

4.2.2.4 Ambiguous Ground Truth

We now discuss the complication of obtaining coverage guarantees on C despite only observing estimates $\widetilde{C} = C + \epsilon$ in the dataset, where $\epsilon \sim \mathcal{N}(0, \Sigma)$. This form of the estimation error can be shown to hold asymptotically under mild assumptions by classical results from least squares estimation, as shown for LTI system identification in [Ljung et al., 1987].

The coverage guarantee result given in Theorem 4.2.3 is the multivariate extension of Theorem A.5 from [Feldman et al., 2023b] and is a novel contribution to the broader space of conformal prediction. Intuitively, we show that if, for all θ, the density $\mathcal{P}(C \mid \theta)$ peaks in $\mathcal{U}(\theta)$, we retain marginal coverage guarantees. If $\mathcal{P}(C \mid \theta)$ is unimodal and radially symmetric about its mode, this condition is satisfied so long as $\mathcal{U}(\theta)$ captures the mode. The map between design parameters θ and A, B is often unimodal, making such a structural assumption reasonable; this was true classically, where a deterministic map was parametrically given by physics (discussed more in Section C.8.6), and remains true of data-driven surrogates in UCCD [Azad and Herber, 2023a, Azad et al., 2024]. $\mathcal{U}(\theta)$ capturing the mode is also a weak assumption assuming a zero-centered distribution for ϵ, since it then amounts to capturing the mode of $\mathcal{P}(\widetilde{C} \mid \theta)$, which holds for any sufficiently accurate predictor. We empirically demonstrate that such assumptions hold and, thus, that the coverage guarantees are retained in Section 4.2.4. The full proof of this theorem is deferred to Section C.8.7.

Theorem 4.2.3. *Let* $\widetilde{C} = C + \epsilon$ *where vec*$(\epsilon) \sim \mathcal{N}(0, \Sigma)$*, where* $\epsilon \perp\!\!\!\perp (\Theta, C)$*. Assume* $\mathcal{U}(\theta) = \{C' \mid ||f(\theta) - C'||_{\text{op}} \leq \widehat{q}\}$ *satisfies* $\mathcal{P}_{\Theta,\widetilde{C}}(\widetilde{C} \in \mathcal{U}(\Theta)) \geq 1 - \alpha$*, where* $||\cdot||_{\text{op}}$ *denotes the matrix operator norm. If for any* $\theta \in \Theta$ *and* $\delta > 0$*,* $\mathcal{P}(\widehat{q}^2 - \delta \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta) > \mathcal{P}(\widehat{q}^2 \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta)$*, then*

$$\mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq \mathcal{P}_{\Theta,\widetilde{C}}(\widetilde{C} \in \mathcal{U}(\Theta)) \geq 1 - \alpha.$$

4.2.2.5 Optimization Algorithm

Due to our generalization over traditional approaches to robust control (discussed in detail in Section 4.2.3), the standard approaches of solution used in those cases, namely generalized algebraic Riccati equations (GARE) or policy gradient, cannot be applied without modification. We, thus, now discuss how policy gradient can be adapted to efficiently solve the problem of interest and then demonstrate corresponding convergence results in Section 4.2.2.6. Given the novelty of the framing of Equation (4.2) over previous framings, specifically in the geometry of the uncertainty regions, the policy gradient expressions derived here too are novel, highlighted below. We frame this discussion around the deterministic, discrete-time, infinite time horizon setting, in which $x_{t+1} = (A - BK)x_t$. We assume that the initial state is drawn from a known distribution $x(0) \sim \mathcal{N}(0, X_0)$. Naively, computing the gradient would require estimation of the infinite sum in J; however, it is well known that the gradient can be computed using a Lyapunov formulation, given by

$$\nabla_K J(K, A, B) = 2((R + B^\top P_K B)K - B^\top P_K A)X_K, \tag{4.3}$$

where X_K and P_K respectively solve the two Lyapunov equations $\ell_X(X_K, \Delta_K) = 0$ and $\ell_P(P_K, \Delta_K, K) = 0$ for specified Q, R, K, and $\Delta_K := A - BK$, where

$$\begin{aligned} \ell_X(X_K, \Delta_K) &:= \Delta_K X_K \Delta_K^\top - X_K + X_0 \\ \ell_P(P_K, \Delta_K, K) &:= \Delta_K^\top P_K \Delta_K + Q + K^\top R K - P_K \end{aligned} \tag{4.4}$$

Note that, while ℓ_P also depends on the choice of Q and R, we do not explicitly note this in the notation as they remain fixed throughout the problem. If the continuous-time setting is of interest instead, there are analogous Lyapunov equations and gradient expressions to those respectively in Equation (4.3) and Equation (4.4). To solve Equation (4.2), we wish to perform gradient updates on K instead with respect to $\phi(K) := \max_{\widehat{C} \in \mathcal{U}(\theta)} J(K, \widehat{C})$. Naively, one could proceed through the remaining analysis by leveraging Danskin's Theorem to compute the gradient of $\nabla_K \phi(K)$, which would result in the expression $\nabla_K \phi(K) = \nabla_K J(K, C^*(K))$, where $C^*(K) := \arg\max_{\widehat{C} \in \mathcal{U}(\theta)} J(K, \widehat{C})$; however, the existence of such a gradient requires that $C^*(K)$ be the unique maximizer. Such an assumption is unlikely to hold in practice; for this reason, we instead relax this assumption and proceed using subgradients. That is, we suppose $C^*(K) := \mathrm{Arg}\max_{\widehat{C} \in \mathcal{U}(\theta)} J(K, \widehat{C})$ instead is a *set* and denote by $\partial_K \phi(K) := \{\nabla_K J(K, C_K) : C_K \in C^*(K)\}$ the *vertices* of the subdifferential.

Thus, robust policy optimization proceeds by iteratively updating K with *any* vertex of the subdifferential, namely by evaluating Equation (4.3) with $[A_K, B_K] := C_K$ for some $C_K \in C^*(K)$ and (X_K^*, P_K^*), the solutions to the Lyapunov equations when $\Delta_K^* := A_K - B_K K$. We initial-

ize this procedure with the optimal controller for the nominally predicted dynamics, i.e. $K^{(0)} := K^*(f(\theta))$. Extending LQR policy gradient methods to the robust setting, therefore, reduces to being able to efficiently solve the maximization problem of C_K over $\mathcal{U}(\theta) := \mathcal{B}_{\widehat{q}}(f(\theta))$. This can be estimated with gradient ascent, where the Lyapunov expression $\nabla_C J(K, C) = 2P_K C W X_K W^\top$ is derived in Section C.8.8. The use of subgradients for policy optimization and the derivation of the explicit $\nabla_C J(K, C)$ expression in its Lyapunov formulation are novel contributions to the robust LQR space; these aspects were heretofore unstudied as previously studied robust formulations (in Section 4.2.3) could be translated into GAREs and, therefore, did not require algorithmic innovation. The algorithm is given in Algorithm 6.

Algorithm 6 CONFORMALIZED PREDICT-THEN-CONTROL (CPC)

procedure CPC($\theta, f(\theta), \widehat{q}, \eta_K, \eta_C, T_K, T_C$)
  Inputs: Design θ, Predictor $f(\theta)$, Conformal quantile $\widehat{q}$, Step sizes η_K, η_C, Max steps T_K, T_C
  $\widehat{C} := f(\theta), K^{(0)} \leftarrow \text{SOLVEARE}(\widehat{C})$
  for $t_K \in \{0, \dots T_K - 1\}$ **do**
    $[A^{(0)}, B^{(0)}] := C^{(0)} \leftarrow \widehat{C}$
    for $t_C \in \{0, \dots T_C - 1\}$ **do**
      $\Delta^{(t_C)} := A^{(t_C)} - B^{(t_C)} K^{(t_K)}$
      $X^{(t_C)} \leftarrow \text{Solve}(\ell_X(X, \Delta^{(t_C)}) = 0; X)$
      $P^{(t_C)} \leftarrow \text{Solve}(\ell_P(P, \Delta^{(t_C)}, K^{(t_K)}) = 0; P)$
      $C^{(t_C+1)} \leftarrow \Pi_{\mathcal{B}_{\widehat{q}}(\widehat{C})}\big(C^{(t_C)} + \eta_C \left(2P^{(t_C)} C^{(t_C)} W^{(t_K)} X^{(t_C)} (W^{(t_K)})^\top\right)\big)$
    end for
    $\Delta^* := A_K - B_K K^{(t_K)}$ $\quad \triangleright\, C^* := C^{(T_C)}$
    $X^* \leftarrow \text{Solve}(\ell_X(X, \Delta^*) = 0; X)$
    $P^* \leftarrow \text{Solve}(\ell_P(P, \Delta^*, K^{(t_K)}) = 0; P)$
    $K^{(t_K+1)} \leftarrow K^{(t_K)} - \eta_K\big(2((R + (B_K)^\top P^* B_K) K^{(t_K)} - (B_K)^\top P^* A_K) X^*\big)$
  end for
  Return $K^{(T_K)}$
end procedure

4.2.2.6 Policy Gradient Convergence Guarantees

We now wish to demonstrate this policy gradient approach retains the desired convergence properties it satisfies in the nominal case. Convergence guarantees surprisingly hold in the standard case despite the nonconvexity of the problem in K due to a property known as "gradient dominance" [Gravell et al., 2020a]. A function $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ is gradient-dominated if, for some $\mu > 0$, $f(x) - f(x^*) \leq \mu ||\nabla_x f(x)||_F^2$, where $x^* := \arg\min_x f(x)$.

We proceed through the analysis similarly leveraging gradient dominance; however, our analysis has the novel problem of having to handle the non-uniqueness of the subgradient being used,

namely that our algorithm may perform updates with one of the collection of subdifferential vertices than using the uniquely defined gradient. For this reason, we instead consider a generalized notion of *subgradient* domination, defined as $\exists$ some $\mu > 0$ such that $f(x) - f(x^*) \leq \mu \min_{g \in \partial f(x)} ||g||_F^2$. We show that ϕ satisfies subgradient dominance and that this then produces convergence guarantees for Algorithm 6 in Theorem C.8.12.

The full proof is deferred to Section C.8.9 and parallels the proof strategy presented in [Fazel et al., 2018]; the main technical challenges are in demonstrating that bounds on expressions related to $J(K, C)$ and $\nabla_K J(K, C)$ are retained in our robust setting and that the non-uniqueness of the maximizer does not interfere with convergence. Intuitively, the non-uniqueness manifests as a looser gradient dominance constant and, thus, convergence decay rate, since μ must be taken to be the loosest constant amongst those of the maximizing set. In line with [Fazel et al., 2018], we assume $X_K \succcurlyeq 0$ across $\widehat{C} \in \mathcal{C}$ and $K \in \mathcal{K}(\mathcal{C})$. This is true if the system is controllable for any $\widehat{C} \in \mathcal{C}$, which holds if the nominal dynamics are well-behaved and the predictor $f(\theta)$ is sufficiently accurate, resulting in a small $\mathcal{C}$ set. The statements below are made for a general set of dynamics $\mathcal{C}$, though we are interested in $\mathcal{C} := \mathcal{U}(\theta)$. We defer the presentation of the explicit poly-expression in Theorem 4.2.4 to Section C.8.9.

Theorem 4.2.4. *Let $J(K, C) := \sum_{t=0}^{\infty} (x_t^\top (Q + K^\top R K) x_t)$ for $w = 0$. Let $K^{(t)}$, $\phi(K) := \max_{C \in \mathcal{C}} J(K, C)$, and $K^*_{\text{rob}}(\mathcal{C}) := \arg\min_{K \in \mathcal{K}(\mathcal{C})} \phi(K)$ be the t-th iterate of Algorithm 6. Assume for each iterate t, the optimization over C converges, i.e. $C^{(T_C)} = C^*(K^{(t)})$, that $K^{(t)} \in \mathcal{K}(\mathcal{C})$, and that $X_K \succcurlyeq 0$ for all $\widehat{C} \in \mathcal{C}$ and $K \in \mathcal{K}(\mathcal{C})$. Denote $\nu := \min_{\widehat{C} \in \mathcal{C}} \min_{K \in \mathcal{K}(\mathcal{C})} \sigma_{\min}(X_K)$. If in Algorithm 6 $\eta_K \leq \min_{[\widehat{A}, \widehat{B}] \in \mathcal{C}} \text{poly}(\frac{\nu \sigma_{\min}(Q)}{J(K^{(0)}, C)}, \frac{1}{||\widehat{A}||}, \frac{1}{||\widehat{B}||}, \frac{1}{||R||}, \sigma_{\min}(R))$, then, there exists a $\gamma > 0$ such that $\phi(K^{(T)}) - \phi(K^*_{\text{rob}}(\mathcal{C})) \leq (1 - \gamma)^T (\phi(K_0) - \phi(K^*_{\text{rob}}(\mathcal{C})))$.*

Formally, such convergence is guaranteed only if iterates $K^{(t)}$ remain within $\mathcal{K}(\mathcal{U}(\theta))$. One modification to Algorithm 6 would involve projecting intermediate iterates to this stabilizing set by solving

$$\Pi_{\mathcal{K}(\mathcal{U}(\theta))}(\widetilde{K}^{(t)}) := \arg\min_K ||K - \widetilde{K}^{(t)}||_{\text{op}}$$
$$\text{s.t.} \max_{[\widehat{A}, \widehat{B}] := \widehat{C} \in \mathcal{U}(\theta)} \max_i \text{Re}(\lambda_i(\widehat{A} + \widehat{B}K)) < 0.$$

There, however, is no known efficient algorithm to solve this projection step. Despite being of theoretical concern, this instability issue fails to be practically relevant, since the controller iterates remain well within the set of stabilization for sufficiently accurate predictors $f(\theta)$. If instabilities arise, an approximate solution can be obtained by replacing $\max_{\widehat{C} \in \mathcal{U}(\theta)}$ of Equation (4.4) with a finite sampling $\{\widehat{C}^{(i)}\}$ over $\mathcal{U}(\theta)$.

4.2.3 Related Works

Robust control can be broadly categorized into trajectory-based and trajectory-free robustness. The former adjusts an initially posited control scheme in an *online* fashion based on feedback measurements [Azad and Herber, 2022, Seiler et al., 2020, Paraskevopoulos, 2017], whereas the latter directly incorporates desired robustness into the optimization problem *prior* to deployment [Gravell and Summers, 2020, Gravell et al., 2022]. Given that control co-design seeks to identify a controller *prior* to deploying a design, we specifically highlight methods of trajectory-free robust control.

A popular classical trajectory-free method is $\mathcal{H}_\infty$ control. $\mathcal{H}_\infty$ is typically formulated as minimizing $||T_{wz}||_\infty$, i.e. the frequency space transfer function from $w \to z$ for some performance state z. By defining $z = \begin{bmatrix} Q^{1/2}x & R^{1/2}u \end{bmatrix}$, we recover a recognizable LQR formulation, with the objective replaced by

$$u^* = \min_{\{u_t\}} \max_{\{w_t\}} \sum_{t=0}^{\infty} (x_t^\top Q x_t + u_t^\top R u_t - \gamma^2 w_t^\top w_t), \tag{4.5}$$

which can be solved via generalized Riccati equations [Başar and Bernhard, 2008]. Here, γ can either be fixed to perform suboptimal $\mathcal{H}_\infty$ synthesis or it can be determined via bisection to identify the smallest γ such that a solution exists. Notably, the nominal $\mathcal{H}_\infty$ formulation seeks additive, unstructured disturbance rejection. Of interest herein, however, was robustness to *multiplicative* uncertainties through the system dynamics. Towards this end, μ-synthesis offers an extension to $\mathcal{H}_\infty$ control by allowing users to specify norm-bounded uncertainties on system dynamics [Bevrani et al., 2015, Chen et al., 2014].

This need for manual specification in μ-synthesis, however, incurs conservatism or controller instability if poorly specified, resulting in increasing interest in data-driven specifications. In this vein, a formulation known as LQR with multiplicative noise (LQRm), has recently become of interest, where the controller is:

$$\begin{aligned} K^* &:= \arg\min_K \mathbb{E}_{\{\delta_i\},\{\gamma_i\}}[J(K,A,B)] \\ \dot{x} &:= \left(A + \sum_{i=1}^{p} \delta_i A_i\right) x + \left(B + \sum_{i=1}^{q} \gamma_i B_i\right) u + w, \end{aligned} \tag{4.6}$$

where $\{A_i\}$ and $\{B_i\}$ and the distributions of $\{\delta_i\} \sim \mathcal{D}_\delta$ and $\{\gamma_i\} \sim \mathcal{D}_\gamma$ can be specified, either with data-free or with data-driven estimation. Most common among data-free specifications are so-called "margin methods." Briefly, margin methods specify $\{\delta_i\}$ and $\{\gamma_i\}$ by finding those $\{\delta_i\}$ and $\{\gamma_i\}$ that result in borderline-stable dynamics when paired with the corresponding, manually

Table 4.3: Each of the results below are the p-values of paired t-tests conducted pairwise between methods testing $H_1 : \mathcal{R}_{\%}^{(\mathrm{CPC})} < \mathcal{R}_{\%}^{(\mathrm{alt})}$ over 1,000 i.i.d. test samples. For any comparison method with $> 80\%$ unstable cases (see Table C.11 for percentages), we have marked the entry with "—".

	Airfoil	Load Positioning	Furuta Pendulum	DC Microgrids	Fusion Plant
Random Critical	—	—	—	—	—
Random OL MSS (Weak)	**0.0003**	—	—	—	—
Random OL MSUS	—	—	—	—	—
Row-Col Critical	—	—	—	—	—
Row-Col OL MSS (Weak)	**0.0117**	—	—	—	—
Row-Col OL MSUS	**0.0009**	—	—	—	—
Shared Lyapunov	**0.0001**	**0.0000**	**0.0112**	0.0913	**0.0023**
Auxiliary Stabilizer	**0.0001**	**0.0005**	**0.0055**	**0.0428**	0.0630
$\mathcal{H}_\infty$	**0.0009**	**0.0000**	**0.0071**	**0.0004**	**0.0013**

specified $(\{A_i\}, \{B_i\})$ and some choice of controller: the particular controller varies across margin strategies. A full description of the margin methods considered is given in Section C.8.10.

As with $\mathcal{H}_\infty$, such data-free LQRm methods sacrifice stability or risk conservatism in ignoring the nature of the dynamics predictor misspecification, resulting in recent works that give data-driven parameterizations [Gravell et al., 2020b]. Here, two approaches were proposed to learn the margin parameters, which we refer to as the "Shared Lyapunov" and "Auxiliary Stabilizer" approaches, described fully in their paper. While such approaches improve upon the conservatism of classical data-free margin methods, they still require the hand specification of the perturbation matrices $\{A_i\}$ and $\{B_i\}$.

4.2.4 Experiments

We now study five setups of interest in the infinite horizon, discrete-time, deterministic setting, namely LQR control of an airfoil [Chrif and Kadda, 2014], a load positioning system [Ahmadi et al., 2023, Jiang et al., 2016], a Furuta pendulum [Arulmozhi and Victorie, 2022], a DC microgrid [Liu et al., 2023a], and a nuclear plant [Kirgni and Wang, 2023]. The dimensions of (θ, A, B) are $(\mathbb{R}^{15}, \mathbb{R}^{4\times4}, \mathbb{R}^{4\times2})$ for the airfoil, $(\mathbb{R}^{5}, \mathbb{R}^{4\times4}, \mathbb{R}^{4\times1})$ for the load positioning system, $(\mathbb{R}^{9}, \mathbb{R}^{4\times4}, \mathbb{R}^{4\times1})$ for the Furuta pendulum, $(\mathbb{R}^{17}, \mathbb{R}^{9\times9}, \mathbb{R}^{9\times1})$ for the DC microgrid, and $(\mathbb{R}^{26}, \mathbb{R}^{8\times8}, \mathbb{R}^{8\times1})$ for the nuclear plant. The full setup details are provided in Section C.8.11.

We compare against $\mathcal{H}_\infty$ control with γ bisection, the data-free margin methods, and the data-based methods for the LQRm setup as discussed in Section 4.2.3. The data-free margin methods are as implemented by [Gravell and Summers, 2020] and are fully described in Section C.8.10, of which we specifically consider "Random Critical," "Random OL MSS (Weak)," "Random OL MSUS," "Row-Col Critical," "Row-Col OL MSS (Weak)," and "Row-Col OL MSUS." The data-

based methods are the "Shared Lyapunov" and "Auxiliary Stabilizer" approaches from [Gravell et al., 2020b]. As discussed, we are considering the trajectory-*free* setting, so we do not compare against methods that achieve robustness adaptively over trajectories, such as those in [Gravell and Summers, 2020, Gravell et al., 2022].

In the experiments, we construct $\mathcal{D}$ as per Section 4.2.2.1, using random gain matrices $K^{(i)}$. N was taken to be $2,000$ with $|\mathcal{D}_{\mathcal{C}}| = 400$ and the remaining $\mathcal{D}_T$ used to train $f(\theta)$, taken to be feed-forward neural networks.

4.2.4.1 Robust Control Regret & Stability

We first study the empirical regret across the aforementioned systems and robust control methods over $1,000$ i.i.d. test points from $\mathcal{P}(\Theta, C)$. To make results comparable across $\theta^{(i)}$, we normalize each trial by its nominal objective, i.e., $\mathcal{R}_{\%} = \mathcal{R}(\Theta, C)/J(K^*(C), C)$, as in [Sun et al., 2023]. If the uncertainty regions of the robust problems are poorly specified, i.e. if the regions of robustness do not capture the true dynamics, the resulting robust controller may have unbounded cost, i.e. $J(K^*_{\mathrm{rob}}, C) = \infty$. We, thus, only compute $\mathcal{R}_{\%}$ over the stabilizing controllers and separately report the proportion of destabilized cases. Lower values are desirable for both.

For each comparison method, we report the result of a one-side paired t-test of $H_1 : \mathcal{R}_{\%}^{(\mathrm{CPC})} < \mathcal{R}_{\%}^{(\mathrm{alt})}$ in Table 4.3. We defer the presentation of the raw regret values and the percent of cases with stabilized dynamics to Section C.8.12 due to space constraints. Notably, the alternative approaches generally incur greater regret than CPC. For $\mathcal{H}_\infty$, this is expected as the misspecification here is in the dynamics matrices, differing from the adversarial exogenous noise that $\mathcal{H}_\infty$ is designed to protect against. Similarly, the data-free margin methods protect against perturbations that are misaligned with the true dynamics misspecification, which result in significant instability for the higher-dimensional problems (i.e. the "Furuta Pendulum," "DC Microgrids," and "Fusion Plant" tasks). The data-driven LQRm methods improve significantly upon these margin approaches in stability, yet they are too conservative as they do not make use of the anticipated structures of the errors made in the predictions by $\widehat{f}(\theta)$.

4.2.4.2 Ambiguous Ground Truth Calibration

To validate the results of Theorem 4.2.3 and demonstrate the empirical validity of the associated assumption, we computed the empirical coverages across various levels of desired coverage $\alpha \in (0, 1)$ for the experimental setups. As previously discussed, the calibration here was performed using a calibration set of *estimated* dynamics $\widetilde{\mathcal{D}} = \{(\theta^{(i)}, \widetilde{C}^{(i)})\}$ but coverage was assessed on the *true* dynamics $\{C^{(i)}\}$. We computed this in the manner described in Section 4.2.4 for α varying by increments of 0.05. For assessing coverage, we again used $1,000$ test points drawn i.i.d. from

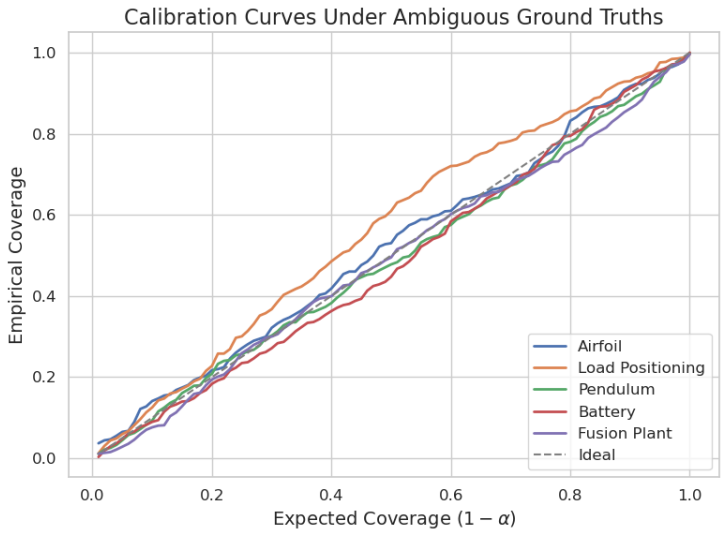

Figure 4.1: Calibration plots for the tasks, assessed on 1,000 i.i.d. test samples of C with calibration performed using the estimated $\widetilde{C}$, affirming Theorem 4.2.3.

$\mathcal{P}(\Theta, C)$ and measured the proportion of samples for which $s(\theta^{(i)}, C^{(i)}) \leq \widehat{q}$. The results are shown in Figure 4.1, where we see the desired calibration under calibration with estimated dynamics.

CHAPTER 5

Future Directions

> If we knew what it was we were doing, it would not be called research, would it?
>
> Albert Einstein

Over the previous chapters of this thesis, we have developed a generalizable framework for robust decision-making that leverages conformal prediction, demonstrating its use in scientific decision-making, single-stage decision making problems for both standard and ensembled predictors, and finally robust linear controls problems. Critically, while this developed framework encompasses a diverse collection of applications, there remain a number of routes for further development. We specifically highlight some directions of recent investigation that seem particularly compelling and present the associated preliminary results along such directions. Notably, these directions all generalize the decision-making framework along the lines of leveraging conformal prediction over infinite-dimensional function spaces.

5.1 Non-Parameteric Conformal Distributionally Robust Optimization

As a direct extension on the ideas presented in Chapter 3, we propose leveraging conformal prediction to make guarantees on *stochastic* contextual predict-then-optimize tasks. Stochastic optimization is a mature field often used in safety-critical situations, such as in the deployment of self-driving cars [Ben-Tal et al., 2009, Gabrel et al., 2014, Beyer and Sendhoff, 2007, Rahimian and Mehrotra, 2019]. Traditionally, problems are framed as seeking decisions $w \in \mathcal{W}$ that minimize $\mathbb{E}_{\mathcal{P}(C)}[f(w, C)]$ for some objective f and random information dictating the decision quality w.

In cases where contextual information is present, this requires explicit modeling of the posterior $\mathcal{P}(C \mid X)$ distribution. Similar to the settings discussed in Chapter 2, amortized variational inference is frequently employed, with decisions subsequently made against $\mathbb{E}_{q_{\varphi(x)}(C)}[f(w, C)]$. While variational inference provides a computationally efficient alternative to MCMC, it has been repeatedly noted that a major shortcoming of VI is its lack of any theoretical guarantees and tendency to produce biased posterior estimates [Blei et al., 2017, Murphy, 2022, Zhang et al., 2018, Yao et al., 2018].

Separately, distributionally robust optimization (DRO) arose, in which solutions of this optimization set are instead sought over an ambiguity set $\mathcal{U}(\mathcal{P})$ of distributions [Rahimian and Mehrotra, 2019, Kuhn et al., 2019, Lin et al., 2022]. Significant progress has been achieved in this vein, but DRO requires a priori knowledge of plausible ambiguity sets or noise distributions to produce answers that are practically useful. For instance, an overly conservative ambiguity set will likely result in suboptimal performance in typical circumstances. Towards this end, data-driven DRO has recently become of interest, in which plausible ambiguity sets are learned empirically [Delage and Ye, 2010, Mohajerin Esfahani and Kuhn, 2018, Chen et al., 2022].

We, therefore, extend CPO and propose CDPO (Conformal-Distributional-Predict-Then-Optimize), a procedure that leverages conformal prediction to produce prediction regions over probability measures and thereby produces guarantees on stochastic decision-making algorithms that rely on amortized variational inference, in turn unifying the fields of distributionally robust optimization and predict-then-optimize decision-making.

5.1.1 CDPO: Score Function

Let $c \in \mathcal{C}$, where $(\mathcal{C}, d)$ is a general metric space, and $\mathcal{F}$ be the σ-field of $\mathcal{C}$. We again consider general convex-concave objective functions $f(w, c)$ that are L-Lipschitz in c under the metric d for any fixed w. With this generalization, the robust formulation of stochastic predict-then-optimize

can be stated as

$$\begin{aligned} w^*(x) := \inf_{w \in \mathcal{W}} \sup_{\widehat{\mathcal{Q}} \in \mathcal{U}(x)} \quad & \mathbb{E}_{\widehat{\mathcal{Q}}}[f(w, C)] \\ \text{s.t.} \quad & \mathcal{P}_{X, \mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha, \end{aligned} \tag{5.1}$$

where $\mathcal{U} : \mathcal{X} \to \mathcal{M}(\mathcal{F})$ is a uncertainty region predictor over the space of probability measures on $\mathcal{F}$. Exact solution of this problem is intractable, as no practical methods exist to optimize over the measure space $\mathcal{U}$. For any fixed $\mathcal{U}$, this robust counterpart to the stochastic predict-then-optimize problem produces a valid upper bound if we use the following score function:

$$s(x, \mathcal{P}_C) = \mathcal{W}_1(\widehat{\mathcal{Q}}_{\varphi(x)}(C), \mathcal{P}_C), \tag{5.2}$$

where $\mathcal{W}_1$ represents the 1-Wasserstein distance. To compute the quantile $\widehat{q}$ of such a score over $\mathcal{D}_{\mathcal{C}}$, we assume the recovery of the exact posterior $\mathcal{P}(C \mid x)$ for a subset of x, namely via MCMC methods. From here, $\mathcal{C}(x) = \{\mathcal{Q} \mid s(x, \mathcal{Q}) \leq \widehat{q}(\alpha)\}$ has marginal guarantees in the form $\mathcal{P}_{X, \mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$. Notably, even computing $\mathcal{W}$ for multi-dimensional distributions is a computationally challenging task; however, we can use the well-known equivalence between computing $\mathcal{W}_1$ and the Assignment Problem, which can be solved in $\mathcal{O}(N^3)$ with the Hungarian Algorithm [Peyré et al., 2019].

With this choice of score function, we can bound the nominal stochastic optimal value. Let

$$\Delta(x, \mathcal{P}_C) := \inf_{w \in \mathcal{W}} \sup_{\widehat{\mathcal{Q}} \in \mathcal{U}(x)} \mathbb{E}_{\widehat{\mathcal{Q}}}[f(w, C)] - \inf_{w \in \mathcal{W}} \mathbb{E}_{\mathcal{P}_C}[f(w, C)].$$

We clearly see $\Delta(x, \mathcal{P}_C) \geq 0$ if $\mathcal{P}_C \in \mathcal{U}(x)$. This framing again makes clear the consequences of leveraging *efficient* prediction regions with guaranteed coverage, formalized below.

Lemma 5.1.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X, \mathcal{P}_C}(\mathcal{P}_C \in \mathcal{U}(X)) \geq 1 - \alpha$ with $\sup_{q \in \mathcal{U}(x)} \mathcal{W}_1(\mathcal{Q}, \mathcal{P}_C) = \operatorname{diam}(\mathcal{U}(x))$. Then,*

$$\mathcal{P}_{X, \mathcal{P}_C}\left(\Delta(X, \mathcal{P}_C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \tag{5.3}$$

The proof is explicitly provided in Section D.1. Thus, $1 - \alpha$ validity of the prediction region ensures the result of the RO procedure is a valid bound with probability $1 - \alpha$, and greater efficiency of the prediction region translates to a tighter upper bound.

5.2 Uncertainty Quantification for Dynamic NeRFs

The ultimate application aim of this line of distribution-free uncertainty quantification is for volumetric reconstruction of proteins from Cryo-EM. We pose the reconstruction task as an inverse

problem of recovering $\phi : \mathbb{R}^3 \to \mathbb{R}$ from observations of the following generative model:

$$\mathcal{I} = H * \int_{-\infty}^{\infty} \phi(R_i^T \vec{x} + \vec{t_i}) dz + \eta, \tag{5.4}$$

where $\mathcal{I} \in \mathbb{R}^{n \times n}$ are the observed images, H is the point-spread function, $R_i, \vec{t_i}$ are the rotation and translation of the density, and η is observation noise. Such reconstruction closely mirrors the traditional problem of volumetric reconstruction from computer vision, specifically in the field of photogrammetry. We, therefore, will pursue progress in this vein prior to redirecting it to the more complex Cryo-EM data.

Volumetric reconstruction inspired by or directly using neural radiance fields (NeRFs) is becoming increasingly prevalent in scientific applications, such as in Cryo-EM reconstruction [Zhong et al., 2021a,b, Levy et al., 2022b, Gupta et al., 2021, Nashed et al., 2021, Levy et al., 2022a, Scheres, 2012, Ullrich et al., 2019, Chen and Ludtke, 2021]. Applications of NeRFs have traditionally been in artistic domains, where recovery of precision is of secondary concern to the quality of user experience. However, in scientific inquiries, precision and accompanying uncertainty quantification are of utmost importance.

Neural radiance fields have demonstrated incredible fidelity in novel view synthesis from input views [Mildenhall et al., 2021], having spawned an entire sub-domain of research purely focused on such NeRF reconstruction [Gao et al., 2022]. However, such reconstruction tasks are often inherently underconstrained problems, meaning a whole distribution of depths could map to the observed images. NeRFs excel at producing point estimates of the volume. They, however, give no sense of the uncertainty associated with such predictions.

The rising need for such uncertainty quantification can be seen in the recent work in the NeRF community on Bayesian uncertainty quantification [Ritter and Karaletsos, 2021, Shen et al., 2021, Sünderhauf et al., 2022, Shen et al., 2022, Hoffman et al., 2023]. These works, however, modify the training process and requisite architectures greatly, limiting their adoption by end users. In this vein, a recent work [Goli et al., 2023] introduced a post-hoc method for uncertainty quantification relying on Laplace approximations. This work, however, is limited to static reconstruction, making it inapplicable to Cryo-EM reconstruction.

We, therefore, propose an extension to this method to enable its use for dynamic object reconstruction. We view such a contribution as a first step toward providing principled uncertainty quantification for Cryo-EM reconstruction methods and as the first post-hoc method for doing uncertainty quantification for dynamic reconstruction.

5.2.1 Neural Radiance Fields (NeRFs)

While modern machine learning has had some impact on SfM algorithms, it has had noticeably more in reconstruction algorithms, particularly in the form of neural radiance fields (NeRFs). At their core, the core idea of NeRFs is to directly estimate the 3D volume rather than going through the intermediary of the depth. In particular, we estimate such a volume as $\Phi : \mathbb{R}^5 \to \mathbb{R}^4$, where the inputs are the point inside the volume $x := (x, y, z)$ and viewing direction $d := (\theta, \phi)$ and the outputs are the color $c := (r, g, b)$ and transmission density τ. An image for a given viewpoint is produced by "volumetric rendering," i.e. for a ray $r(t)$ corresponding to a pixel, we compute

$$C(r) = \int_{t_0}^{t_1} T(t)\tau(r(t))c(r(t), d)dt,$$
$$\text{where } T(t) = \exp\left(-\int_{t_1}^{t} \tau(r(u))du\right).$$

Thus, F is trained against the standard L_2 loss, i.e. $\mathcal{L} := ||C - \widehat{C}||_2^2$, across images. Note that this assumes the camera pose is known a priori, which is in practice typically obtained via conventional SfM algorithms.

For dynamic reconstruction, we extend this pipeline by introducing a deformation field Φ_t, which conceptually warps a static "canonical frame" represented now by Φ_x to the current frame. That is, $\Phi_t(x, t) \to \Delta x$, maps a position x in the reconstructed canonical volume to its adjusted position at time t, meaning the reconstruction at time t is now represented by $\Phi_x(x + \Delta x)$. For instance, a canonical frame may reconstruct a person standing upright, with $\Phi_t(x, t)$ warping their legs to appropriately bend if they are seated in frame t.

5.2.2 Static Uncertainty Quantification

We briefly summarize the method as presented in [Goli et al., 2023] before proposing its extension to dynamic object reconstruction. In this setup, we assume prior knowledge of the training camera poses, $\{\mathcal{T}_n\}$. We are specifically interested in quantifying epistemic uncertainty here, that is, highlighting regions of the reconstruction that are underconstrained by the input views. For instance, a single image is, without prior knowledge, insufficient to determine the depths of contained scene elements. With more viewpoints, the reconstruction becomes more constrained. We, therefore, wish to capture those regions of the reconstruction which, upon perturbation, would result in immaterial changes in the viewpoints observed in the fitting process. For instance, if only frontal views of a person are observed during fitting, any perturbation to regions of the volume corresponding to their back would not be observable in projecting to $\{\mathcal{T}_n\}$, which we wish to highlight.

Such a field conceptually acts much in the way the deformation was described in section 5.2.1. We, therefore, formalize "spatial uncertainty" as a property of this aforementioned perturbation field, detailed as follows. Assume we have fitted $\Phi_{\theta^*} := (c_{\theta^*}, \tau_{\theta^*})$. Denoting the perturbation field as $\mathcal{D}_\varphi(x)$, rendering under perturbation involves rendering with the modified functions:

$$\widetilde{c}_\varphi(x) := c_{\theta^*}(x + \mathcal{D}_\varphi(x), d)$$
$$\widetilde{\tau}_\varphi := \tau_{\theta^*}(x + \mathcal{D}_\varphi(x)).$$

We denote the modified rendering result, i.e. section 5.2.1 using $(\widetilde{c}_\varphi(x), \widetilde{\tau}_\varphi)$ as $\widetilde{C}(r)$. We, therefore, wish to characterize the φ under which observations $\widetilde{C}(r)$ are unchanged. Critically, unlike the deformation field of section 5.2.1, $\mathcal{D}_\varphi(x)$ is explicitly defined, specifically as

$$\mathcal{D}_\varphi(x) := \text{Trilinear}(x, \varphi), \tag{5.5}$$

where $\varphi \in \mathbb{R}^{M^3 \times 3}$ for M a discretization resolution of the reconstructed volume. That is, for each vertex m of a discretized representation of Φ, we associate a corresponding displacement vector φ_m. We, therefore, wish to find the posterior $\mathcal{P}(\varphi \mid X)$ of these parameters, with the variance of the appropriate component of φ being the desired spatial uncertainty characterization of interest. Important to note is that θ^* is *not* modified in this post-hoc analysis.

We now leverage Laplace approximation to obtain the desired posterior, modeling the prior as $\varphi \sim \mathcal{N}(0, \lambda^{-1})$. The derivation is in [Goli et al., 2023], with the final covariance being

$$\mathcal{I}(\varphi)^{-1} \approx \left(\frac{2}{R} \sum_r J_\varphi(r) J_\varphi(r)^T + 2\lambda I \right)^{-1}, \tag{5.6}$$

where $J_\varphi(r)$ is the Jacobian of $\widetilde{C}$ evaluated at r, namely $J_\varphi(r) := \frac{\partial \widetilde{C}}{\partial \varphi}$. Only the diagonal entries of this matrix are considered due to previous studies on the sparsity structure of NeRF parameter correlations, from which we obtain the desired characterization of vertex displacement variances.

The static version immediately lends itself to an extension to dynamic reconstruction. We specifically aim to produce uncertainty quantification in the warped frames, i.e. by characterizing a perturbation field on $\Phi_x(x + \Delta x + \mathcal{D}_\theta(x))$, where $\Delta x = \Phi_t(x, t)$.

5.3 Conformally Robust Engineering Design

Much of engineering design centers on optimizing design parameters under a PDE-constrained functional, such as optimizing car or aircraft designs for drag minimization or structural designs for withstanding stress [Sokolowski et al., 1992, Hsu, 1994, Challis and Guest, 2009, Dunning and

Kim, 2015]. Traditionally, workflows would require repeatedly running domain-specific numerical PDE solvers to evaluate the designs as they were iteratively refined [Zhang et al., 2006, Caron et al., 2025]. Such workflows, however, suffered from slow iteration time, as numerical PDE solvers incur a significant computational cost that cannot be amortized over their runs across different designs. For this reason, significant interest has arisen in neural surrogate models that learn a "flow map," with which a PDE can be efficiently approximately solved across different input conditions with an inference pass [Pathak et al., 2022, Wen et al., 2022, Bonev et al., 2023].

One concern with purely relying on such surrogate models in design optimization pipelines is that they lack any guarantees of recovering the true solution functions, unlike classical numerical solvers. This, in turn, has led to the proliferation of methods to provide uncertainty estimates for such models [Zou and Karniadakis, 2023, Psaros et al., 2023, Zhu et al., 2019, Martina-Perez et al., 2021, Tripathy and Bilionis, 2018]. Such uncertainty quantification methods, however, rely on distributional assumptions, whose utility fails with distributional misspecification [Sahin et al., 2024, Mollaali et al., 2023]. For this reason, recent efforts have been directed towards developing distribution-free, data-driven approaches to uncertainty quantification by leveraging conformal prediction [Gopakumar et al., 2024, Ma et al., 2024], a principled framework for producing distribution-free prediction regions with marginal frequentist coverage guarantees [Angelopoulos and Bates, 2021, Shafer and Vovk, 2008].

These initial efforts to leverage conformal prediction, however, fall short along two axes. The first is that they fail to provide coverage of the infinite-dimensional functions being predicted by neural operators. Initial works in this direction only produced coverage guarantees on predictions of a fixed-discretization, in turn sacrificing the discretization-invariant property that is central to neural operators [Gopakumar et al., 2024, Ma et al., 2024]. A recent work took steps towards addressing this deficiency by guaranteeing simultaneous coverage up to some maximally observed resolution; this approach, however, still fails to provide coverage over the untruncated function space [Gray et al.]. In addition, such uncertainty quantification has yet to be leveraged for downstream use cases. The space of finite-dimensional conformal prediction followed a similar trend, with the proliferation of methods that produce calibrated regions with only more recent works discovering their applicability to decision-making tasks [Lekeufack et al., 2024, Cresswell et al., 2024, Kiyani et al., 2025, Cortes-Gomez et al., 2024]. We present initial results toward this end in Section D.2.

APPENDIX A

Amortized Variational Inference with Coverage Guarantees

A.1 Group Conditional CANVI

We now discuss how CANVI can be easily extended to provide a stronger notion of group conditional coverage over the purely marginal coverage over $\mathcal{X}$ that was provided in the paper. Here, instead of defining a global $\widehat{q}$, a collection $\{\widehat{q}_i\}_{i=1}^K$ is obtained by defining centroids over the $\mathcal{X}$ space $\{c_i\}_{i=1}^K$, constructing balls of radius ϵ around them $\mathcal{B}_\epsilon(c_i) \subset \mathcal{X}$, and performing calibration using $\mathcal{D}_\mathcal{C}^i := \{(x_j, \theta_j)\}_{j=1}^{\mathcal{N}_C}$ with $x_j \in \mathcal{B}_\epsilon(c_i)$.

To compare, we define a *global* quantile $\widehat{q}$ using KN samples. Specifically, we take $N = 100,000$ and $K = 5$, meaning calibration was performed using 500,000 i.i.d. samples for the *overall* calibration set. Coverage was assessed over $\alpha \in [0, 1]$ discretized at steps of .05 over $K = 5$ regions. Coverage was assessed over 10 batches of 100,000 i.i.d. test samples.

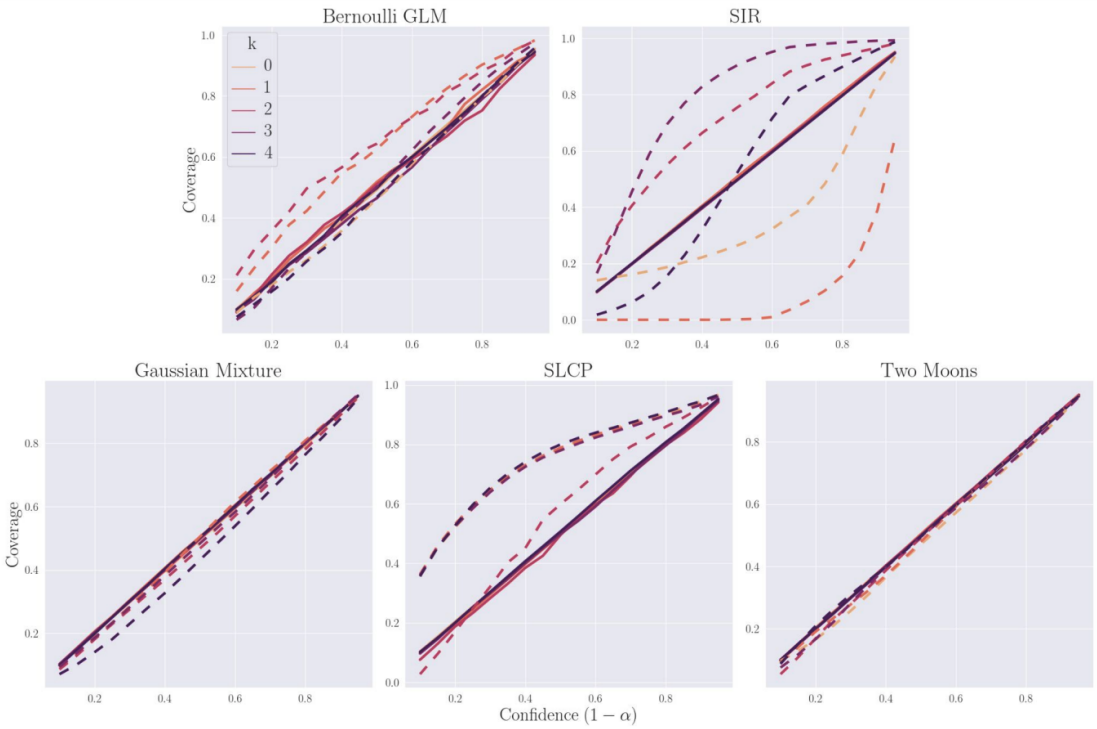

Figure A.1: Calibration on the SBI benchmarks with overall quantile and region-specific quantiles, respectively the dashed and solid lines. Region-specific conformalized lines are slightly difficult to distinguish, as they all lie along the desired $y = x$ curve. Error bars from coverage assessments across test batches are plotted, although they are difficult to see due to the low variance between estimates across batches.

Figure A.1 demonstrates the miscalibration of the regions produced using the overall quantile and subsequent correction with region-specific calibration. We additionally plot the calibration across the X space of the Two Moons task to demonstrate the non-uniformity of coverage across regions and subsequent corrections in Figure A.2.

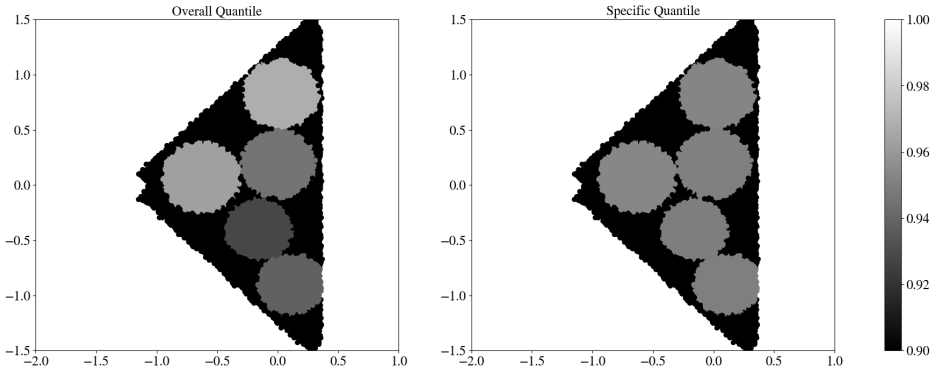

Figure A.2: Coverage for Two Moons with the overall and region-specific quantiles with $\alpha = 0.05$.

A.2 CANVI Validity

Lemma A.2.1. *Let* $\alpha \in (0,1)$ *and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \\ \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X,\Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

Let $(x', \theta') \sim \mathcal{P}(X,\Theta) \perp\!\!\!\perp \mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{R}} \cup \mathcal{D}_{\mathcal{T}}$*, with* $\mathcal{D}$ *being the data used to train* $\{q^{(t)}(\Theta|X)\}_{t=1}^{T}$*. Then* $1-\alpha \leq \mathcal{P}(1/q^{(*)}(\theta'|x') \leq \widehat{q}_{\mathcal{R}}^{(*)}(\alpha))$.

Proof. Consider the score function $s^{(*)}(x,\theta) := 1/q^{(*)}(\theta \mid x)$. Observe that $(x',\theta') \cup \mathcal{D}_{\mathcal{R}}$ are jointly sampled i.i.d. from $\mathcal{P}(X,\Theta)$, independent of the datasets used to design $s^{(*)}(x,\theta)$, namely $\mathcal{D} \cup \mathcal{D}_{\mathcal{C}} \cup \mathcal{D}_{\mathcal{T}}$. Denoting $\mathcal{S}_{\mathcal{R}} := \{s^{(*)}(x_i,\theta_i)\}_{(x_i,\theta_i)\in\mathcal{D}_{\mathcal{R}}}$, scores $s^{(*)}(x',\theta') \cup \mathcal{S}_{\mathcal{R}}$, thus, too are i.i.d and, hence, exchangeable. The coverage guarantee then follows from the general theory of conformal prediction, presented in [Angelopoulos and Bates, 2021]. ■

A.3 Riemannian Manifolds

Let $(\mathcal{M}, g)$ denote a Riemannian manifold, with g denoting the metric tensor associated with this space. By definition, a manifold $\mathcal{M}$ is locally isomorphic to Euclidean space, from which we can define the notion of a tangent space $\mathcal{T}_z\mathcal{M}$ at each point $z \in \mathcal{M}$. The metric tensor g then defines a *local* notion of distance, namely $g : \mathcal{T}_z\mathcal{M} \times \mathcal{T}_z\mathcal{M} \to \mathbb{R}$. This, therefore, induces a local notion of length, namely for $x \in \mathcal{T}_z\mathcal{M}, ||x||_z = \sqrt{g(x,x)}$.

Global distances over the manifold, therefore, can then be denoted by integrating such a local notion across a path γ. Concretely, a path is defined between $a, b \in \mathcal{M}$ by $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = a, \gamma(1) = b$. The length of such a path is then $\ell(\gamma) := \int_0^1 ||\gamma'(t)||_{\gamma(t)} dt$. A natural choice of distance, therefore, is the minimum length of a constant velocity path, formally

$$d(a,b) = \inf_{\gamma : ||\gamma'(t)||=1, \gamma(0)=a, \gamma(1)=b} \ell(\gamma) \tag{A.1}$$

A.4 CANVI: Efficiency Analysis

A.4.1 Lipschitz Continuity of Efficiency

Lemma A.4.1. *Let* $\ell(q,\tau)$ *be as defined in Equation 2.2. If* $\ell_x(q,\tau)$ *is L-Lipschitz continuous in* τ *for any* $x \in \mathcal{X}$*, then* $\ell(q,\tau)$ *is L-Lipschitz continuous.*

Proof.

$$
\begin{aligned}
|\ell(q,\tau_1) - \ell(q,\tau_2)| &= \left|\mathbb{E}_X\left[\mathcal{L}\left(\{\theta : 1/q(\Theta \mid X) \le \tau_1\}\right) - \mathcal{L}\left(\{\theta : 1/q(\Theta \mid X) \le \tau_2\}\right)\right]\right| \\
&= \left|\int \mathcal{P}(x)\left[\ell_x(q,\tau_1) - \ell_x(q,\tau_2)\right]\right| \le \int \mathcal{P}(x)\left|\ell_x(q,\tau_1) - \ell_x(q,\tau_2)\right| dx \\
&\le L\left|\tau_1 - \tau_2\right| \int \mathcal{P}(x)dx = L\left|\tau_1 - \tau_2\right|,
\end{aligned}
$$

completing the proof as desired. ■

A.4.2 Predictive Efficiency of CANVI

The proof emerges through a reduction of CANVI to a special case of the "Validity First Conformal Prediction for the Smallest Prediction Set" (VFCP) algorithm presented in the [Yang and Kuchibhotla, 2021]. The VFCP algorithm is provided for convenience in Algorithm 7. Note that we made the appropriate replacements in notations to match those presented in the main body of our paper for clarity.

To further underscore the parallel, we present an immaterially modified version of VFCP, where it is assumed the data are split *prior* to the algorithm execution into disjoint sets $\mathcal{D}, \mathcal{D}_{\mathcal{C}}$, and $\mathcal{D}_{\mathcal{R}}$. We assume input prediction methods $\mathcal{A}^{(1)}, ..., \mathcal{A}^{(T)}$ have been trained on $\mathcal{D}$. Finally, the presentation in [Yang and Kuchibhotla, 2021] allows for generic definitions of the measure function of $\mathcal{C}(x)$, referred to as "Width" therein. We present it with the particular Lebesgue measure function as presented in the main body to avoid confusion. These three modifications in the algorithm presentation have no manifest effects in the proof of [Yang and Kuchibhotla, 2021], meaning all results as presented there apply with the appropriate choices of inputs in Algorithm 7.

We similarly consider an immaterially modified form of CANVI, that is, one in which $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ are pre-generated in precisely the fashion outlined Algorithm 1 and provided as input. Thus, inputs to CANVI parallel those of VFCP, namely $\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, 1-\alpha, D_{\mathcal{C}}$, and $D_{\mathcal{R}}$.

Intuitively, the proof follows from the fact that CANVI is a special case of Algorithm 7, where the particular structural assumptions we make imply the corresponding assumptions as stated in the theorem from [Yang and Kuchibhotla, 2021]. Algorithm 7 is presented using the so-called "nested set formulation" of conformal prediction. Briefly, the nested sets formulation starts by requiring the user to specify a *set-valued* function F_r instead of a score function. Sets must be nested over increasing r. Despite being seemingly more expressive, the equivalence of the nested set and split score conformal methods was demonstrated in [Gupta et al., 2022]. In particular, for a chosen score function $s(x,\theta)$, the corresponding nested set function would be $F_\tau^{(t)} = \{(x,\theta) \mid s^{(t)}(x,\theta) \le \tau\}$, where $s^{(t)}$ denotes the score function as defined with $\mathcal{A}^{(t)}$. The presentation of

Algorithm 7 using the nested set formulation was, thus, a means to allow for the employment of relevant proof strategies from classical learning theory. We now proceed through the formal equivalence of CANVI and a special case of VFCP.

We first state for reference the original theorem for VFCP, which we leverage in the proof of theorem.

Theorem A.4.2. *Suppose Assumption 1 holds. Let $\alpha \in (0,1)$, $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ be drawn i.i.d. from $\mathcal{P}(X,\Theta)$, and*

$$\mathcal{C}_{\mathcal{R}}^{(*)} = \text{VFCP}\left(\mathcal{A}^{(t)}, 1-\alpha, D_{\mathcal{C}}, D_{\mathcal{T}}, s(x,\theta)\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ and $\delta \in [0,1]$, Assumption 2 holds, then with probability at least $(1-\delta)$,

$$\mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{R}}^{(*)}(X))] \leq \min_{1\leq t\leq T} \mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{R}}^{(t)}(X))] + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right], \quad \text{(A.2)}$$

where γ, L_W, and $L_{[T]} = \max_{1\leq t\leq T} L_t$ are constants defined in Assumptions 2 and 1 [Yang and Kuchibhotla, 2021].

We now present the proof of our theorem. The proof proceeds in two steps. We first demonstrate the CANVI and VFCP algorithms are equivalent if we assume access to the exact $\ell(q,\tau)$, which, coupled with the assumptions imposed on q, allows us to directly leverage the results of Theorem A.4.2. We then demonstrate, under Assumption 3, we recover the desired bound even if t^* is chosen with $\widehat{\ell}(q,\tau)$.

Theorem A.4.3. *Suppose for any $x \in \mathcal{X}$ and $t = 1,...,T$, $q^{(t)}(\theta \mid x) \in \mathcal{C}^3(\mathbb{R}^n)$ is bounded above and for $\theta \neq 0$, $\mathcal{L}(\{\theta : \nabla_\theta q^{(t)}(\theta|x)\}) = 0$. Further assume $P(X,\Theta)$ is bounded above. Let $\alpha \in (0,1)$ and*

$$q^{(*)}(\Theta \mid X), \widehat{q}_{\mathcal{R}}^{(*)}(\alpha) = \\ \text{CANVI}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, \mathcal{P}(X,\Theta), 1-\alpha, N_{\mathcal{C}}, N_{\mathcal{T}}\right)$$

If, for $r \geq \max\{\sqrt{\log(4T/\delta)/2N_{\mathcal{C}}}, 2/N_{\mathcal{C}}\}$ and $\delta \in [0,1]$, Assumption 2 holds and for $\Delta, \epsilon > 0$ Assumption 3 holds, then with probability at least $(1-\epsilon)(1-\delta)$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1\leq t\leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}}\right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}}\right)^{\gamma}\right], \quad \text{(A.3)}$$

where γ, L_W, and $L_{[T]} = \max_{1\leq t\leq T} L_t$ are constants defined in Assumptions 2 and 1.

Algorithm 7 VALIDITY FIRST CONFORMAL PREDICTION (VFCP) [Yang and Kuchibhotla, 2021]

procedure VFCP

Inputs: Predictors $\{\mathcal{A}^{(t)}\}_{t=1}^{T}$, Target coverage $1-\alpha$, Calibration set $\mathcal{D}_{\mathcal{C}}$, Recalibration set $\mathcal{D}_{\mathcal{R}}$, Score $s(x,\theta)$

Using $\mathcal{A}^{(t)}$, construct an increasing (nested) sequence of sets $\{F_\tau^{(t)}\}_{\tau\in\mathcal{T}}$, where $\mathcal{T}\subset\mathbb{R}$ and

$$F_\tau^{(t)} = \{(x,\theta) \mid s^{(t)}(x,\theta) \le \tau\},$$

where $s^{(t)}$ denotes the score function as defined with $\mathcal{A}^{(t)}$

Compute conformal prediction set $\mathcal{C}^{(t)}$ based on $\{F_\tau^{(t)}\}_{\tau\in\mathcal{T}}$. Specifically, for each $(x_i,\theta_i)\in\mathcal{D}_{\mathcal{C}}$ and $t\in\{1,2,...,T\}$, denote its corresponding score as

$$s^{(t)}(x_i,\theta_i) := \inf_{\tau\in\mathcal{T}}\{(x_i,\theta_i)\in F_\tau^{(t)}\}$$

Compute the corresponding conformal prediction set as

$$\mathcal{C}_{\mathcal{C}}^{(t)} := \{(x,\theta) : s^{(t)}(x,\theta) \le \widehat{q}_{\mathcal{C}}^{(t)}(\alpha)\},$$

where $\widehat{q}_{\mathcal{C}}^{(t)}(\alpha)$ is the $\lceil(|\mathcal{D}_{\mathcal{C}}|+1)(1-\alpha)\rceil$-th largest element of $\{s^{(t)}(x_i,\theta_i)\}_{i\in\mathcal{D}_{\mathcal{C}}}$

Let $\mathcal{C}_{\mathcal{C}}^{(t)}(x) := \{\theta : (x,\theta)\in\mathcal{C}_{\mathcal{C}}^{(t)}\}$. Set

$$t^* := \operatorname*{arg\,min}_{1\le t\le T}\mathbb{E}_X[\mathcal{L}(\mathcal{C}_{\mathcal{C}}^{(t)}(X))]$$

For each $(x_i,\theta_i)\in\mathcal{D}_{\mathcal{R}}$, define the conformal score

$$s^{(*)}(x_i,\theta_i) := \inf_{\tau\in\mathcal{T}}\{(x_i,\theta_i)\in F_\tau^{(t^*)}\}$$

Compute the corresponding conformal prediction set as

$$\mathcal{C}_{\mathcal{R}}^{(*)} := \{(x,\theta) : s^{(*)}(x,\theta) \le \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)\},$$

where $\widehat{q}_{\mathcal{R}}^{(*)}(\alpha) := \lceil(|\mathcal{D}_{\mathcal{R}}|+1)(1-\alpha)\rceil$-th largest element of $\{s^{(*)}(x_i,\theta_i)\}_{i\in\mathcal{D}_{\mathcal{R}}}$

Return the prediction set $\mathcal{C}_{\mathcal{R}}^{(*)}$

end procedure

Proof. Let $\mathcal{C}_{\mathcal{R}}^{(*)} = \text{VFCP}\left(\{q^{(t)}(\Theta \mid X)\}_{t=1}^{T}, 1-\alpha, D_{\mathcal{C}}, D_{\mathcal{T}}, 1/q(\theta \mid x)\right)$. We first wish to demonstrate $\mathcal{C}_{\mathcal{C},\mathcal{R}}^{(t),\text{VFCP}}(x) = \mathcal{C}_{\mathcal{C},\mathcal{R}}^{(t),\text{CANVI}}(x)$ for any fixed x, where we use the $\mathcal{C}, \mathcal{R}$ condensed notation to mean this equality holds both under $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$. This equivalence can be shown in demonstrating the equivalence in corresponding scores, as the resulting empirical score distributions and hence quantiles over $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{D}_{\mathcal{R}}$ and finally future prediction regions follow to then be equivalent by the algorithm structure.

This equivalence follows as a straightforward instance of the equivalence of the set-valued and standard conformal prediction frameworks. In particular, we recover the original score formulation, as:

$$s^{(t)}(x_i, \theta_i) := \inf_{\tau \in \mathcal{T}} \{(x_i, \theta_i) \in F_{\tau}^{(t)}\} = \inf_{\tau \in \mathcal{T}} \{(x_i, \theta_i) \in \{(x, \theta) \mid 1/q^{(t)}(\theta \mid x) \leq \tau\}\} = 1/q^{(t)}(\theta_i \mid x_i).$$

The bound under access to the exact $\ell(q, \tau)$ then follows from Theorem A.4.2 under demonstration of the appropriate assumptions. Assumption 2 holds by assumption. Assumption 1 holds for any $\vartheta^{(t)} := \{\theta : \nabla_{\theta} q^{(t)}(\theta|x)\}$ by Corollary 2.3.3. In the application of Assumption 1 for the proof of Theorem A.4.2, it suffices for $\mathcal{F}_t^{-1}(1-\alpha) \in \vartheta^{(t)}$ and $\mathcal{F}_t^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1)) \in \vartheta^{(t)}$. Since $\mathcal{L}(\{\theta : \nabla_{\theta} q^{(t)}(\theta|x)\}) = 0$ and $\mathcal{P}(X, \Theta)$ is bounded above, this means $\mathcal{P}_{X,\Theta}(\mathcal{F}_t^{-1}(1-\alpha) \in \vartheta^{(t)}) = 1$ and $\mathcal{P}_{X,\Theta}(\mathcal{F}_t^{-1}(1-\alpha+1/(N_{\mathcal{C}}+1)) \in \vartheta^{(t)}) = 1$. Thus, if t^* is chosen in CANVI using $\ell(q, \tau)$, by Theorem A.4.2, with probability $1(1-\delta) = 1-\delta$,

$$\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}} \right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}} \right)^{\gamma} \right].$$

The extension of this to the case of interest, where t^* is chosen using $\widehat{\ell}(q, \tau)$, is now a straightforward application of Assumption 3, from which we have that $\exists\, \Delta, \epsilon > 0$, such that with probability at least $1-\epsilon$

$$\left| \ell(q^{(\widehat{t^*})}, \widehat{q}_{\mathcal{R}}^{(\widehat{t^*})}(\alpha)) - \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)) \right| < \Delta \implies \ell(q^{(\widehat{t^*})}, \widehat{q}_{\mathcal{R}}^{(\widehat{t^*})}(\alpha)) < \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)) + \Delta. \quad \text{(A.4)}$$

Switching back to denoting $\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) := \ell(q^{(\widehat{t^*})}, \widehat{q}_{\mathcal{R}}^{(\widehat{t^*})}(\alpha))$, this implies that, with probability $(1-\epsilon)(1-\delta)$,

$$\begin{aligned}
&\ell(q^{(*)}, \widehat{q}_{\mathcal{R}}^{(*)}(\alpha)) \leq \ell(q^{(t^*)}, \widehat{q}_{\mathcal{R}}^{(t^*)}(\alpha)) + \Delta \\
&\leq \min_{1 \leq t \leq T} \ell(q^{(t)}, \widehat{q}_{\mathcal{R}}^{(t)}(\alpha)) + \Delta + 3L_W L_{[T]} \left[\left(\frac{\log(4T/\delta)}{N_{\mathcal{C}}} \right)^{\gamma/2} + \left(\frac{2}{N_{\mathcal{C}}} \right)^{\gamma} \right],
\end{aligned}$$

completing the proof as desired.

■

A.5 Gaussian Hölder Continuity

Theorem A.5.1. *Let Θ and X be zero-mean unit-variance Gaussian random variables with correlation ρ. Let $q^{(t)}(\theta|x) = \mathcal{N}(\theta; tx, 1-\rho^2)$. Let $\kappa := t^2 - 2t\rho + 1$ and $r > 0$. Then $F_t^{-1}(z)$, is 1-Hölder continuous on $[1-\alpha, 1-\alpha+r]$ with Hölder constant*

$$\frac{\kappa\Phi^{-1}(\frac{1-\alpha}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{1-\alpha}{2})^2 - \frac{(1-\alpha)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}} \tag{A.5}$$

Proof. Notice that in the bivariate Gaussian case, we have closed forms for the following:

$$\Theta \mid x \sim \mathcal{N}(\rho x, 1-\rho^2) \qquad X \mid \theta \sim \mathcal{N}(\rho\theta, 1-\rho^2)$$
$$\Theta \sim \mathcal{N}(0,1) \qquad X \sim \mathcal{N}(0,1).$$

We wish to find the distribution of $s(X,\Theta) = 1/q(\Theta \mid X)$ jointly over X, Θ to find $F_t^{-1}(z)$ explicitly. The CDF of this score can be computed as follows:

$$\mathcal{P}\left(1/q_t(\Theta \mid X) \le q\right) = \mathcal{P}\left(\sqrt{2\pi(1-\rho^2)}e^{\frac{(tX-\Theta)^2}{2(1-\rho^2)}} \le q\right)$$
$$= \mathcal{P}\left(R^2 \le 2(1-\rho^2)\log\left(\sqrt{\frac{q^2}{2\pi(1-\rho^2)}}\right)\right) = \mathcal{P}\left(R \le \sqrt{(1-\rho^2)\log\left(\frac{q^2}{2\pi(1-\rho^2)}\right)}\right),$$

where $R := |tX - \Theta|$. From the above calculation, $F_t^{-1}(z)$ must satisfy:

$$\mathcal{P}\left(R \le \sqrt{(1-\rho^2)\log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right)}\right) = z.$$

Notice now that, since (X, Θ) are bivariate Gaussian:

$$tX - \Theta \sim \mathcal{N}(0, t^2 + 1 - 2t\rho) \implies R \sim \text{HalfNormal}(t^2 + 1 - 2t\rho).$$

Therefore, the z quantile of R is $\sqrt{t^2+1-2t\rho}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$. Solving for $F_t^{-1}(z)$ in this quantile

produces the final threshold:

$$\sqrt{(1-\rho^2)\log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right)} = \sqrt{t^2+1-2t\rho}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$
$$\implies \log\left(\frac{F_t^{-1}(z)^2}{2\pi(1-\rho^2)}\right) = \frac{t^2+1-2t\rho}{1-\rho^2}\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2$$
$$\implies F_t^{-1}(z) = \sqrt{2\pi(1-\rho^2)\exp\left(\frac{t^2+1-2t\rho}{1-\rho^2}\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)^2\right)}.$$

The Hölder constant follows from bounding the derivative of $F_t^{-1}(z)$, namely

$$\frac{\kappa\Phi^{-1}(\frac{z}{2})\sqrt{\exp\left(\frac{\kappa}{1-\rho^2}\Phi^{-1}(\frac{z}{2})^2-\frac{(z)^2}{2}\right)}}{\sqrt{(1-\rho^2)/2}}$$

This expression is monotonically decreasing in z and hence maximized for $z = 1-\alpha$ in the interval, giving the desired expression.

■

A.6 Simulation-Based Inference Benchmarks

The benchmark tasks are a subset of those provided by [Lueckmann et al., 2021]. For convenience, we provide brief descriptions of the tasks curated by this library; however, a more comprehensive description of these tasks can be found in their manuscript.

A.6.1 Gaussian Linear

10-dimensional Gaussian model with a Gaussian prior:

$$\textbf{Prior}: \mathcal{N}(0, 0.1 \odot I)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

A.6.2 Gaussian Linear Uniform

10-dimensional Gaussian model with a uniform prior:

$$\textbf{Prior}: \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w \sim \mathcal{N}(x \mid w, 0.1 \odot I)$$

A.6.3 SLCP with Distractors

Simple Likelihood Complex Posterior (SLCP) with Distractors has uninformative dimensions in the observation over the standard SLCP task:

$$\textbf{Prior}: \mathcal{U}(-3, 3)$$
$$\textbf{Simulator}: x \mid w = p(y) \text{ where } p \text{ reorders}$$
$$y \text{ with a fixed random order}$$
$$y_{[1:8]} \sim \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} w_3^4 & w_3^2 w_4^2 \tanh(w_5) \\ w_3^2 w_4^2 \tanh(w_5) & w_4^4 \end{bmatrix}\right),$$
$$y_{9:100} \sim \frac{1}{20} \sum_{i=1}^{20} t_2(\mu^i, \Sigma^i), \mu^i \sim \mathcal{N}(0, 15^2 I),$$
$$\Sigma^i_{j,k} \sim \mathcal{N}(0, 9), \Sigma^i_{j,j} = 3e^a, a \sim \mathcal{N}(0, 1),$$

A.6.4 Bernoulli GLM Raw

10-parameter GLM with Bernoulli observations and Gaussian prior. Observations are not sufficient statistics, unlike the standard "Bernoulli GLM" task:

$$\textbf{Prior}: \beta \sim \mathcal{N}(0, 2), f \sim \mathcal{N}(0, (F^T F)^{-1})$$
$$F_{i,i-2} = 1, F_{i,i-1} = -2$$
$$F_{i,i} = 1 + \sqrt{\frac{i-1}{9}}, F_{i,j} = 0; i \leq j$$
$$\textbf{Simulator}: x^{(i)} \mid w \sim \text{Bern}(\eta(v_T^{(i)} f + \beta)),$$
$$\eta(\odot) = \exp(\odot)/(1 + \exp(\odot))$$

A.6.5 Gaussian Mixture

A mixture of two Gaussians, with one having a much broader covariance structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-10, 10)$$
$$\textbf{Simulator}: x \mid w \sim 0.5\mathcal{N}(x \mid w, I) + 0.5\mathcal{N}(x \mid w, .01I)$$

A.6.6 Two Moons

Task with a posterior that has both global (bimodal) and local (crescent-shaped) structure:

$$\textbf{Prior}: \beta \sim \mathcal{U}(-1, 1)$$
$$\textbf{Simulator}: x \mid w =$$
$$\begin{bmatrix} r\cos(\alpha) + 0.25 \\ r\sin(\alpha) \end{bmatrix} + \begin{bmatrix} -|w_1 + w_2|/\sqrt{2} \\ (-w_1 + w_2)/\sqrt{2} \end{bmatrix}$$
$$\alpha \sim \mathcal{U}(-\pi/2, \pi/2), r \sim \mathcal{N}(0.1, 0.01^2)$$

A.6.7 SIR

Epidemiology model with S (susceptible), I (infected), and R (recovered). A contact rate β and mean recovery rate of γ are used as follows:

$$\textbf{Prior}: \beta \sim \text{LogNormal}(\log(0.4), 0.5),$$
$$\gamma \sim \text{LogNormal}(\log(1/8), 0.2)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10}; x^{(i)} \mid w \sim \text{Bin}(1000, \frac{I}{N}),$$
where I is simulated from:
$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \qquad \frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I, \qquad \frac{dR}{dt} = \gamma I$$

A.6.8 Lotka-Volterra

An ecological model commonly used in describing dynamics of competing species. w parameterizes this interaction as $w = (\alpha, \beta, \gamma, \delta)$:

$$\textbf{Prior}: \alpha \sim \text{LogNormal}(-.125, 0.5)$$
$$\beta \sim \text{LogNormal}(-3, 0.5), \gamma \sim \text{LogNormal}(-.125, 0.5)$$
$$\delta \sim \text{LogNormal}(-3, 0.5)$$
$$\textbf{Simulator}: x = (x^{(i)})_{i=1}^{10},$$
$$x_{1,i} \mid w \sim \text{LogNormal}(\log(X), 0.1),$$
$$x_{2,i} \mid w \sim \text{LogNormal}(\log(Y), 0.1)$$
where X, Y is simulated from:
$$\frac{dX}{dt} = \alpha X - \beta XY, \qquad \frac{dY}{dt} = -\gamma Y + \delta XY$$

A.6.9 ARCH

The two-dimensional parameter $\theta = (\theta_1, \theta_2)$ includes both an autoregressive component (θ_1) and a component controlling the level of conditional noise (θ_2). Given a full realization of the time series $y_{1:T}$, we aim to amortize inference over θ. Priors are taken to be $\theta_1 \sim \text{Unif}(-1, 1)$ and $\theta_2 \sim \text{Unif}(0, 1)$. One important change from the model of [Thomas et al., 2022] is that we fix $e^{(0)} = 0$ rather than drawing this quantity from a standard Gaussian.

The Adam optimizer was used with a learning rate of 0.0001 for 25,000 training steps for each of the three methods: IWBO ($K = 10$), ELBO, FAVI. Due to constraints arising from the uniform priors on the parameters, the IWBO and ELBO implementations rely on logit transformations of the latent random variables θ to avoid zero-density regions that result in undefined gradients. The encoder network is trained to learn distributions on the unconstrained space of the transformed random variable θ', and visualizations are produced by performing the inverse transformation. While FAVI avoids these issues because it is likelihood-free, for an apples-to-apples comparison we also implement FAVI on the unconstrained latent space as well.

A.7 Training Details

All encoders were implemented in PyTorch [Paszke et al., 2019] with a Neural Spline Flow architecture. The NSF was built using code from [Durkan et al., 2020a]. Specific architecture hyperparameter choices were taken to be the defaults from [Durkan et al., 2020a] and are available in the code. Optimization was done using Adam [Kingma and Ba, 2014] with a learning rate of 10^{-3} over 5,000 training steps. Minibatches were drawn from the corresponding prior $\mathcal{P}(\Theta)$ and simulator $\mathcal{P}(X \mid \Theta)$ as specified per task in the preceding section. Training these models required

between 10 minutes and two hours using an Nvidia RTX 2080 Ti GPUs for each of the SBI tasks.

A.8 Prediction Regions

Credible regions for $q_\varphi(\Theta \mid x)$ (for a single $x \sim \mathcal{P}(X)$) on a subset of the SBI benchmark tasks are plotted below over varying degrees of training, as indicated in the figures. As expected, the efficiency of the prediction regions improves over training across all tasks, resulting in *smaller* regions for each target coverage $1 - \alpha$.

A.8.1 Gaussian Linear

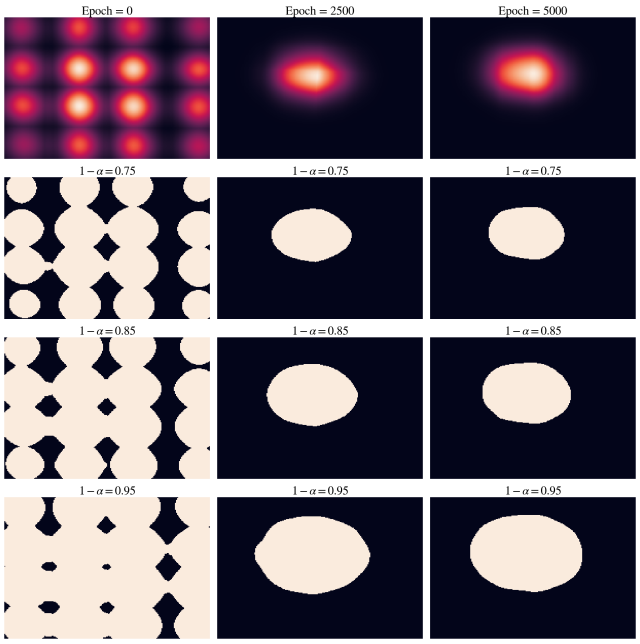

Figure A.3: Gaussian Linear conformal prediction regions

A.8.2 Gaussian Linear Uniform

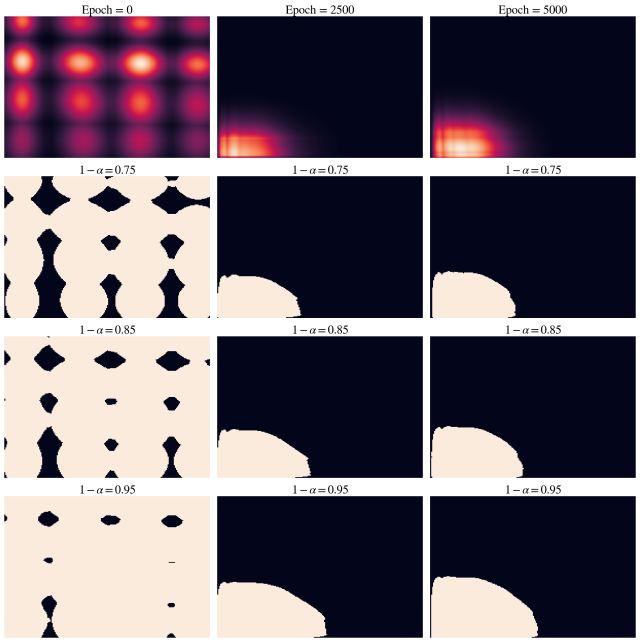

Figure A.4: Gaussian Linear Uniform conformal prediction regions

A.8.3 SLCP

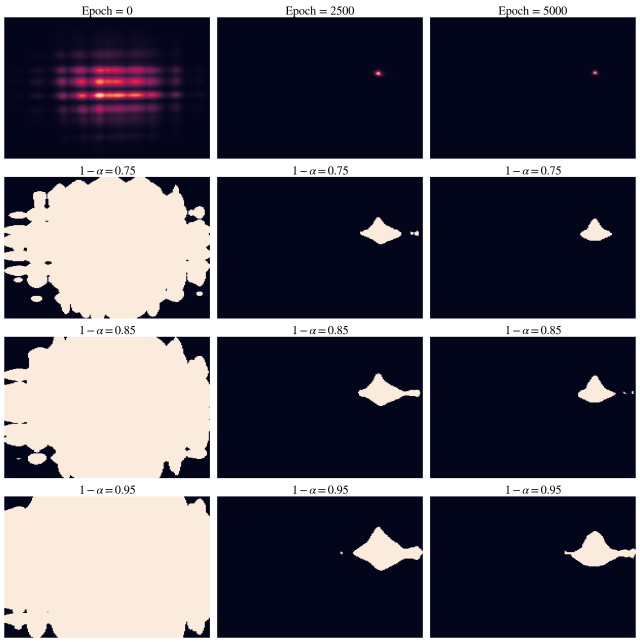

Figure A.5: SLCP conformal prediction regions

A.8.4 Bernoulli GLM Raw

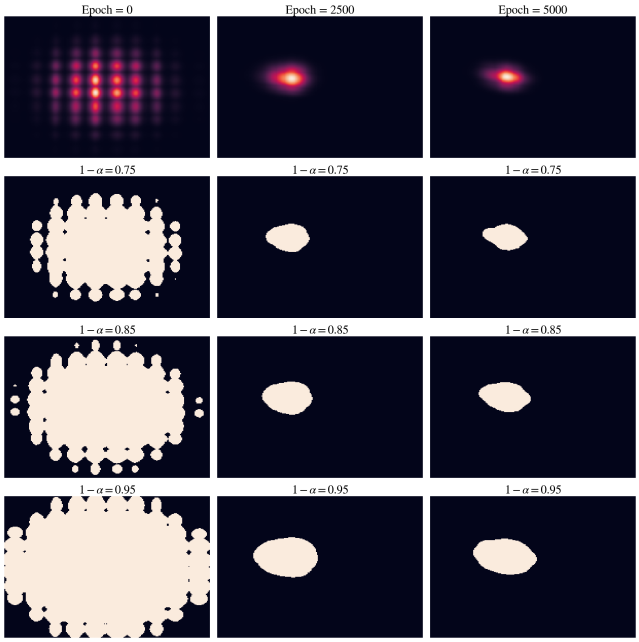

Figure A.6: Bernoulli GLM Raw conformal prediction regions

A.8.5 Gaussian Mixture

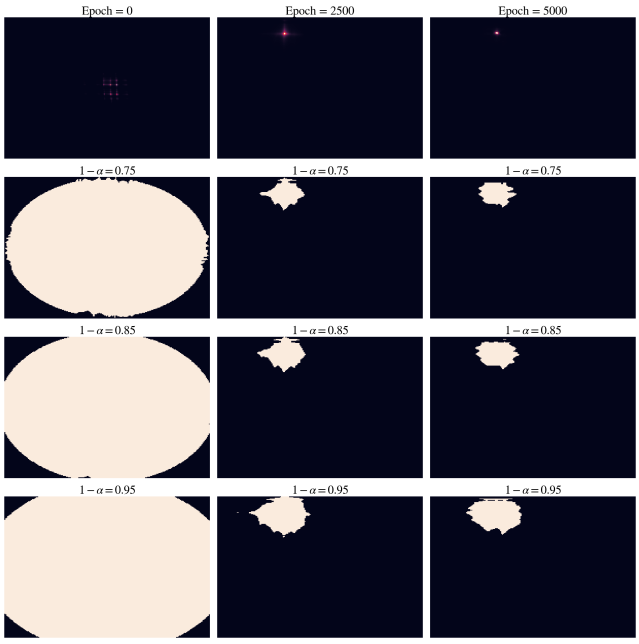

Figure A.7: Gaussian Mixture conformal prediction regions

A.8.6 Two Moons

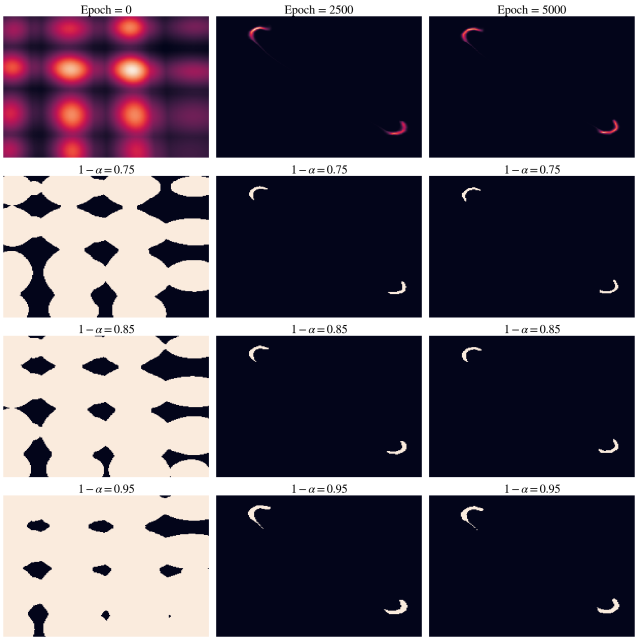

Figure A.8: Two Moons conformal prediction regions

A.9 Posteriors

We provide visualizations of approximate and reference posteriors (produced with MCMC from [Lueckmann et al., 2021]) to justify the overdispersion claims made on the variational approximation procedure.

A.9.1 Gaussian Linear

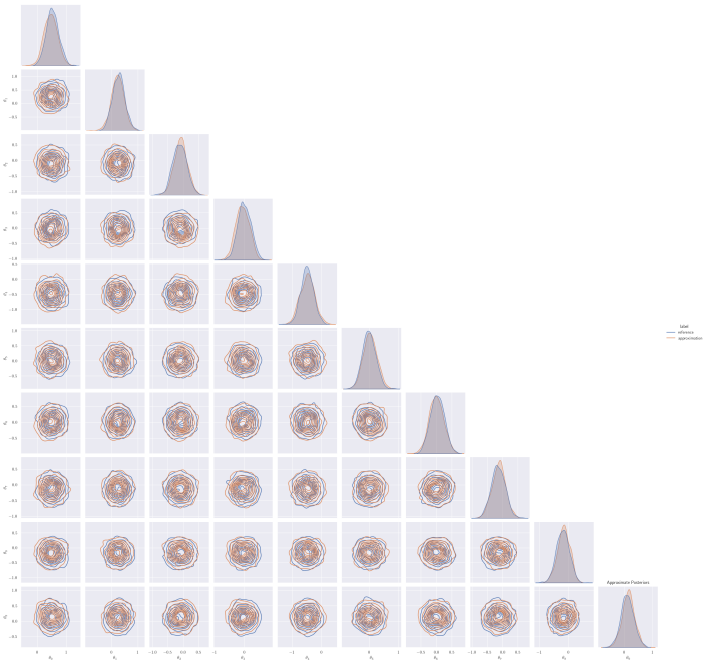

Figure A.9: Gaussian Linear true vs. approximate posterior distributions

A.9.2 Gaussian Mixture

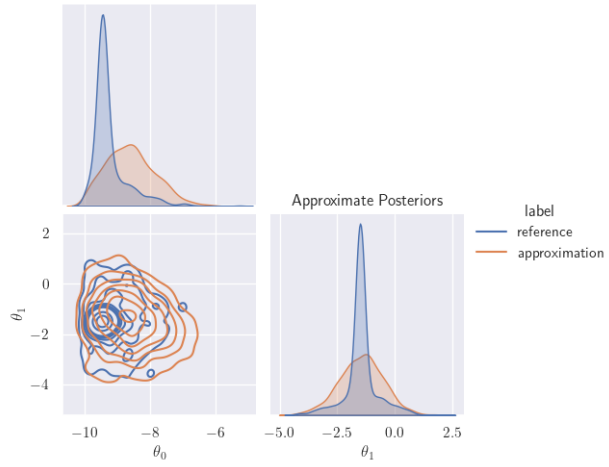

Figure A.10: Gaussian Mixture true vs. approximate posterior distributions

A.9.3 Gaussian Linear Uniform

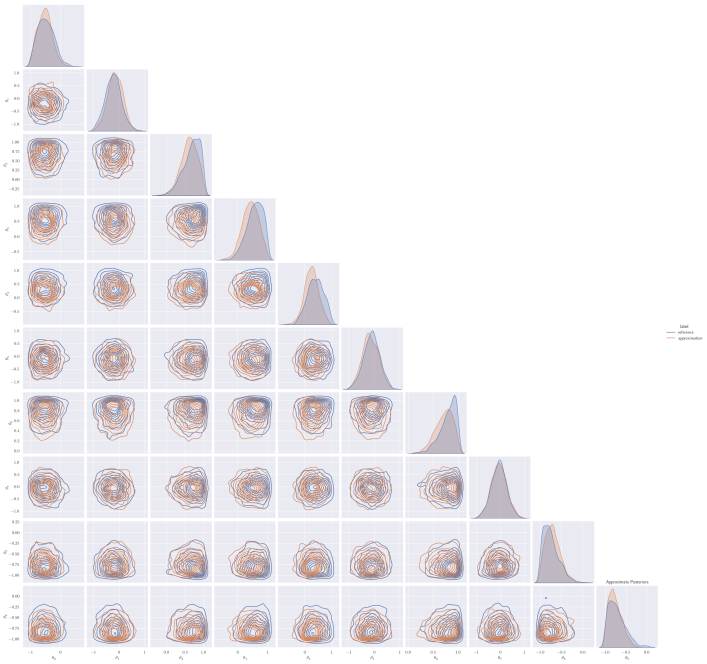

Figure A.11: Gaussian Linear Uniform true vs. approximate posterior distributions

A.9.4 Two Moons

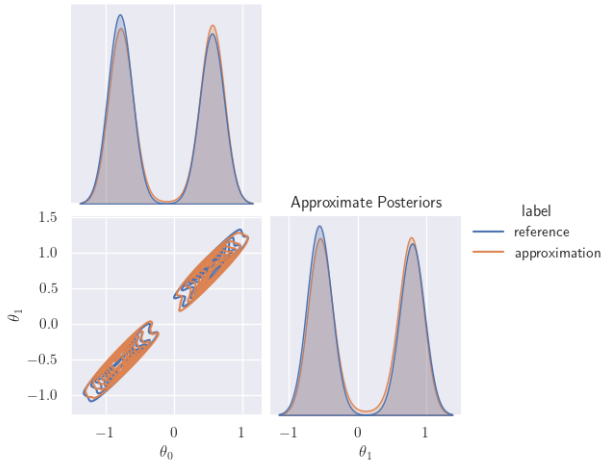

Figure A.12: Two Moons true vs. approximate posterior distributions

A.9.5 SLCP

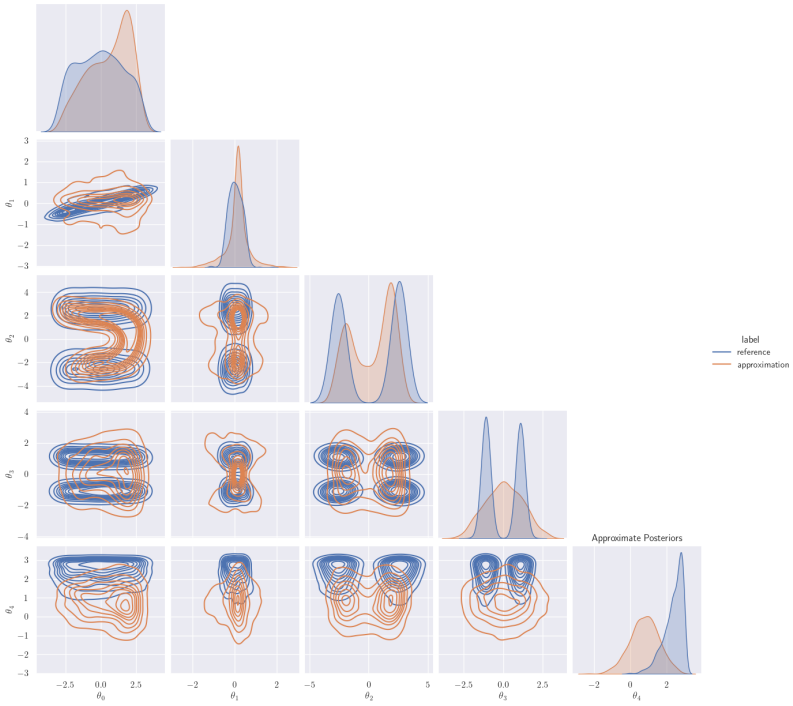

Figure A.13: SLCP true vs. approximate posterior distributions

A.9.6 Bernoulli GLM

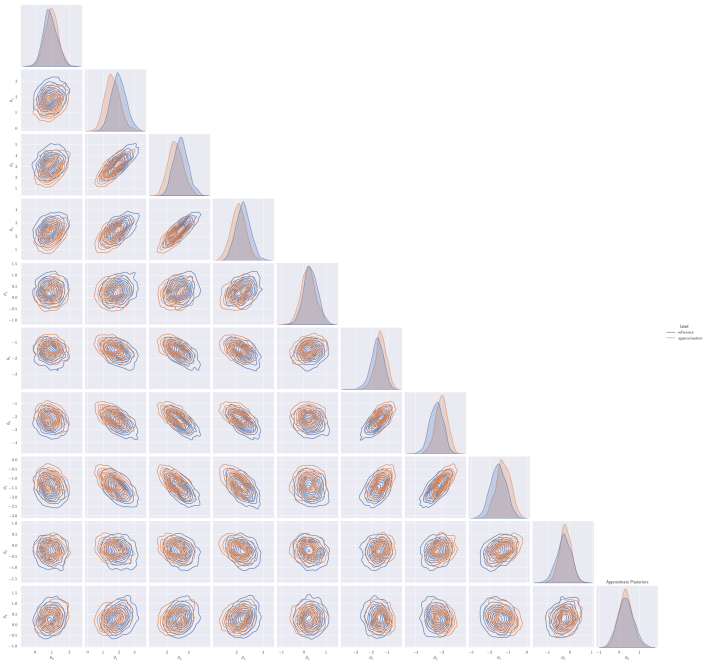

Figure A.14: Bernoulli GLM true vs. approximate posterior distributions

A.9.7 ARCH

For selected points from the training sets, we show the exact and approximate posteriors obtained by training according to the ELBO, IWBO, and FAVI objectives. The ground-truth parameter value is noted in red.

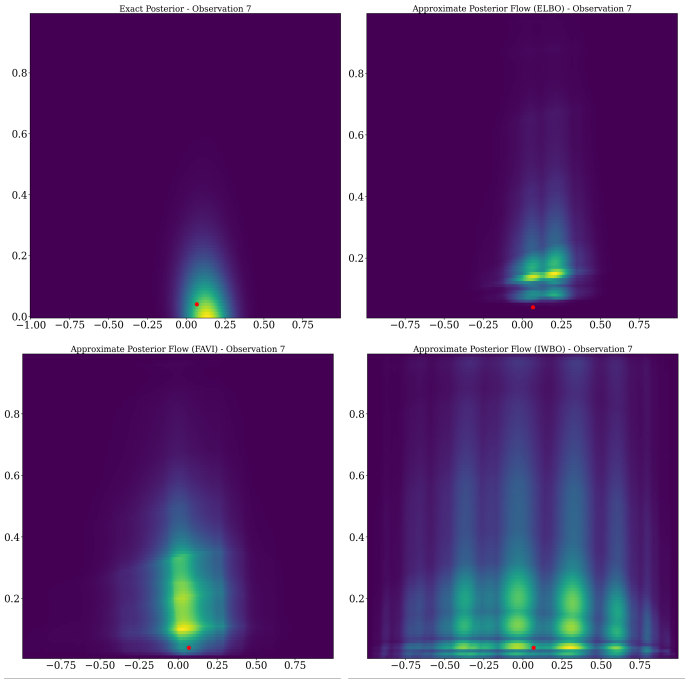

Figure A.15: ARCH true vs. approximate posterior distributions under ELBO (top-right), FAVI (bottom-left), and IWBO (bottom-right) training objectives.

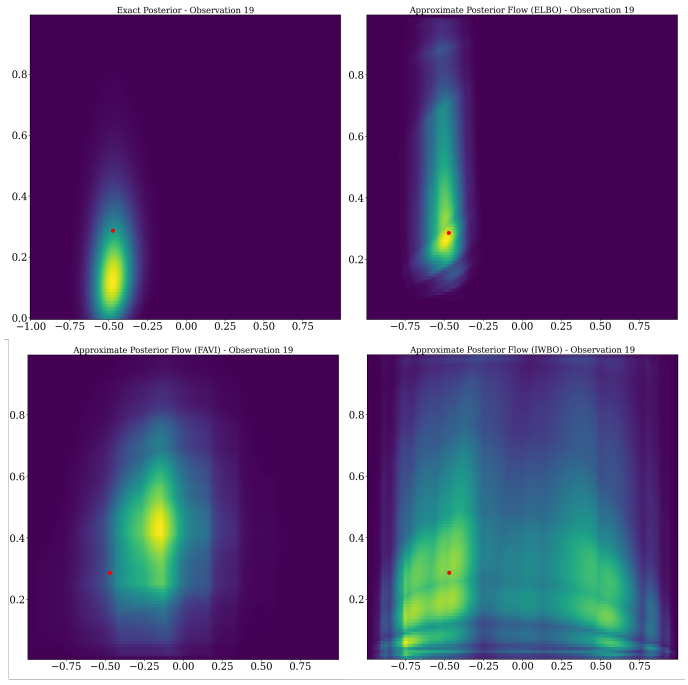

Figure A.16: ARCH true vs. approximate posterior distributions under ELBO (top-right), FAVI (bottom-left), and IWBO (bottom-right) training objectives.

Depending on the miscalibration of the variational posterior on either per-objective or per-point basis, the conformalized $1 - \alpha$ high-density region (HDR) either shrinks or expands, and can be visualized in two dimensions. Figure A.17, Figure A.18 show examples where the 50% prediction regions obtained from the FAVI-learned variational posterior shrinks and grows after applying CANVI.

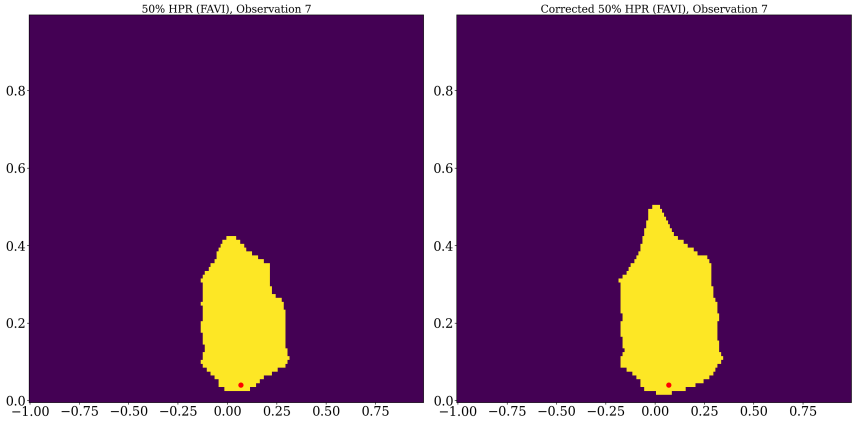

Figure A.17: 50% prediction region, observation 7, before (left) and after (right) applying CANVI.

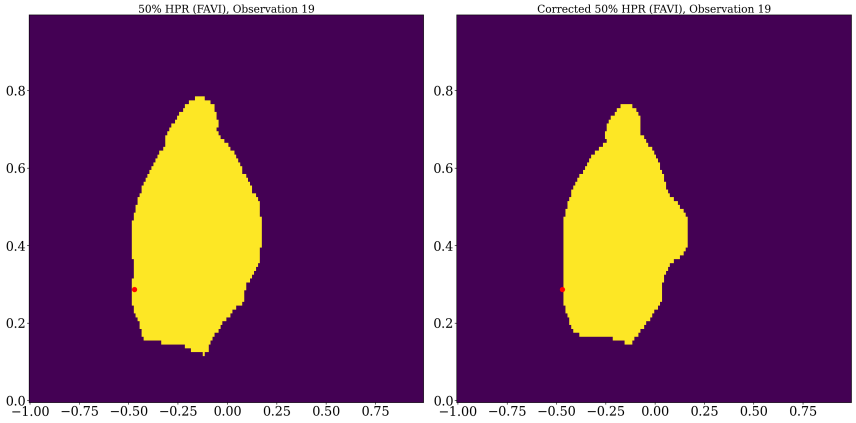

Figure A.18: 50% prediction region, observation 8, before (left) and after (right) applying CANVI.

We visualize the efficiency of the iterates q_φ across training iterations in Figure A.19. Recalling the form of Equation (2.5), it is unsurprising that FAVI tends to be more efficient, as it trains using simulated data (over which Equation (2.5) is computed). To estimate Equation (2.5), at every 500 training steps, we simulate 20 θ, x pairs from the forward model. A larger number of Monte Carlo samples can be used but at an increased cost. As the resulting estimates are noisy, we smooth the resulting series with a Savitzky-Golay filter with a window length of 10 and third-degree polynomial order for better visualization.

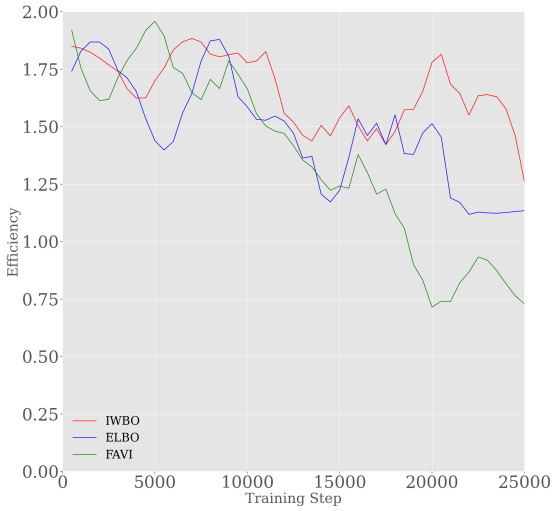

Figure A.19: Efficiency estimates for variational posteriors trained by ELBO, IWBO, and FAVI across training iterations.

A.10 SED Experimental Details

The PROVABGS emulator (Section 2.4.3) was trained to minimize the MSE using normalized simulated PROVABGS outputs with fixed log stellar mass parameter [Hahn et al., 2023]. Training data were generated from PROVABGS with a fixed magnitude parameter ($\theta_0 = 10.5$), resampled onto a 5 Angstrom grid, and normalized to integrate to one. After training, forward passes through the emulator are significantly faster than the base simulator. We provide two simulated draws from our emulator in Figure A.20. We use the recommended priors from [Hahn et al., 2023] on the remaining eleven parameters. As these are highly constrained (uniform priors, vastly different scales, and a 4-dimensional vector on the simplex), we similarly operate on an unconstrained, 10-dimensional space by invertible transformations. All three methods were trained for 10,000 steps using the Adam optimizer with learning rate 0.0001. $K = 1000$ was used for the IWBO.

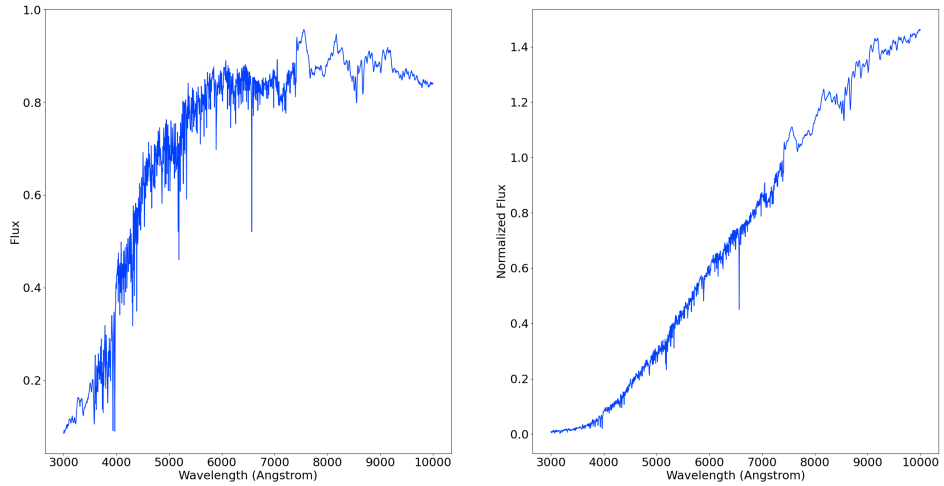

Figure A.20: Two example draws from our neural network emulator of PROVABGS.

We let the output of our simulator be a mean parameter $\mu \in \mathbb{R}^{1400}$, and generate observed data independently binwise via $x_i \sim \mathcal{N}(\mu_i, |\mu_i|^2\sigma^2)$ for a fixed hyperparameter $\sigma = .1$ and all $i = 1, \ldots, 1400$. We adopt this noise model for simplicity, as it imposes a fixed signal-to-noise ratio (SNR) of $1/\sigma = 10$ across all spectra and wavelength bins.

APPENDIX B

Conformal Contextual Robust Optimization

B.1 Prediction Region Validity Lemma

Lemma B.1.1. *Consider any $f(w, c)$ that is L-Lipschitz in c under the metric d for any fixed w. Assume further that $\mathcal{P}_{X,C}(C \in \mathcal{U}(X)) \geq 1 - \alpha$. Then,*

$$\mathcal{P}_{X,C}\left(\Delta(X, C) \leq L \operatorname{diam}(\mathcal{U}(X))\right) \geq 1 - \alpha. \tag{B.1}$$

Proof. We consider the event of interest conditionally on a pair (x, c) where $c \in \mathcal{U}(x)$:

$$\begin{aligned}
&\min_w \max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - \min_w f(w, c) \\
&\leq \max_w |\max_{\widehat{c} \in \mathcal{U}(x)} f(w, \widehat{c}) - f(w, c)| \\
&\leq L \max_{\widehat{c} \in \mathcal{U}(x)} d(\widehat{c}, c) \leq L \operatorname{diam}(\mathcal{U}(x)).
\end{aligned}$$

Since we have the assumption that $\mathcal{P}(C \in \mathcal{U}(X)) \geq 1 - \alpha$, the result immediately follows. ■

B.2 Optimization Convergence Lemma

We first begin by citing a standard result of projected gradient descent, from which the result of interest immediately follows.

Lemma B.2.1. *Let K be a closed convex set, and $f : K \to \mathbb{R}$ be convex, differentiable, and L-Lipschitz. Let $x^* \in K$ be a minimizer of f, and define $T := \frac{L^2 ||x_0 - x^*||}{\epsilon^2}$ and $\eta := \frac{||x_0 - x^*||}{L\sqrt{T}}$. Then the iterates $\{x_t\}_{t=0}^T$ returned by projected gradient descent satisfy*

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} x_t\right) - f(x^*) \leq \epsilon. \tag{B.2}$$

Lemma B.2.2. *Let* $\phi(w) := \max_{\widehat{c} \in \bigcup_{k=1}^{K} \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$ *for* $\{\widehat{c}_k\}_{k=1}^{K} \subset \mathcal{C}$, $\widehat{q} \in \mathbb{R}^+$, *and* $f(w, c)$ *convex-concave and* L*-Lipschitz in* c *for any fixed* w. *Let* $w^* \in \mathcal{W}$ *be a minimizer of* ϕ. *For any* $\epsilon > 0$, *define* $T := \frac{L^2 ||w_0 - w^*||}{\epsilon^2}$ *and* $\eta := \frac{||w_0 - w^*||}{L\sqrt{T}}$. *Then the iterates* $\{w_t\}_{t=0}^{T}$ *returned by Algorithm 3 satisfy*

$$\phi\left(\frac{1}{T+1}\sum_{t=0}^{T} w_t\right) - \phi(w^*) \leq \epsilon. \tag{B.3}$$

Proof. Notice that $\phi(w)$ is convex by Danskin's Theorem by assumption of the convexity of f in w. By Danskin's Theorem, $\nabla_w \phi(w) = \nabla_w f(w, c^*)$, where $c^* := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. Further notice

$$\phi(w) := \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c}) = \max_{k} \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c}). \tag{B.4}$$

Denote $\phi_k(w) := \max_{\widehat{c} \in \mathcal{B}_{\widehat{q}}(\widehat{c}_k)} f(w, \widehat{c})$. Clearly, $\phi_k(w)$ is L-Lipschitz by assumption on the structure of f. Further, as the point-wise maximum of L-Lipschitz functions is itself L-Lipschitz, it follows that $\phi(w) = \max_k \phi_k(w)$ is also L-Lipschitz. The conclusion, thus, follows by applying Theorem B.2.1 to $\phi(w)$. ■

B.3 SBI Representative Points

B.3.1 Gaussian Mixture

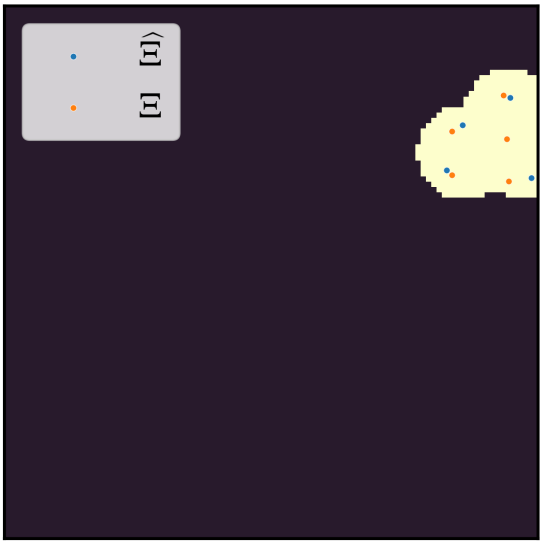

Figure B.1: Recovery of the true representative points under approximation algorithm given in Algorithm 5 for a Gaussian Mixture posterior.

B.3.2 Two Moons

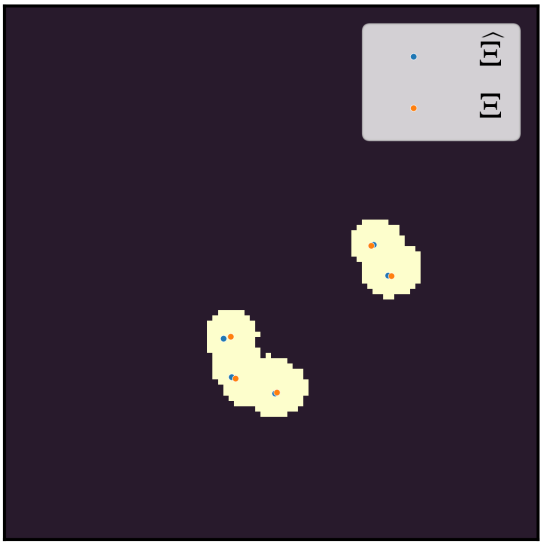

Figure B.2: Recovery of the true representative points under approximation algorithm given in Algorithm 5 for a Two Moons posterior.

B.4 Robust Vehicle Routing Setup

The routing graph of Manhattan was extracted using OSMnx, with local highway speeds extracted using OpenStreetMap [Boeing, 2017]. Highway speed imputation was performed on edges where such information was not available, specifically by averaging over those highways of comparable categorization, namely "residential," "secondary," or "tertiary." Doing so defined a nominal travel cost $\widetilde{c}$.

We now wish to modify these nominal travel costs to account for the weather predictions made upstream. That is, we wish to account for the precipitation map $\widetilde{Y} \in \mathbb{R}^{W \times H}$ in these edge weights. To do so, we use the global coordinates $(c_x^v, c_y^v) \in \mathbb{R}^2$ of each $v \in \mathcal{V}$ to find the precipitation at the corresponding location. Concretely, we determine the pixel coordinate by scaling the coordinate to the range of the region that was forecasted. So, for a forecast over the window $(c_x^{\min}, c_x^{\max}) \times (c_y^{\min}, c_y^{\max})$, the corresponding pixel lookup is:

$$p_x^v = \lfloor \frac{c_x^v - c_x^{\min}}{c_x^{\max} - c_x^{\min}} \rfloor \times W \qquad p_y^v = \lfloor \frac{c_y^v - c_y^{\min}}{c_y^{\max} - c_y^{\min}} \rfloor \times H.$$

The corresponding precipitation associated with each vertex, therefore, is $\widetilde{Y}_{p_x^v, p_y^v}$. We define the final travel cost for each edge $e \in \mathcal{E}$ with endpoints (e_s, e_t) as:

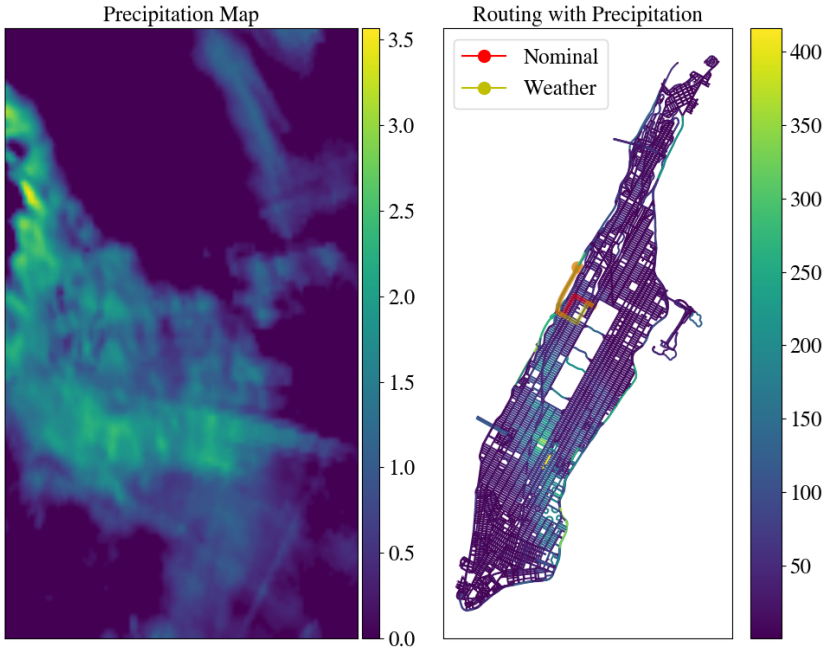

Figure B.3: Precipitation maps (left) are converted to edge weights (right) as per Equation (B.5). Solving the shortest paths problem (SPP) on this newly weighted graph, therefore, can produce distinct routes from that based on the nominal travel-time SPP, as highlighted by the two distinct paths under the nominal and weather-weighted graphs on the right.

$$c_e := \widetilde{c}_e \cdot \exp\left\{ \frac{\widetilde{Y}_{p_x^{e_s}, p_y^v} + \widetilde{Y}_{p_x^{e_t}, p_y^{e_t}}}{2} \right\}. \tag{B.5}$$

We then solve SPP on the weighted directed graph with edge weights c_e. An example of this weighting and the corresponding shortest path is illustrated in Figure B.3.

APPENDIX C

Applications of Conformal Decision Making

C.1 Conformal Decision Making for Ensembles

C.2 CSA Predict-Then-Optimize Visual Walkthrough

We now present a visual accompaniment of the predict-then-optimize algorithm presented in Section 4.1.3. We once again take $K = 2$ for visual clarity in this walkthrough, where the predictors are as discussed in Section 4.1.3, namely assumed to be generative predictors $q_k(C \mid X)$ where the number of samples per predictor are fixed to be $\{J_k\}$. For illustration, we assume $J_1 = 5$ and $J_2 = 3$, meaning predictions with the first model are made by drawing 5 samples and 3 for the second. We assume the CSA calibration of Section 4.1.2 has already been performed, from which a collection of projection directions and quantiles $\{(u_m, \widehat{q}_m)\}_{m=1}^{M}$ are available that implicitly define an acceptance region $\widehat{\mathcal{Q}}$. We further assume the individual predictor score functions are all the GPCP score given in Equation (3.3), with d_k from Equation (3.3) specifically here taken to simply be the standard Euclidean 2-norm, giving

$$s_k(x, c) = \min_{j \in 1,\ldots,J_k} ||\widehat{c}_{kj} - c||. \tag{C.1}$$

We now wish to compute $c^* = \max_{\widehat{c} \in \mathcal{C}(x)} f(w, \widehat{c})$. To do so, we must start by defining this region $\mathcal{C}(x)$ for the test point x, which we do by drawing the respective number of samples from the two models, producing samples $\{\widehat{c}_{1j}\}_{j=1}^{5}$ and $\{\widehat{c}_{2j}\}_{j=1}^{3}$, as shown in Figure C.1.

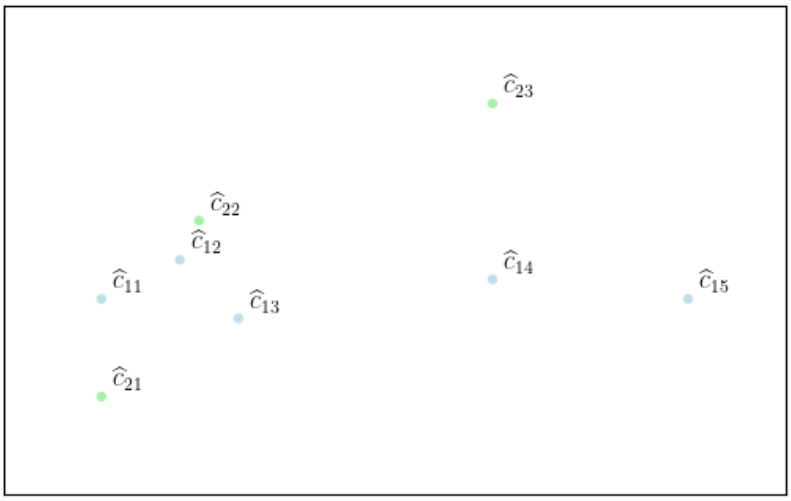

Figure C.1: Samples drawn from the two generative models $\{\widehat{c}_{1j}\}_{j=1}^{5} \sim q_1(C \mid x)$ (blue) and $\{\widehat{c}_{2j}\}_{j=1}^{3} \sim q_2(C \mid x)$ (green). Note that this is a visualization in the $\mathcal{C}$ space, i.e. *not* the space of multivariate scores.

By definition, $\forall c \in \mathcal{C}(x)$,

$$u_m^\top \left(\min_{j_1=1,...,5} ||\widehat{c}_{1j_1} - c||, \min_{j_2=1,...,3} ||\widehat{c}_{2j_2} - c|| \right) \leq \widehat{q}_m \qquad \forall m = 1, ..., M. \tag{C.2}$$

As a result, we must have that, $\forall c \in \mathcal{C}(x)$, $\exists\ j_1 = 1, ..., 5$ and $j_2 = 1, ..., 3$ such that $u_m^\top \left(||\widehat{c}_{1j_1} - c||, ||\widehat{c}_{2j_2} - c|| \right) \leq \widehat{q}_m\ \forall m = 1, ..., M$. Solving for c^*, therefore, amounts to considering each pair $\vec{j} := (j_1, j_2) \in \mathcal{J}$, where $\mathcal{J} := \{1, ..., 5\} \times \{1, ..., 3\}$, and solving

$$\begin{aligned} c^*_{\vec{j}} &:= \arg\max_c f(w, c) \\ \text{s.t.} \quad & u_m^\top \left(||\widehat{c}_{1\vec{j}_1} - c||, ||\widehat{c}_{2\vec{j}_2} - c|| \right) \leq \widehat{q}_m \quad \forall m \in \{1, ..., M\} \end{aligned} \tag{C.3}$$

Notice that, for any fixed $\vec{j}$, this is a standard convex optimization problem with a convex feasible region. We illustrate how the feasible region would be constructed for a fixed $\vec{j}$ in Figure C.2. Note that the construction of this feasible is never explicitly done in practice and is only implicitly used by convex solver routines in practice. We can, therefore, then solve Equation (C.3) over all possible $\vec{j} \in \mathcal{J}$ and aggregate the maxima to compute c^*.

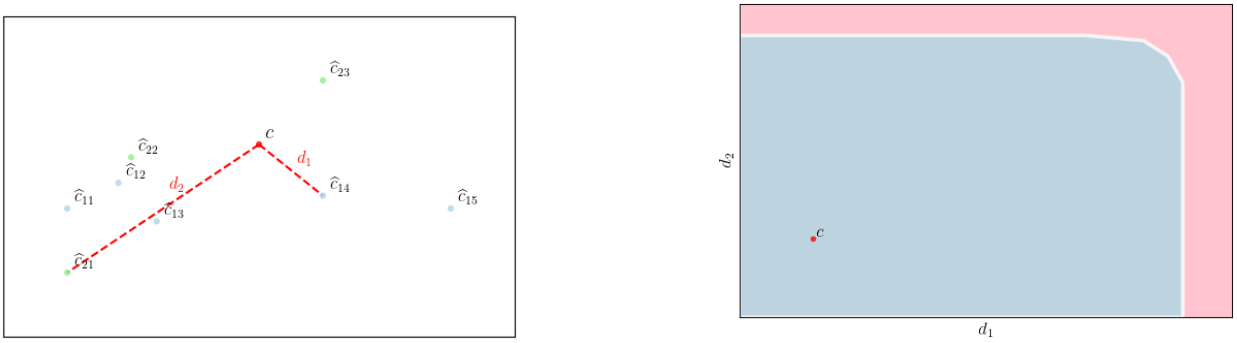

Figure C.2: A candidate $c \in \mathcal{C}$ is in prediction region if the projections of its distances (d_1, d_2) from *at least one* pair of points indexed by $\vec{j} := (j_1, j_2)$ is in the acceptance region $\widehat{\mathcal{Q}}$. Here, we illustrate a point c that lies in the feasible region from its proximity to the $\vec{j} := (4, 1)$ pair of points. Note that the left is again a visualization over the $\mathcal{C}$ space, whereas the right is of the *score* space.

C.2.1 CSA Predict-Then-Optimize Algorithm

We provide the condensed presentation of the predict-then-optimize algorithm below.

Algorithm 8 PREDICT-THEN-OPTIMIZE UNDER CSA

procedure PREDICT-THEN-OPTIMIZE UNDER CSA
Inputs: Context x, Predictors $\{q_k(C \mid X)\}_{k=1}^K$, Optimization steps T, Sample counts $\{J_k\}_{k=1}^K$, CSA quantile $\{(u_m, \widehat{q}_m)\}_{m=1}^M$
  $\{\{\widehat{c}_{kj}\}_{j=1}^{J_k} \sim q_k(C \mid X)\}_{k=1}^K, \mathcal{J} = \prod_{k=1}^K [J_k]$
  $w^{(0)} \sim U(\mathcal{W})$
  for $t \in \{1, \dots T\}$ **do**
    for $\vec{j} \in \mathcal{J}$ **do** $c^*_{\vec{j}} \leftarrow \arg\max_c f(w^{(t)}, c) \quad \text{s.t.} \quad \forall m \in 1, ..., M \quad u_m^\top s(\widehat{c}_{\vec{j}}, c) \leq \widehat{q}_m$
    $c^* \leftarrow \arg\max_{c^*_{\vec{j}}} f(w^{(t)}, c^*_{\vec{j}})$
    $w^{(t)} \leftarrow \Pi_{\mathcal{W}}(w^{(t-1)} - \eta \nabla_w f(w^{(t-1)}, c^*))$
  end for
  Return $w^{(T)}$
end procedure

C.3 Computational Efficiency

C.3.1 Vectorized Score Computation

We now discuss the vectorized form of the computations discussed in Section 4.1.2. In particular, putting the scores into a matrix $S_C^{(1)} = [s_1, ..., s_{N_{C_1}}]^\top \in \mathbb{R}^{N_{C_1} \times K}$ and directions into a matrix

$U = [u_1, ..., u_M]^\top \in \mathbb{R}^{K \times M}$, all the projections $u_m^\top s_i$ can be computed as $S_C^{(1)} U \in \mathbb{R}^{N_{C_1} \times M}$, where $[S_C^{(1)} U]_{i,m}$ is precisely $u_m^\top s_i$. $\tilde{q} \in \mathbb{R}^M := \{\tilde{q}_m := \text{quantile}([S_C^{(1)} U]_{:,m}; 1-\alpha)\}$ is then the quantile per row.

For any test point $s' \in \mathbb{R}^K$, we can then very efficiently check if it falls into the region by checking if it satisfies $Us' \leq \tilde{q}$ component-wise. Each iteration of the loop to find β^*, therefore, is very fast, and we find the search typically converges in 5-10 iterations.

The final step is computing $\widehat{t}$. Computing this follows similarly to above, where we take the scores $S_C^{(2)}$, compute projections $S_C^{(2)} U \in \mathbb{R}^{N_{C_2} \times M}$, find $\widetilde{T} := S_C^{(2)} U / \tilde{q} \in \mathbb{R}^{N_{C_2} \times M}$, where division is interpreted as being defined component-wise along the rows, computing the maxima similarly along the rows $[T^*]_i := \max \widetilde{T}_{i,:}$, and finally computing $\widehat{t}$ as the $1-\alpha$ quantile of T^*.

C.3.2 Empirical Efficiency Validation

To demonstrate the computational efficiency of this vectorized approach, we reran the experiment on the "Parkinsons" UCI task with both varying numbers of predictors (K) and projection directions (M); for each combination, we measured the total time taken to compute the quantile (i.e. to run Algorithm 1) and to perform the projection to assess coverage for the test points. The additional predictors were taken to be random forests with different numbers of trees. K is given in the left column and M in each column heading, with the entry for each (K, M) pair being reported in seconds. As expected, by the vectorized nature of the computations, as discussed in the main paper, the performance scales gracefully over M at roughly $\mathcal{O}(M)$ and remains roughly constant in K.

Table C.1: Performance values for varying K (number of predictors) and M (number of projection directions). All values are reported in seconds.

K \ M	**10**	**100**	**1000**	**10000**
6	0.111668	0.373029	2.32803	37.2561
8	0.0961056	0.327051	2.16211	36.7216
10	0.117146	0.373464	2.73875	37.2603
12	0.123772	0.384527	2.35386	37.0735

C.4 Full OpenML Results

Table C.2: Average coverages across tasks for $\alpha = 0.05$ are shown in the top row and average prediction set lengths in the bottom row, where both were assessed over a batch of i.i.d. test samples (20% of the dataset size). Standard deviations and means were computed across 5 randomizations of draws of the training, calibration, and test sets. In cases where the method failed to achieve sufficient coverage (defined as < 0.93), we do not include it in comparison for set length. Similarly, the single-stage approach fails to achieve coverage due to lack of exchangeability with test points.

Dataset	Metrics	Linear Model	LASSO	Random Forest	XGBoost	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	*CSA (Single-Stage)*	CSA
361234	Coverage	0.97 (0.011)	0.966 (0.011)	0.939 (0.002)	0.954 (0.006)	0.956 (0.011)	0.948 (0.01)	0.96 (0.013)	0.95 (0.006)	*0.955 (0.013)*	0.957 (0.01)
	Length	9.673 (0.160)	9.645 (0.154)	10.080 (0.160)	9.157 (0.052)	9.196 (0.123)	8.703 (0.086)	9.524 (0.056)	17.759 (0.275)	*7.646 (0.073)*	**7.688 (0.181)**
361235	Coverage	0.947 (0.0)	0.945 (0.005)	0.968 (0.016)	0.95 (0.005)	0.955 (0.016)	0.897 (0.005)	0.953 (0.011)	0.932 (0.021)	*0.745 (0.011)*	0.984 (0.005)
	Length	20.961 (0.651)	24.241 (0.246)	**10.096 (0.587)**	11.387 (0.452)	11.782 (0.057)	—	16.088 (0.118)	15.823 (1.272)	*6.162 (0.458)*	11.695 (0.266)
361236	Coverage	0.975 (0.008)	0.975 (0.008)	0.961 (0.0)	0.948 (0.012)	0.948 (0.012)	0.938 (0.012)	0.965 (0.008)	0.934 (0.004)	*0.94 (0.004)*	0.963 (0.004)
	Length	44407.071 (1173.758)	44509.269 (1229.817)	50820.568 (385.951)	41045.069 (1221.808)	43185.942 (1002.516)	40905.295 (1089.411)	44437.938 (851.862)	60509.250 (2410.320)	*30953.589 (2482.065)*	**33439.322 (1275.213)**
361237	Coverage	0.969 (0.023)	0.969 (0.023)	0.981 (0.0)	0.923 (0.0)	0.954 (0.015)	0.9 (0.008)	0.969 (0.023)	0.885 (0.038)	*0.8 (0.015)*	0.977 (0.008)
	Length	44.019 (0.990)	44.069 (1.115)	27.035 (1.014)	—	26.524 (1.244)	—	31.967 (1.118)	—	*14.473 (0.503)*	**23.145 (0.199)**
361241	Coverage	0.954 (0.001)	0.956 (0.001)	0.944 (0.005)	0.957 (0.002)	0.954 (0.002)	0.923 (0.0)	0.952 (0.0)	0.949 (0.001)	*0.917 (0.006)*	0.951 (0.001)
	Length	19.133 (0.062)	20.245 (0.095)	18.102 (0.055)	18.482 (0.062)	17.958 (0.062)	—	18.932 (0.034)	29.548 (0.191)	*15.199 (0.427)*	**17.328 (0.097)**
361242	Coverage	0.944 (0.004)	0.955 (0.0)	0.947 (0.004)	0.942 (0.0)	0.948 (0.0)	0.914 (0.003)	0.944 (0.001)	0.949 (0.003)	*0.9 (0.006)*	0.944 (0.002)
	Length	70.248 (0.304)	84.510 (0.282)	**50.442 (0.421)**	54.844 (0.036)	54.217 (0.115)	—	65.635 (0.094)	61.613 (0.372)	*44.602 (0.170)*	57.935 (0.070)
361243	Coverage	0.922 (0.03)	0.952 (0.015)	0.956 (0.022)	0.952 (0.015)	0.937 (0.022)	0.893 (0.044)	0.937 (0.022)	0.919 (0.022)	*0.748 (0.126)*	0.956 (0.022)
	Length	—	71.388 (0.152)	75.924 (2.291)	72.877 (0.729)	**68.493 (0.993)**	—	72.048 (0.024)	—	*43.742 (10.285)*	**68.220 (1.605)**
361244	Coverage	0.97 (0.022)	0.97 (0.022)	0.97 (0.022)	0.97 (0.022)	0.97 (0.022)	0.97 (0.022)	0.97 (0.022)	0.974 (0.015)	*0.963 (0.037)*	0.956 (0.015)
	Length	3.274 (0.004)	3.274 (0.004)	3.336 (0.023)	3.284 (0.010)	3.272 (0.000)	3.269 (0.003)	3.289 (0.002)	4.854 (0.293)	*0.287 (0.008)*	**0.287 (0.008)**
361247	Coverage	0.96 (0.001)	0.953 (0.003)	0.94 (0.001)	0.951 (0.003)	0.963 (0.001)	0.903 (0.007)	0.951 (0.0)	0.954 (0.001)	*0.843 (0.003)*	0.943 (0.006)
	Length	0.025 (0.000)	0.038 (0.000)	**0.006 (0.000)**	0.016 (0.000)	0.015 (0.000)	—	0.022 (0.000)	0.013 (0.000)	*0.005 (0.000)*	**0.008 (0.000)**
361249	Coverage	0.96 (0.002)	0.956 (0.002)	0.962 (0.003)	0.972 (0.007)	0.953 (0.005)	0.938 (0.002)	0.965 (0.005)	0.936 (0.01)	*0.931 (0.0)*	0.953 (0.005)
	Length	3.008 (0.006)	3.068 (0.009)	2.800 (0.000)	2.780 (0.025)	2.775 (0.006)	**2.558 (0.019)**	2.894 (0.000)	4.706 (0.099)	*2.216 (0.020)*	**2.614 (0.043)**

Table C.3: Average coverages across tasks for $\alpha = 0.025$ are shown in the top row and average prediction set lengths in the bottom row, where both were assessed over a batch of i.i.d. test samples (20% of the dataset size). Standard deviations and means were computed across 5 randomizations of draws of the training, calibration, and test sets. In cases where the method failed to achieve sufficient coverage (defined as < 0.96), we do not include it in comparison for set length. Similarly, the single-stage approach fails to achieve coverage due to lack of exchangeability with test points.

Dataset	Metrics	Linear Model	LASSO	Random Forest	XGBoost	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	*CSA (Single-Stage)*	CSA
361234	Coverage	0.987 (0.008)	0.987 (0.008)	0.974 (0.004)	0.977 (0.008)	0.982 (0.008)	0.971 (0.01)	0.981 (0.01)	0.97 (0.008)	*0.976 (0.01)*	0.973 (0.006)
	Length	11.939 (0.137)	11.871 (0.084)	12.484 (0.168)	11.972 (0.009)	11.587 (0.110)	11.157 (0.086)	11.965 (0.050)	25.598 (0.974)	*9.306 (0.259)*	**8.855 (0.059)**
361235	Coverage	0.987 (0.0)	0.982 (0.011)	0.979 (0.011)	0.984 (0.005)	0.989 (0.005)	0.966 (0.016)	0.976 (0.005)	0.958 (0.021)	*0.889 (0.011)*	0.989 (0.005)
	Length	24.595 (0.825)	28.841 (1.129)	**11.811 (0.992)**	14.237 (0.786)	14.472 (0.172)	**12.278 (0.026)**	19.231 (0.356)	—	*7.719 (0.467)*	**12.563 (0.766)**
361236	Coverage	0.992 (0.004)	0.992 (0.004)	0.981 (0.0)	0.965 (0.008)	0.975 (0.008)	0.965 (0.008)	0.977 (0.012)	0.955 (0.008)	*0.955 (0.012)*	0.973 (0.004)
	Length	48591.496 (873.946)	48578.821 (867.718)	56132.760 (368.832)	46623.663 (1565.744)	47630.890 (810.373)	45714.911 (1062.420)	49188.464 (753.205)	—	*32881.580 (3110.734)*	**35777.096 (1949.538)**
361237	Coverage	0.981 (0.0)	0.981 (0.0)	0.981 (0.0)	0.977 (0.008)	0.962 (0.0)	0.962 (0.0)	0.977 (0.008)	0.965 (0.008)	*0.927 (0.031)*	0.981 (0.0)
	Length	47.738 (0.542)	47.440 (0.959)	30.785 (0.037)	**26.208 (0.897)**	30.554 (0.561)	**27.182 (0.803)**	35.982 (0.619)	67.660 (6.380)	*18.214 (0.436)*	**26.897 (0.515)**
361241	Coverage	0.979 (0.001)	0.978 (0.001)	0.976 (0.001)	0.978 (0.001)	0.978 (0.0)	0.964 (0.002)	0.977 (0.0)	0.972 (0.002)	*0.958 (0.003)*	0.979 (0.0)
	Length	21.772 (0.085)	23.089 (0.106)	21.543 (0.009)	21.454 (0.109)	20.862 (0.088)	**19.291 (0.060)**	21.905 (0.041)	40.082 (0.045)	*17.765 (0.329)*	19.897 (0.062)
361242	Coverage	0.977 (0.003)	0.978 (0.001)	0.975 (0.002)	0.968 (0.003)	0.973 (0.004)	0.955 (0.002)	0.971 (0.002)	0.975 (0.0)	*0.936 (0.001)*	0.977 (0.001)
	Length	83.892 (0.143)	99.811 (0.866)	**65.672 (0.212)**	68.119 (0.187)	68.155 (0.357)	—	80.032 (0.354)	85.678 (0.371)	*53.549 (0.585)*	69.388 (0.128)
361243	Coverage	0.985 (0.007)	0.985 (0.007)	0.985 (0.007)	0.956 (0.022)	0.985 (0.007)	0.97 (0.015)	0.985 (0.007)	0.978 (0.007)	*0.748 (0.126)*	0.985 (0.007)
	Length	92.698 (1.567)	87.949 (0.147)	88.993 (2.345)	—	84.569 (0.557)	**79.879 (0.739)**	87.950 (0.308)	137.673 (15.214)	*46.504 (12.957)*	**79.976 (2.585)**
361244	Coverage	0.974 (0.015)	0.974 (0.015)	0.974 (0.015)	0.974 (0.015)	0.974 (0.015)	0.974 (0.015)	0.974 (0.015)	0.989 (0.022)	*0.963 (0.037)*	0.974 (0.015)
	Length	5.274 (0.004)	5.274 (0.004)	5.336 (0.023)	5.284 (0.010)	5.272 (0.000)	5.269 (0.003)	5.289 (0.002)	11.283 (1.039)	*0.287 (0.008)*	**0.287 (0.008)**
361247	Coverage	0.98 (0.0)	0.976 (0.003)	0.969 (0.004)	0.974 (0.001)	0.977 (0.003)	0.945 (0.004)	0.971 (0.003)	0.982 (0.001)	*0.906 (0.003)*	0.974 (0.004)
	Length	0.029 (0.000)	0.042 (0.000)	**0.009 (0.000)**	0.019 (0.000)	0.018 (0.000)	—	0.025 (0.000)	0.016 (0.000)	*0.008 (0.000)*	0.012 (0.000)
361249	Coverage	0.981 (0.003)	0.981 (0.003)	0.984 (0.002)	0.991 (0.002)	0.981 (0.003)	0.976 (0.002)	0.981 (0.003)	0.971 (0.007)	*0.962 (0.011)*	0.977 (0.003)
	Length	3.645 (0.023)	3.674 (0.026)	3.600 (0.000)	3.322 (0.019)	3.402 (0.005)	3.201 (0.019)	3.543 (0.000)	6.533 (0.252)	*2.719 (0.164)*	**2.972 (0.064)**

C.5 CSA Prediction Region Visualizations

We now visualize some of the prediction regions corresponding to some of the trials run in Section C.4. While we find these intervals to be connected across these tasks, we expect visualizations over multivariate output spaces, i.e. for 2D regression problems, would reveal sets to be non-connected.

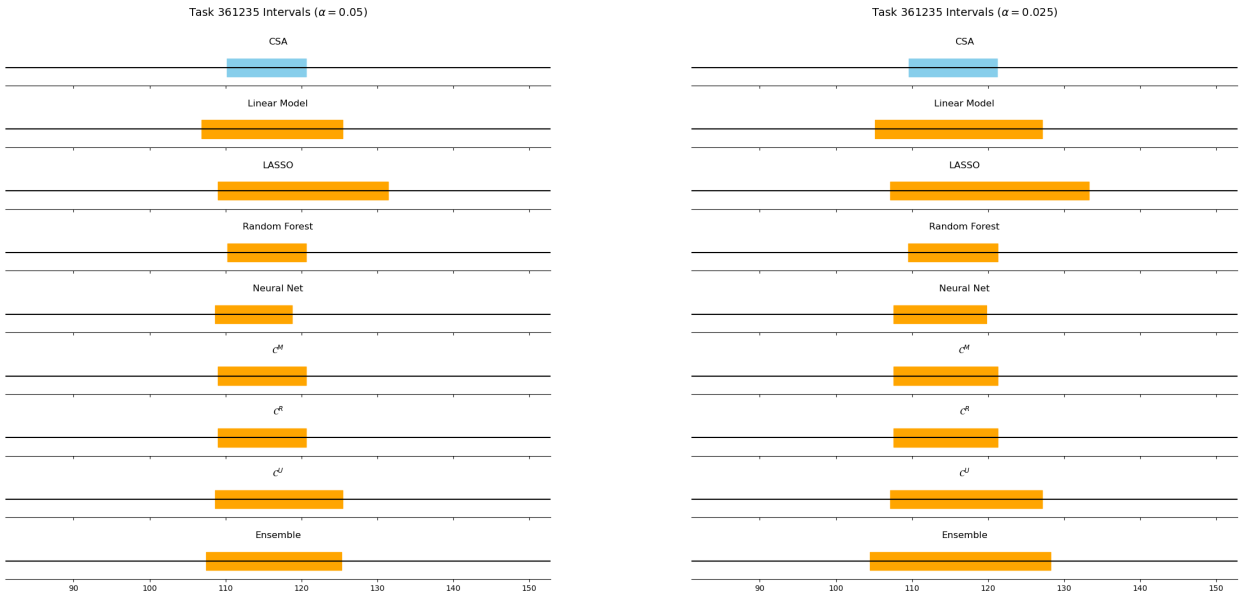

Figure C.3: Prediction regions across methods for task 361235 for $\alpha = 0.05$ (left) and 0.025 (right).

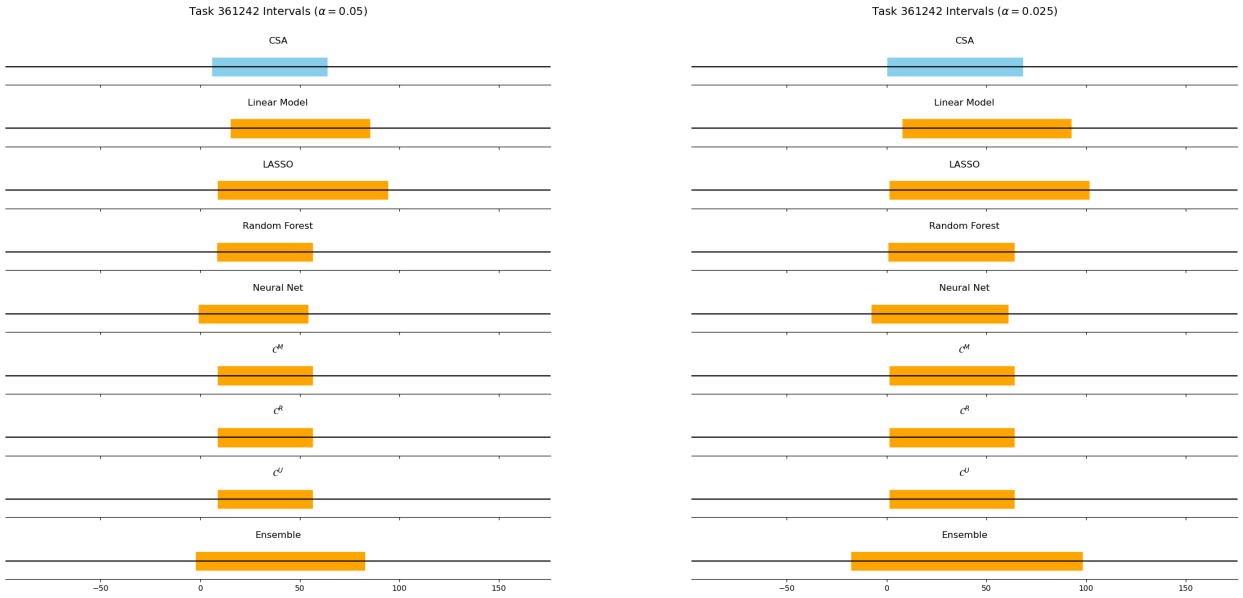

Figure C.4: Prediction regions across methods for task 361242 for $\alpha = 0.05$ (left) and 0.025 (right).

C.6 UCI Results

We consider those regression tasks from the UCI repository [Asuncion et al., 2007] that have at least 1,000 samples. The complete collection of results is presented in Table C.4. As discussed in the main text, across nearly all the UCI benchmark tasks, we find that the conformalized random forest does optimally and that ensembling methods provide no further benefit over simply taking this single predictor. In this degenerate case, we would expect an optimal aggregator to simply then return this optimal single predictor. We then find CSA to consistently significantly outperform other aggregation strategies and to return a prediction region of size comparable to that of the nominal random forest.

Table C.4: Average coverages across tasks for $\alpha = 0.05$ are shown in the top row and average prediction set lengths in the bottom row, where both were assessed over a batch of i.i.d. test samples (10% of the dataset size). We are highlighting the robustness compared to other aggregation strategies here and so bold the best performing amongst the aggregation methods. Standard deviations and means were computed across 5 randomizations of draws of the training, calibration, and test sets. Note that, while the single-stage prediction regions are the smallest, they fail to achieve the desired coverage level and are, therefore, precluded from comparison.

Dataset	Metrics	Linear Model	LASSO	Random Forest	XGBoost	$\mathcal{C}^M$	$\mathcal{C}^R$	$\mathcal{C}^U$	Ensemble	*CSA (Single-Stage)*	CSA
airfoil	Coverage	0.966 (0.011)	0.926 (0.011)	0.934 (0.026)	0.966 (0.011)	0.945 (0.021)	0.916 (0.016)	0.934 (0.026)	0.958 (0.032)	*0.805 (0.058)*	0.953 (0.011)
	Length	19.062 (0.615)	—	11.368 (0.188)	11.770 (0.261)	**12.371 (0.131)**	—	16.227 (0.025)	16.097 (1.052)	*8.609 (0.719)*	14.075 (0.734)
bike	Coverage	0.944 (0.007)	0.947 (0.004)	0.954 (0.002)	0.958 (0.003)	0.938 (0.006)	0.908 (0.0)	0.946 (0.003)	0.955 (0.0)	*0.963 (0.007)*	0.947 (0.003)
	Length	3.178 (0.011)	3.343 (0.021)	0.065 (0.000)	0.111 (0.000)	**0.111 (0.001)**	—	1.722 (0.007)	0.682 (0.004)	*0.156 (0.019)*	0.134 (0.007)
concrete	Coverage	0.977 (0.008)	0.977 (0.008)	0.981 (0.0)	0.915 (0.023)	0.962 (0.0)	0.915 (0.023)	0.981 (0.0)	0.915 (0.023)	*0.854 (0.054)*	0.977 (0.008)
	Length	44.295 (0.424)	44.470 (0.326)	26.053 (2.114)	—	**26.488 (1.141)**	—	32.455 (1.165)	—	*19.712 (6.592)*	**25.302 (3.244)**
kin40k	Coverage	0.949 (0.002)	0.949 (0.001)	0.946 (0.002)	0.948 (0.003)	0.945 (0.002)	0.919 (0.002)	0.949 (0.0)	0.945 (0.002)	*0.907 (0.004)*	0.941 (0.006)
	Length	3.781 (0.017)	3.781 (0.016)	2.333 (0.026)	3.343 (0.026)	3.269 (0.018)	—	3.291 (0.009)	4.985 (0.003)	*2.138 (0.001)*	**2.456 (0.042)**
parkinsons	Coverage	0.937 (0.003)	0.946 (0.0)	0.958 (0.009)	0.953 (0.007)	0.936 (0.001)	0.904 (0.008)	0.936 (0.005)	0.955 (0.004)	*0.873 (0.043)*	0.951 (0.003)
	Length	35.957 (0.323)	36.430 (0.328)	3.254 (0.423)	11.268 (0.114)	10.992 (0.074)	—	21.470 (0.003)	14.161 (0.083)	*3.457 (0.727)*	**4.584 (0.637)**
pol	Coverage	0.944 (0.001)	0.942 (0.002)	0.951 (0.004)	0.955 (0.001)	0.938 (0.001)	0.909 (0.003)	0.946 (0.001)	0.953 (0.005)	*0.884 (0.009)*	0.952 (0.003)
	Length	97.944 (0.150)	97.771 (0.386)	28.000 (0.000)	48.572 (1.279)	45.056 (0.821)	—	69.078 (0.362)	57.432 (0.663)	*24.321 (1.370)*	**33.230 (1.569)**
protein	Coverage	0.958 (0.001)	0.957 (0.002)	0.953 (0.003)	0.959 (0.003)	0.957 (0.003)	0.928 (0.003)	0.953 (0.003)	0.955 (0.0)	*0.91 (0.002)*	0.963 (0.004)
	Length	2.316 (0.000)	2.412 (0.011)	2.151 (0.014)	2.210 (0.006)	2.134 (0.002)	—	2.269 (0.001)	3.707 (0.039)	*1.717 (0.036)*	**1.994 (0.000)**
pumadyn32nm	Coverage	0.961 (0.011)	0.961 (0.01)	0.947 (0.001)	0.962 (0.003)	0.96 (0.007)	0.935 (0.003)	0.95 (0.013)	0.957 (0.013)	*0.906 (0.026)*	0.963 (0.001)
	Length	3.997 (0.024)	3.979 (0.031)	1.525 (0.027)	3.518 (0.058)	3.507 (0.051)	2.350 (0.043)	3.242 (0.033)	5.351 (0.139)	*1.564 (0.048)*	**1.858 (0.078)**
tamielectric	Coverage	0.953 (0.002)	0.953 (0.002)	0.947 (0.003)	0.953 (0.003)	0.952 (0.003)	0.926 (0.007)	0.949 (0.0)	0.948 (0.003)	*0.909 (0.003)*	0.949 (0.002)
	Length	0.950 (0.001)	0.951 (0.001)	1.271 (0.006)	0.953 (0.001)	0.948 (0.001)	—	1.029 (0.003)	4.539 (0.038)	*0.774 (0.005)*	**0.799 (0.004)**
wine	Coverage	0.96 (0.005)	0.943 (0.01)	0.931 (0.015)	0.923 (0.02)	0.928 (0.005)	0.884 (0.015)	0.948 (0.005)	0.946 (0.015)	*0.805 (0.054)*	0.933 (0.015)
	Length	2.352 (0.002)	3.521 (0.025)	2.619 (0.050)	—	—	—	2.621 (0.039)	3.390 (0.042)	*1.683 (0.224)*	**2.291 (0.026)**

C.7 Conformal Aggregation Methods

We now describe the methods from [Gasparin and Ramdas, 2024a] that were compared against experimentally, specifically the standard majority-vote $\mathcal{C}^M$, partially randomized thresholding $\mathcal{C}^R$, and fully randomized thresholding $\mathcal{C}^U$ approaches. As discussed in Section 4.2.3, these methods all follow the structural form of

$$\mathcal{C}(x) := \left\{ y \mid \sum_{k=1}^{K} w_k \mathbb{1}[y \in \mathcal{C}_k(x)] \geq \widehat{a} \right\} \tag{C.4}$$

and largely differ in their choice of weights and thresholds. The standard majority-vote $\mathcal{C}^M$ is the most natural choice, defined by

$$\mathcal{C}^M(x) := \left\{ y \mid \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}[y \in \mathcal{C}_k(x)] > \frac{1}{2} \right\}. \tag{C.5}$$

The randomized methods differ in that independent randomization is leveraged over the threshold, namely with:

$$\mathcal{C}^R(x) := \left\{ y \mid \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}[y \in \mathcal{C}_k(x)] > \frac{1}{2} + \frac{U}{2} \right\} \tag{C.6}$$

$$\mathcal{C}^U(x) := \left\{ y \mid \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}[y \in \mathcal{C}_k(x)] > U \right\}, \tag{C.7}$$

for $U \sim \mathrm{Unif}([0,1])$. Notably, all these methods retain the guarantees typical of conformal prediction.

C.8 Conformal Robust Control of Linear Systems

C.8.1 Prediction Region Validity Lemma

Given that we characterize *both* the continuous- and discrete-time settings below, we produce a generalized definition to that presented in Definition 1.

Definition 2. *Let $M(C, K^*(C)) := A - BK^*(C)$ be the optimal closed-loop matrix. Define*

$$r(C, K^*(C)) := \begin{cases} \frac{\min_i -\mathrm{Re}(\lambda_i(M))}{\kappa(U)\,||W||_2} & \textit{Continuous time setting} \\ \frac{\min_i (1-|\lambda_i(M)|)}{\kappa(U)\,||W||_2} & \textit{Discrete time setting} \end{cases}$$

where $M = U\Lambda U^{-1}$, $\kappa(U)$ is the condition number of U, and $W = [I \ \ -K^*(C)^\top]^\top$.

Lemma C.8.1. *Let $J(K, C)$ be a function such that, for any fixed θ, it is non-negative and L-Lipschitz in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm for any $K \in \mathcal{K}(\mathcal{U}(\theta))$, where $\mathcal{K} : \Omega(\Theta) \to \Omega(\mathcal{C})$ and $\Omega_1 \subset \Omega_2 \implies \mathcal{K}(\Omega_2) \subset \mathcal{K}(\Omega_1)$. Further assume that $\mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq 1 - \alpha$. Then:*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\mathrm{dom}}(\Theta, C)\right) \geq 1 - \alpha. \tag{C.8}$$

Further, if $\widehat{q} < r(C, K^*(C))$, (see Definition 2) $\Delta_{\mathrm{dom}}(\Theta, C) = 0$.

Proof. We consider the event of interest conditionally on a pair (θ, C) where $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$. By assumption, we then have that $\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta))) \subset \mathcal{K}(C)$. As previously noted, the suboptimality here is defined over the *true* C matrix, meaning, unlike previous works, we here wish to bound $J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) - \min_{K \in \mathcal{K}(C)} J(K, C)$ in place of $\min_{K \in \mathcal{K}(C)} \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} J(K, \widehat{C}) - \min_{K \in \mathcal{K}(C)} J(K, C)$, where $K^*(\mathcal{B}_{\widehat{q}}(f(\theta))) := \arg\min_{K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} J(K, \widehat{C})$. We

begin by matching the minimization sets in the terms as follows:

$$\begin{aligned}
&\left|J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) - \min_{K\in\mathcal{K}(C)} J(K, C)\right| \\
&= \left|J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) - \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C) + \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C) - \min_{K\in\mathcal{K}(C)} J(K, C)\right| \\
&\leq \underbrace{\left|J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) - \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C)\right|}_{\text{statistical robustness cost}} + \underbrace{\left|\min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C) - \min_{K\in\mathcal{K}(C)} J(K, C)\right|}_{\text{universal stabilization cost}}
\end{aligned}$$

As discussed in the main text, the error decomposes into two terms: the first from making the controller robust to adversarial dynamics matrices and the second from requiring that such a controller stabilize the whole collection of dynamics. We now bound each term separately, starting with the first term. We first note:

$$J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) \leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), \widehat{C}) =: \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} J(K, \widehat{C})$$

where the first step follows by the assumption $C \in \mathcal{B}_{\widehat{q}}(f(\theta))$ and second by definition of $K^*(\mathcal{B}_{\widehat{q}}(f(\theta)))$. From here,

$$\begin{aligned}
&\left|J(K^*(\mathcal{B}_{\widehat{q}}(f(\theta))), C) - \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C)\right| \\
&\leq \left|\min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} J(K, \widehat{C}) - \min_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} J(K, C)\right| \\
&\leq \max_{K\in\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))} |\max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} J(K, \widehat{C}) - J(K, C)| \\
&\leq L \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} ||\widehat{C} - C||_{\text{op}} \leq 2L\widehat{q}.
\end{aligned}$$

We now demonstrate that the second term vanishes within a radius of "safety," which we do through perturbation analysis of the closed-loop margin. In particular, notice that, since $\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta))) \subset \mathcal{K}(C)$, this term vanishes if the minimizer over $\mathcal{K}(C)$ lies in $\mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$. That is, this difference vanishes if $K^*(C) := \arg\min_{K\in\mathcal{K}(C)} J(K, C)$ satisfies $K^*(C) \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$. Notably, this is equivalent to finding a condition under which $K^*(C)$ stabilizes all the dynamics matrices $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$.

To procure a test of this property, we consider an approach from the theory of switched linear systems, where controllers are sought that stabilize a collection of adjusting dynamics matrices. In particular, recall that $\mathcal{B}_{\widehat{q}}(f(\theta))$ is constructed under an operator norm and, therefore, any $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ can be viewed as a bounded perturbation to the nominal prediction corresponding to θ.

That is, that $\widehat{C} = C + \Delta$ for $||\Delta||_{\text{op}} \leq \widehat{q}$. Thus, we can equivalently view the task as seeking a condition that guarantees that, if $||C - \widehat{C}||_{\text{op}} \leq \widehat{q}$, $K^*(C)$ stabilizes $\widehat{C}$.

We now consider the $W := [I_{n\times n} \; -K^*(C)^\top]^\top$ matrix and analyze the closed-loop stability of $\widehat{C}W$. By definition, we have that $\widehat{C}W := (C + \Delta)W =: CW + E$, where we know that CW is stabilized by $K^*(C)$, since $K^*(C) \in \mathcal{K}(C)$. By this latter point, we know CW satisfies one of two properties, depending on whether the system being analyzed is a continuous- or discrete-time setting. In the case of continuous time, we have that $\min_i -\text{Re}(\lambda_i(CW)) > 0$ and that, for discrete time, $\min_i(1 - |\lambda_i(CW)|) > 0$.

Under the additional assumption of CW being diagonalizable, we have that $CW = U\Lambda U^{-1}$ for $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$. With this, we now return to analyzing the perturbed $CW + E$. By the Bauer–Fike bound, we know that for any eigenvalue λ'_j of $CW + E$, $\min_i |\lambda'_j - \lambda_i| \leq \kappa(U)||E||_{\text{op}}$, where $\kappa(U)$ is the condition number of U. Thus, if these perturbed eigenvalues remain within the stabilized regions, $\widehat{C}$ is stabilized by $K^*(C)$.

In the continuous-time setting, this is guaranteed if $\kappa(U)||E||_{\text{op}} < \min_i -\text{Re}(\lambda_i(CW))$ or, by the fact that $||E|| := ||\Delta W|| \leq ||\Delta|| \cdot ||W|| \leq \widehat{q}||W||$, if

$$\kappa(U)(\widehat{q}||W||) < \min_i -\text{Re}(\lambda_i(CW)) \iff \widehat{q} < \frac{\min_i -\text{Re}(\lambda_i(CW))}{\kappa(U)||W||}$$

The discrete-time setting follows analogously, simply with the stability condition replaced by the discrete-time analog, i.e. if $\kappa(U)||E||_{\text{op}} < \min_i(1 - |\lambda_i(CW)|)$. That is, a sufficient condition for stabilization is that

$$\kappa(U)(\widehat{q}||W||) < \min_i(1 - |\lambda_i(CW)|) \iff \widehat{q} < \frac{\min_i(1 - |\lambda_i(CW)|)}{\kappa(U)||W||}$$

Since we have the assumption that $\mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq 1 - \alpha$, the result immediately follows. ■

C.8.2 Deterministic Discrete-Time Regret Analysis

Theorem C.8.2. *[Deterministic, discrete-time] Let $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top(Q + K^\top RK)x_t)$ with $w = 0$. Assume that $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$. Then, under Assumption 4,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L is the Lipschitz constant of $J(K, \widehat{C})$ in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm. Further, if $\widehat{q} < r(C, K^*(C))$, (see discrete-time in Definition 2) $\Delta_{\text{dom}}(\Theta, C) = 0$.

Proof. We consider any fixed θ and demonstrate that $J(K, \widehat{C})$ is non-negative and L-Lipschitz in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm for any $K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, from which Theorem C.8.1 can

be invoked to arrive at the desired conclusion. Given the assumed determinism of the dynamics, we have that $x_t = (\widehat{C}W)^t x_0$, meaning the above objective setup can equivalently be expressed as:

$$\sum_{t=0}^{\infty} x_0^\top ((\widehat{C}W)^{t\top}(Q + K^\top RK)(\widehat{C}W)^t)x_0, \tag{C.9}$$

J is clearly non-negative by construction. It, therefore, suffices to demonstrate this objective is Lipschitz continuous with an appropriate Lipschitz constant. Notice the Lipschitz constant can be obtained by bounding the magnitude of the gradient with respect to $\widehat{C}$, which we do as follows

$$\begin{aligned}
&\nabla_{\widehat{C}} \left(\sum_{t=0}^{\infty} x_0^\top ((\widehat{C}W)^{t\top}(Q + K^\top RK)(\widehat{C}W)^t)x_0 \right) \\
&= \sum_{t=0}^{\infty} t\mathrm{diag}((Q + K^\top RK)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top \\
&+ \sum_{t=0}^{\infty} t\mathrm{diag}((Q^\top + (RK)^\top K)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top
\end{aligned}$$

We now bound the magnitude of this quantity as follows:

$$\begin{aligned}
L &\leq \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} || \sum_{t=0}^{\infty} t\mathrm{diag}((Q + K^\top RK)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top \\
&\quad + \sum_{t=0}^{\infty} t\mathrm{diag}((Q^\top + (RK)^\top K)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top ||_{\mathrm{op}} \\
&\leq \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t \left|\left| \mathrm{diag}((Q + K^\top RK)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top \right|\right|_{\mathrm{op}} \\
&\quad + \sum_{t=0}^{\infty} t \left|\left| \mathrm{diag}((Q^\top + (RK)^\top K)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\mathrm{diag}(x_0)W^\top \right|\right|_{\mathrm{op}},
\end{aligned}$$

where we have used $\mathrm{diag}(x_0)$ for a vector x_0 to denote a diagonal matrix with x_0 placed along its main diagonal. We now bound each of these two terms separately, although the structure of the two is the same, so we explicitly show steps for bounding the first, from which the same can be repeated on the second. Importantly, we make use of the fact $||\mathrm{diag}(x_0)||_{\mathrm{op}} = ||x_0||_\infty$ and

$||A||_\infty \leq \sqrt{n}||A||_{\text{op}}$ for $A \in \mathbb{R}^{n\times n}$ as follows:

$$\begin{aligned}
&\max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||\text{diag}((Q+K^\top RK)(\widehat{C}W)^t x_0)(\widehat{C}W)^{(t-1)}\text{diag}(x_0)W^\top||_{\text{op}} \\
&\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||\text{diag}((Q+K^\top RK)(\widehat{C}W)^t x_0)||_{\text{op}}||(\widehat{C}W)^{(t-1)}||_{\text{op}}||\text{diag}(x_0)||_{\text{op}}||W^\top||_{\text{op}} \\
&= \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||(Q+K^\top RK)(\widehat{C}W)^t x_0||_\infty||x_0||_\infty||(\widehat{C}W)^{(t-1)}||_{\text{op}}||W||_{\text{op}} \\
&\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||Q+K^\top RK||_\infty||(\widehat{C}W)^t||_\infty||x_0||_\infty||x_0||_\infty||(\widehat{C}W)^{(t-1)}||_{\text{op}}||W||_{\text{op}} \\
&\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t(\sqrt{n}||Q+K^\top RK||_\infty||x_0||_\infty^2||W||_{\text{op}})||(\widehat{C}W)^t||_{\text{op}}||(\widehat{C}W)^{(t-1)}||_{\text{op}}.
\end{aligned}$$

We now collect all terms independent of t into $D(K) = \sqrt{n}||Q+K^\top RK||_\infty||x_0||_\infty^2||W||_{\text{op}}$:

$$\leq D(K) \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||(\widehat{C}W)^t||_{\text{op}}||(\widehat{C}W)^{(t-1)}||_{\text{op}}.$$

Critically, we can now demonstrate that this sum is bounded by virtue of K being a universal stabilizer of the dynamics set $\mathcal{B}_{\widehat{q}}(f(\theta))$. By this stabilization, we know that $\widehat{C}W$ is Schur stable, i.e. $\min_i(1-|\lambda_i(\widehat{C}W)|) > 0$ or $\max_i|\lambda_i(\widehat{C}W)| < 1$. Thus, there exists a $P \succ 0$ such that for some $\tau \in (0,1)$, we have

$$(\widehat{C}W)^\top P(\widehat{C}W) \preceq \tau^2 P.$$

We now denote the norm induced by such a P as $||\cdot||_P$. Then, $||\widehat{C}W||_P \leq \tau$, meaning $||(\widehat{C}W)^t|| \leq \tau^t$. By the norm equivalence between the induced matrix norm and the standard operator norm, we have that

$$||(\widehat{C}W)^t||_{\text{op}} \leq \kappa(P)||(\widehat{C}W)^t||_P \leq \kappa(P)\tau^t,$$

where $\kappa(P) := \sqrt{\lambda_{\max}(P)/\lambda_{\min}(P)}$ is the condition number of P. We now see that the previous sum converges:

$$\begin{aligned}
D(K) \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t||(\widehat{C}W)^t||_{\text{op}}||(\widehat{C}W)^{(t-1)}||_{\text{op}} &\leq D(K)\kappa(P)^2 \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty} t\tau^{2t-1} \\
&\leq D(K)\kappa(P)^2 \frac{\tau}{(1-\tau^2)^2},
\end{aligned}$$

completing the proof. ■

The finite LQR case follows immediately as a corollary of the above, stated below for completeness.

Corollary C.8.3. *[Deterministic, discrete-time, finite-horizon] Let* $J(K, C) := \sum_{t=0}^{T}(x_t^\top(Q + K^\top RK)x_t)$ *with* $w = 0$*. Assume that* $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$*. Then:*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L *is the Lipschitz constant of* $J(K, \widehat{C})$ *in* $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ *under the operator norm. Further, if* $\widehat{q} < r(C, K^*(C))$*, (see discrete-time in Definition 2)* $\Delta_{\text{dom}}(\Theta, C) = 0$*.*

C.8.3 Deterministic Continuous-Time Regret Analysis

The proof follows in much the same manner as the discrete-time case, with modest adjustments to the exact specification of the assumptions regarding the dynamics.

Theorem C.8.4. *[Deterministic, continuous-time] Let* $J(K, C) := \int_0^\infty (x(t)^\top (Q + K^\top RK)x(t))dt$ *for* $w = 0$*. Assume that* $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$*. Then, under Assumption 4,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L *is the Lipschitz constant of* $J(K, \widehat{C})$ *in* $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ *under the operator norm. Further, if* $\widehat{q} < r(C, K^*(C))$*, (see continuous-time in Definition 2)* $\Delta_{\text{dom}}(\Theta, C) = 0$*.*

Proof. We consider any fixed θ and demonstrate the desired properties that $J(K, C)$ is non-negative and L-Lipschitz in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm for any $K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, from which Theorem C.8.1 can be invoked to arrive at the desired conclusion. Given the assumed determinism of the dynamics, we further have that $x(t) = e^{\widehat{C}Wt}x(0)$, meaning the above objective setup can equivalently be expressed with the uncertainty sets related to the objective function, namely as:

$$\int_0^\infty x(0)^\top (e^{\widehat{C}Wt})^\top (Q + K^\top RK)(e^{\widehat{C}Wt})x(0)dt. \tag{C.10}$$

J is clearly non-negative by construction. It, therefore, suffices to demonstrate this objective is Lipschitz continuous with an appropriate Lipschitz constant. We again proceed by bounding the

norm of the gradient as follows:

$$\begin{aligned}
&\nabla_{\widehat{C}} \left(\int_0^\infty x(0)^\top (e^{\widehat{C}Wt})^\top (Q + K^\top RK)(e^{\widehat{C}Wt}) x(0) dt \right) \\
&= \int_0^\infty t \mathrm{diag}((Q + K^\top RK) e^{\widehat{C}Wt} x(0)) e^{\widehat{C}Wt} \mathrm{diag}(x(0)) W^\top dt \\
&+ \int_0^\infty t \mathrm{diag}((Q^\top + (RK)^\top K) e^{\widehat{C}Wt} x(0)) e^{\widehat{C}Wt} \mathrm{diag}(x(0)) W^\top dt
\end{aligned}$$

We again bound each of these two terms separately, as follows:

$$\begin{aligned}
&\max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \left|\left| \int_0^\infty t \mathrm{diag}((Q + K^\top RK) e^{\widehat{C}Wt} x(0)) e^{\widehat{C}Wt} \mathrm{diag}(x(0)) W^\top dt \right|\right|_{\mathrm{op}} \\
&\leq \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \int_0^\infty t \left|\left| \mathrm{diag}((Q + K^\top RK) e^{\widehat{C}Wt} x(0)) \right|\right|_{\mathrm{op}} \left|\left| e^{\widehat{C}Wt} \right|\right|_{\mathrm{op}} ||\mathrm{diag}(x(0))||_{\mathrm{op}} \left|\left| W^\top \right|\right|_{\mathrm{op}} dt \\
&= \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \int_0^\infty t \left|\left| (Q + K^\top RK) e^{\widehat{C}Wt} x(0) \right|\right|_{\infty} \left|\left| e^{\widehat{C}Wt} \right|\right|_{\mathrm{op}} ||x(0)||_{\infty} ||W||_{\mathrm{op}} dt \\
&\leq \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \int_0^\infty t (\sqrt{n} \left|\left| Q + K^\top RK \right|\right|_{\infty} ||W||_{\mathrm{op}} ||x(0)||_{\infty}^2) \left|\left| e^{\widehat{C}Wt} \right|\right|_{\mathrm{op}}^2 dt.
\end{aligned}$$

Collecting all terms independent of t into a constant $D(K) = \sqrt{n}||Q+K^\top RK||_\infty ||W||_{\mathrm{op}} ||x(0)||_\infty^2$ and using the bound $||e^{\widehat{C}Wt}|| \leq \beta(\widehat{C}) e^{-\alpha(\widehat{C})t}$, we reach the conclusion as:

$$= \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} D(K) \int_0^\infty t \left|\left| e^{\widehat{C}Wt} \right|\right|_{\mathrm{op}}^2 dt \leq D(K) \beta(\widehat{C})^2 \int_0^\infty t e^{-2\alpha(\widehat{C})t} dt = \max_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))} \frac{D(K)\beta(\widehat{C})^2}{4\alpha(\widehat{C})^2},$$

as desired. ■

Once again, the proof in the finite time horizon case follows equivalently.

Corollary C.8.5. *[Deterministic, continuous-time, finite-horizon] Let* $J(K, C) := \int_0^T (x(t)^\top (Q + K^\top RK) x(t)) dt$ *for* $w = 0$*. Assume that* $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$*. Then, under Assumption 4,*

$$\mathcal{P}_{\Theta,C} \left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\mathrm{dom}}(\Theta, C) \right) \geq 1 - \alpha,$$

where L *is the Lipschitz constant of* $J(K, \widehat{C})$ *in* $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ *under the operator norm. Further, if* $\widehat{q} < r(C, K^*(C))$*, (see continuous-time in Definition 2)* $\Delta_{\mathrm{dom}}(\Theta, C) = 0$*.*

C.8.4 Stochastic Discrete-Time Regret Analysis

We introduce below the discrete-time analog of the continuous-time decay assumption made in Assumption 5 below.

Assumption 6. *For any θ, $\exists$ constants $\alpha_1, \beta_1 > 0$ such that for all $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$, $K \in \mathcal{K}(\widehat{C})$, and $t \geq 0$, $||Q(t) + K^\top R(t)K|| \leq \beta_1 e^{-\alpha_1 t}$ and $\min_{\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))}(2\alpha_2(\widehat{C}) + \alpha_1) > 0$ where $\alpha_2(\widehat{C}) := \inf_{K \in \mathcal{K}(\widehat{C})}(-\log \rho(\widehat{C}W)) > 0$.*

Theorem C.8.6. *[Stochastic, discrete-time] Let $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top (Q + K^\top R K)x_t)$ with $w_t \sim \mathcal{N}(0, \Sigma)$ i.i.d. across t such that $D_2(K) := ||\Sigma||_{\text{op}}||W||_{\text{op}} < \infty$. Assume further that $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1 - \alpha$. Then, under Assumption 4 and Assumption 6,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta, C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta, C)\right) \geq 1 - \alpha,$$

where L is the Lipschitz constant of $J(K, \widehat{C})$ in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm. Further, if $\widehat{q} < r(C, K^*(C))$, (see discrete-time in Definition 2) $\Delta_{\text{dom}}(\Theta, C) = 0$.

Proof. We again consider any fixed θ and demonstrate the desired properties that $J(K, \widehat{C})$ is non-negative and L-Lipschitz in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm for any $K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, from which Theorem C.8.1 can be invoked to arrive at the desired conclusion. Notice the objective can be reformulated in the standard manner as follows:

$$\begin{aligned}
J(K, \widehat{C}) &:= \mathbb{E}[\sum_{t=0}^{\infty}(x_t^\top (Q_t + K^\top R_t K)x_t)] = \sum_{t=0}^{\infty}\mathbb{E}[(x_t^\top (Q_t + K^\top R_t K)x_t)] \\
&= \sum_{t=0}^{\infty}\mathbb{E}[\text{Tr}(x_t^\top (Q_t + K^\top R_t K)x_t)] = \sum_{t=0}^{\infty}\mathbb{E}[\text{Tr}((Q_t + K^\top R_t K)x_t x_t^\top)]] \\
&= \sum_{t=0}^{\infty}\text{Tr}((Q_t + K^\top R_t K)\mathbb{E}[x_t x_t^\top]) = \sum_{t=0}^{\infty}\text{Tr}((Q_t + K^\top R_t K)(\mathbb{E}[x_t]\mathbb{E}[x_t]^\top + \text{Var}(x_t))).
\end{aligned}$$

We now use the following computations to evaluate this final expression:

$$\begin{aligned}
\mathbb{E}[x_t] &= \mathbb{E}[(\widehat{C}W)x_{t-1} + w_t] \\
&= \mathbb{E}[(\widehat{C}W)x_{t-1}] + \mathbb{E}[w_t] = \mathbb{E}[(\widehat{C}W)((\widehat{C}W)x_{t-2} + w_{t-1})] \\
&= \mathbb{E}[(\widehat{C}W)^2 x_{t-2}] + (\widehat{C}W)\mathbb{E}[w_{t-1})] = ... = (\widehat{C}W)^t x_0.
\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(x_t) &= \mathrm{Var}((\widehat{C}W)x_{t-1} + w_t) = (\widehat{C}W)\mathrm{Var}(x_{t-1})(\widehat{C}W)^\top + \Sigma \\ &= (\widehat{C}W)\mathrm{Var}(x_{t-2})(\widehat{C}W)^\top + (\widehat{C}W)\Sigma(\widehat{C}W)^\top + \Sigma = ... = \sum_{k=0}^{t-1}(\widehat{C}W)^k\Sigma(\widehat{C}W)^{k\top}.\end{aligned}$$

With these simplifications, we are left with:

$$\begin{aligned}J(K, \widehat{C}) = &\sum_{t=0}^{\infty} x_0^\top(\widehat{C}W)^{t\top}(Q_t + K^\top R_t K)(\widehat{C}W)^t x_0 \\ &+ \sum_{t=0}^{\infty}\sum_{k=0}^{t-1}\mathrm{Tr}((Q_t + K^\top R_t K)(\widehat{C}W)^k\Sigma(\widehat{C}W)^{k\top}))\end{aligned}$$

J is clearly non-negative by construction. We demonstrate this quantity is Lipschitz continuous with an appropriate Lipschitz constant, again by bounding the gradient. The bound for the first term was demonstrated in the proof of Theorem C.8.2, for which reason we solely present that of the second term as follows:

$$\begin{aligned}&\sum_{t=0}^{\infty}\sum_{k=0}^{t-1}\nabla_{\widehat{C}}\mathrm{Tr}((Q_t + K^\top R_t K)(\widehat{C}W)^k\Sigma(\widehat{C}W)^{k\top})) \\ &= \sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k((((Q_t + K^\top R_t K)^\top(\widehat{C}W)^k\Sigma^\top) \odot (\widehat{C}W)^{(k-1)})W^\top + \\ &\sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k(((Q_t + K^\top R_t K)(\widehat{C}W)^k\Sigma) \odot (\widehat{C}W)^{(k-1)})W^\top\end{aligned}$$

We now bound each of these two terms separately, although the structure of the two is the same, so we explicitly show steps for bounding the first, from which the same can be repeated on the second.

$$\begin{aligned}&\max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} ||\sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k((((Q_t + K^\top R_t K)^\top(\widehat{C}W)^k\Sigma^\top) \odot (\widehat{C}W)^{(k-1)})W^\top||_{\mathrm{op}} \\ &\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k||((Q_t + K^\top R_t K)^\top(\widehat{C}W)^k\Sigma^\top) \odot (\widehat{C}W)^{(k-1)}||_{\mathrm{op}}||W||_{\mathrm{op}} \\ &\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k||Q_t + K^\top R_t K||_{\mathrm{op}}||(\widehat{C}W)^k||_{\mathrm{op}}||\Sigma||_{\mathrm{op}}||(\widehat{C}W)^{(k-1)}||_{\mathrm{op}}||W||_{\mathrm{op}} \\ &\leq \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty}\sum_{k=0}^{t-1}k\beta_1 e^{-\alpha_1 t}||(\widehat{C}W)^k||_{\mathrm{op}}||\Sigma||_{\mathrm{op}}||(\widehat{C}W)^{(k-1)}||_{\mathrm{op}}||W||_{\mathrm{op}}\end{aligned}$$

We again repeat the proof technique leveraged in Section C.8.2, where we first collect all terms independent of t into a constant $D_2(K) = ||\Sigma||_{\text{op}}||W||_{\text{op}}$. By precisely the same argument as presented there, namely that K stabilizes $\widehat{C}$, we have that there is a $P \succ 0$ such that for some $\tau \in (0,1)$, we have

$$(\widehat{C}W)^\top P(\widehat{C}W) \preceq \tau^2 P. \implies ||(\widehat{C}W)^t||_{\text{op}} \leq \kappa(P)\tau^t,$$

where $\kappa(P)$ is the condition number of P. By Assumption 6, we have that $\alpha_2(\widehat{C}) \leq -\log(\tau)$ or $e^{-\alpha_2(\widehat{C})} \geq \tau$. Therefore, we have that the prior sum is bounded by

$$\begin{aligned}
&\leq D_2(K)\kappa(P)^2 \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty}\sum_{k=0}^{t-1} k\tau^{2k-1} \\
&\leq D_2(K)\kappa(P)^2 \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{t=0}^{\infty}\beta_1 e^{-\alpha_1 t}\sum_{k=0}^{t-1} k e^{-\alpha_2(\widehat{C})(2k-1)} \\
&\leq D_2(K)\kappa(P)^2\beta_1 \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{k=0}^{t-1} k e^{-\alpha_2(\widehat{C})(2k-1)}\sum_{t=k}^{\infty} e^{-\alpha_1 t} \\
&= D_2(K)\kappa(P)^2\beta_1 \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{k=0}^{t-1} k e^{-\alpha_2(\widehat{C})(2k-1)}\left(\frac{e^{\alpha_1-\alpha_1 k}}{e^{\alpha_1}-1}\right) \\
&= D_2(K)\kappa(P)^2\beta_1\left(\frac{1}{e^{\alpha_1}-1}\right) \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \sum_{k=0}^{t-1} k(e^{-\alpha_2(\widehat{C})})^{(2k-1)}(e^{-\alpha_1})^{k-1} \\
&= D_2(K)\kappa(P)^2\beta_1\left(\frac{1}{e^{\alpha_1}-1}\right) \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} e^{-\alpha_2(\widehat{C})}\sum_{k=0}^{t-1} k(e^{-(2\alpha_2(\widehat{C})+\alpha_1)})^{k-1} \\
&= D_2(K)\kappa(P)^2\beta_1\left(\frac{1}{e^{\alpha_1}-1}\right) \max_{\widehat{C}\in\mathcal{B}_{\widehat{q}}(f(\theta))} \frac{e^{-\alpha_2(\widehat{C})}}{(1-e^{-(2\alpha_2(\widehat{C})+\alpha_1)})^2},
\end{aligned}$$

thus completing the proof. ■

C.8.5 Stochastic Continuous-Time Regret Analysis

Theorem C.8.7. *[Stochastic, continuous-time] Let* $J(K,C) := \mathbb{E}\left[\int_0^\infty (x(t)^\top(Q(t)+K^\top R(t)K)x(t))dt\right]$ *with* $w(t)$ *a white noise process with spectral density* Σ *such that* $D_2(K) := ||\Sigma||_{\text{op}}||W||_{\text{op}} < \infty$. *Assume further that* $\mathcal{P}_{\Theta,C}(C \in \mathcal{B}_{\widehat{q}}(f(\Theta))) \geq 1-\alpha$. *Then, under Assumption 4 and Assumption 5,*

$$\mathcal{P}_{\Theta,C}\left(0 \leq \mathcal{R}(\Theta,C) \leq 2L\widehat{q} + \Delta_{\text{dom}}(\Theta,C)\right) \geq 1-\alpha,$$

where L is the Lipschitz constant of $J(K, \widehat{C})$ in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm. Further, if $\widehat{q} < r(C, K^*(C))$, (see continuous-time in Definition 2) $\Delta_{\mathrm{dom}}(\Theta, C) = 0$.

Proof. We again consider any fixed θ and demonstrate the desired properties that $J(K, \widehat{C})$ is non-negative and L-Lipschitz in $\widehat{C} \in \mathcal{B}_{\widehat{q}}(f(\theta))$ under the operator norm for any $K \in \mathcal{K}(\mathcal{B}_{\widehat{q}}(f(\theta)))$, from which Theorem C.8.1 can be invoked to arrive at the desired conclusion. Notice the objective can be reformulated in the standard manner as follows:

$$
\begin{aligned}
&J(K, \widehat{C}) := \mathbb{E}\left[\int_0^\infty (x(t)^\top (Q(t) + K^\top R(t) K) x(t)) dt\right] \\
&= \int_0^\infty \mathbb{E}[(x(t)^\top (Q(t) + K^\top R(t) K) x(t))] dt = \int_0^\infty \mathbb{E}[\mathrm{Tr}(x(t)^\top (Q(t) + K^\top R(t) K) x(t))] dt \\
&= \int_0^\infty \mathbb{E}[\mathrm{Tr}((Q(t) + K^\top R(t) K) x(t) x(t)^\top)]] dt = \int_0^\infty \mathrm{Tr}((Q(t) + K^\top R(t) K) \mathbb{E}[x(t) x(t)^\top]) dt \\
&= \int_0^\infty \mathrm{Tr}((Q(t) + K^\top R(t) K)(\mathbb{E}[x(t)]\mathbb{E}[x(t)]^\top + \mathrm{Var}(x(t)))) dt.
\end{aligned}
$$

We now obtain the expressions for $\mathbb{E}[x(t)]$ and $\mathrm{Var}(x(t))$ using standard results from stochastic differential equations. For a full review on this topic, see [Särkkä and Solin, 2019]:

$$
\begin{aligned}
J(K, \widehat{C}) = &\int_0^\infty x(0)^\top (e^{\widehat{C} W t})^\top (Q(t) + K^\top R(t) K)(e^{\widehat{C} W t}) x(0) dt \\
&+ \int_0^\infty \int_0^t \mathrm{Tr}((Q(t) + K^\top R(t) K) e^{\widehat{C} W k} \Sigma e^{\widehat{C} W k^\top}) dk dt
\end{aligned}
$$

J is clearly non-negative by construction. We demonstrate this quantity is Lipschitz continuous with an appropriate Lipschitz constant, again by bounding the gradient in much the same manner as the above bound. The bound for the first term was demonstrated in the proof of Theorem C.8.2, which holds under the finiteness assumption of $D_1(K)$. We, thus, solely present that of the second term as follows:

$$
\begin{aligned}
&\int_0^\infty \int_0^t \nabla_{\widehat{C}} \mathrm{Tr}((Q(t) + K^\top R(t) K) e^{(\widehat{C} W) k} \Sigma e^{(\widehat{C} W) k^\top}) dk dt \\
&= \int_0^\infty \int_0^t k(((Q(t)^\top + (R(t) K)^\top K) e^{k \widehat{C} W} \Sigma^\top) \odot e^{k \widehat{C} W}) W^\top dk dt \\
&+ \int_0^\infty \int_0^t k(((Q(t) + K^\top R(t) K) e^{k \widehat{C} W} \Sigma) \odot e^{k \widehat{C} W}) W^\top
\end{aligned}
$$

We now bound each of these two terms separately, although the structure of the two is the same, so we explicitly show steps for bounding the first, from which the same can be repeated on the

second.

$$
\begin{aligned}
L &\leq \max_{\widehat{C}} || \int_0^\infty \int_0^t k(((Q(t)^\top + (R(t)K)^\top K) e^{k\widehat{C}W}\Sigma^\top) \odot e^{k\widehat{C}W}) W^\top dk dt||_{\text{op}} \\
&\leq \max_{\widehat{C}} \int_0^\infty \int_0^t k||Q(t) + K^\top R(t) K||_{\text{op}} ||\Sigma||_{\text{op}} ||W||_{\text{op}} ||e^{k\widehat{C}W}||_{\text{op}}^2 dk dt
\end{aligned}
$$

We again now collect all terms independent of t into a constant $D_2(K) = ||\Sigma||_{\text{op}}||W||_{\text{op}}$, leaving

$$
\begin{aligned}
&\leq \max_{\widehat{C}} D_2(K) \int_0^\infty \beta_1 e^{-\alpha_1 t} \int_0^t k\beta_2(\widehat{C})^2 e^{-2\alpha_2(\widehat{C})k} dk dt \\
&= \max_{\widehat{C}} \frac{D_2(K)\beta_1\beta_2(\widehat{C})^2}{4\alpha_2^2(\widehat{C})} \int_0^\infty e^{-\alpha_1 t} \left(1 - 2\alpha_2(\widehat{C}) t e^{-2\alpha_2(\widehat{C})t} - e^{-2\alpha_2(\widehat{C})t}\right) dt \\
&= \max_{\widehat{C}} \frac{D_2(K)\beta_1\beta_2(\widehat{C})^2}{4\alpha_2^2(\widehat{C})} \left(\frac{1}{\alpha_1} - \frac{2\alpha_2(\widehat{C})}{(\alpha_1 + 2\alpha_2(\widehat{C}))^2} - \frac{1}{\alpha_1 + 2\alpha_2(\widehat{C})}\right) = \max_{\widehat{C}} \frac{D_2(K)\beta_1\beta_2(\widehat{C})^2}{\alpha_1(\alpha_1 + 2\alpha_2(\widehat{C}))^2}
\end{aligned}
$$

We, therefore, again have the desired upper bound on the Lipschitz constant, as desired. ■

C.8.6 Unimodal Assumption Explanation

In classical engineering design, one would prescribe the dynamics of the system by explicitly writing out the physics of the system; this is so universally done that it may not even feel like an assumption in engineering design. This is the sense in which we mean that the design parameters commonly have some "unimodal" (often Dirac) measure in mapping to the system dynamics. For instance, if one is studying a cart pole system with a position x and angle θ, a common characterization (see Chapter 3 of [Tedrake, 2023]) is given by:

$$
\begin{aligned}
\ddot{x} =& \frac{1}{m_c + m_p \sin^2\theta} \left[f_x + m_p \sin\theta (l\dot{\theta}^2 + g\cos\theta)\right] \\
\ddot{\theta} =& \frac{1}{l(m_c + m_p \sin^2\theta)} \left[-f_x \cos\theta - m_p l\dot{\theta}^2 \cos\theta \sin\theta - (m_c + m_p) g \sin\theta\right]
\end{aligned}
$$

Here, the parameters m_c, m_p, and l could all be viewed as the "design parameters" θ depending on what one, as an engineer, has control over. Knowing that the underlying reality can be described by single set of dynamics is, therefore, what motivates using a unimodal model, as we wished to highlight with the models used in previous UCCD works.

C.8.7 Coverage Guarantees Under Noisy Observations

Theorem C.8.8. *Let $\widetilde{C} = C + \epsilon$ where vec$(\epsilon) \sim \mathcal{N}(0, \Sigma)$, where $\epsilon \perp\!\!\!\perp (\Theta, C)$. Assume $\mathcal{U}(\theta) = \{C' \mid ||f(\theta) - C'||_{\text{op}} \leq \widehat{q}\}$ satisfies $\mathcal{P}_{\Theta,\widetilde{C}}(\widetilde{C} \in \mathcal{U}(\Theta)) \geq 1 - \alpha$, where $||\cdot||_{\text{op}}$ denotes the matrix operator norm. If for any $\theta \in \Theta$ and $\delta > 0$, $\mathcal{P}(\widehat{q}^2 - \delta \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta) > \mathcal{P}(\widehat{q}^2 \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta)$, then*

$$\mathcal{P}_{\Theta,C}(C \in \mathcal{U}(\Theta)) \geq \mathcal{P}_{\Theta,\widetilde{C}}(\widetilde{C} \in \mathcal{U}(\Theta)) \geq 1 - \alpha.$$

Proof. Given that $\mathcal{P}_{\Theta,\widetilde{C}}(\widetilde{C} \in \mathcal{U}(\Theta)) \geq 1 - \alpha$, it suffices to show that $\mathcal{P}(C \in \mathcal{U}(\theta) \mid \Theta = \theta) \geq \mathcal{P}(\widetilde{C} \in \mathcal{U}(\theta) \mid \Theta = \theta)$ for all θ, as the conclusion can be drawn by the law of total probability:

$$\begin{aligned}
&\mathcal{P}(\widetilde{C} \in \mathcal{U}(\Theta) \mid \Theta = \theta) \\
&= \mathcal{P}\left(\left\|\widetilde{C} - f(\theta)\right\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&= \mathcal{P}\left(\sup_{\|x\|=1} \left\{\|Cx + \epsilon x - f(\theta)x\|_2^2\right\} \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&= \mathcal{P}\left(\sup_{\|x\|=1} \left\{\|Cx - f(\theta)x\|_2^2 + 2x^T\epsilon^T\left(Cx - f(\theta)x\right) + \|\epsilon x\|_2^2\right\} \leq \widehat{q}^2 \mid \Theta = \theta\right)
\end{aligned}$$

We now *lower* bound this inner quantity, from which

$$\begin{aligned}
&\sup_{\|x\|=1} \left\{\|Cx - f(\theta)x\|_2^2 + 2x^T\epsilon^T\left(Cx - f(\theta)x\right) + \|\epsilon x\|_2^2\right\} \\
&\qquad \geq \sup_{\|x\|=1} \left\{\|Cx - f(\theta)x\|_2^2 + 2x^T\epsilon^T\left(Cx - f(\theta)x\right)\right\} \\
&\qquad \geq \|Cx' - f(\theta)x'\|_2^2 + 2(x')^T\epsilon^T\left(Cx' - f(\theta)(x')\right),
\end{aligned}$$

for any choice of $x' : ||x'||_2 = 1$. The second line follows by the trivial fact that $\|\epsilon x\|_2^2 \geq 0$ and the third from the fact that the previous line is the supremum of *all* such possible values x'. We now specifically take $x' := \arg\max_x \|Cx - f(\theta)x\|_2^2$ and denote it as x^*. From here, we arrive at the final bound

$$\begin{aligned}
&\sup_{\|x\|=1} \left\{\|Cx - f(\theta)x\|_2^2 + 2x^T\epsilon^T\left(Cx - f(\theta)x\right) + \|\epsilon x\|_2^2\right\} \\
&\qquad \geq \|(C - f(\theta))(x^*)\|_2^2 + 2(x^*)^T\epsilon^T\left(Cx^* - f(\theta)(x^*)\right) \\
&\qquad =: \|C - f(\theta)\|_{\text{op}}^2 + 2(x^*)^T\epsilon^T\left(Cx^* - f(\theta)(x^*)\right)
\end{aligned}$$

Since this is a *lower* bound on the original quantity of interest, we have that the probability this

quantity is upper bounded by $\widehat{q}^2$ is *greater* than that of the original quantity being upper bounded. That is,

$$\begin{aligned}
&\mathcal{P}\left(\sup_{\|x\|=1}\left\{\|Cx - f(\theta)x\|_2^2 + 2x^T\epsilon^T\left(Cx - f(\theta)x\right) + \|\epsilon x\|_2^2\right\} \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&\leq \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 + 2x^{*^T}\epsilon^T\left(Cx^* - f(\theta)x^*\right) \leq \widehat{q}^2 \mid \Theta = \theta\right)
\end{aligned}$$

Let $\delta = 2x^{*^T}\epsilon^T\left(Cx^* - f(\theta)x^*\right)$. Then:

$$\begin{aligned}
&:= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 + \delta \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 + \delta \leq \widehat{q}^2 \mid \Theta = \theta, \delta > 0\right)\mathcal{P}(\delta > 0 \mid \Theta = \theta) \\
&\quad + \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 + \delta \leq \widehat{q}^2 \mid \Theta = \theta, \delta \leq 0\right)\mathcal{P}(\delta \leq 0 \mid \Theta = \theta) \\
&= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 - \delta \mid \Theta = \theta, \delta > 0\right)\mathcal{P}(\delta > 0 \mid \Theta = \theta) \\
&\quad + \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta, \delta > 0\right)\mathcal{P}(\delta \leq 0 \mid \Theta = \theta)
\end{aligned}$$

In this final line, we made use of the fact that

$$\mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 - \delta \mid \Theta = \theta, \delta \leq 0\right) = \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta, \delta > 0\right),$$

which follows since the distribution of δ is symmetric about 0 by the symmetry of the distribution of ϵ. In particular, $\delta = f(C, \epsilon)$; since $C \perp\!\!\!\perp \epsilon$, the joint distributions $\mathcal{P}(C, \epsilon)$ and $\mathcal{P}(C, -\epsilon)$ are identical. Thus, the distribution of $\delta' = f(C, -\epsilon)$ matches that of δ. Using this, we add terms that sum to 0 as follows:

$$\begin{aligned}
&= \mathcal{P}\Big(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\Big) \\
&\quad - \mathcal{P}(\delta > 0 \mid \Theta = \theta)\,\mathcal{P}\Big(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta, \delta > 0\Big) \\
&\quad - \mathcal{P}(\delta \leq 0 \mid \Theta = \theta)\,\mathcal{P}\Big(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta, \delta \leq 0\Big) \Big\} = 0 \\
&\quad + \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 - \delta \mid \Theta = \theta, \delta > 0\right)\mathcal{P}(\delta > 0 \mid \Theta = \theta) \\
&\quad + \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta, \delta > 0\right)\mathcal{P}(\delta \leq 0 \mid \Theta = \theta),
\end{aligned}$$

where these newly added terms will be used for manipulation subsequently. From here, we re-express this expression with expectations, where we again use the symmetry in δ in the second

term:

$$
\begin{aligned}
&= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&\quad - \mathcal{P}\left(\delta > 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta) \mid \delta > 0\right] \\
&\quad - \mathcal{P}\left(\delta \leq 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta) \mid \delta > 0\right] \\
&\quad + \mathcal{P}\left(\delta > 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 - \delta \mid \Theta = \theta\right) \mid \delta > 0\right] \\
&\quad + \mathcal{P}\left(\delta \leq 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta\right) \mid \delta > 0\right].
\end{aligned}
$$

With this rewrite, we can group terms and conclude using the stated assumption on the "peaking" structure of the probability in the prediction region:

$$
\begin{aligned}
&= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&\quad - \mathcal{P}\left(\delta > 0 \mid \Theta = \theta\right) \mathbb{E}\big[\mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&\qquad\qquad - \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 - \delta \mid \Theta = \theta\right) \mid \delta > 0\big] \\
&\quad + \mathcal{P}\left(\delta \leq 0 \mid \Theta = \theta\right) \mathbb{E}\big[\mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta\right) \\
&\qquad\qquad - \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \mid \delta > 0\big] \\
&= \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \\
&\quad - \mathcal{P}\left(\delta > 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\widehat{q}^2 - \delta \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \mid \delta > 0\right] \\
&\quad + \mathcal{P}\left(\delta \leq 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\widehat{q}^2 \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta\right) \mid \delta > 0\right].
\end{aligned}
$$

We, therefore, have that $\mathcal{P}(\widetilde{C} \in \mathcal{U}(\theta) \mid \Theta = \theta) \leq \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) + \Delta$, where

$$
\begin{aligned}
\Delta := &\ \mathcal{P}\left(\delta \leq 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\widehat{q}^2 \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta\right) \mid \delta > 0\right] \\
&- \mathcal{P}\left(\delta > 0 \mid \Theta = \theta\right) \mathbb{E}\left[\mathcal{P}\left(\widehat{q}^2 - \delta \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) \mid \delta > 0\right]
\end{aligned}
$$

By the assumption, we know that for all $\delta > 0$:

$$
\mathcal{P}\left(\widehat{q}^2 - \delta \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 \mid \Theta = \theta\right) > \mathcal{P}\left(\widehat{q}^2 \leq \|C - f(\theta)\|_{\text{op}}^2 \leq \widehat{q}^2 + \delta \mid \Theta = \theta\right).
$$

We also know that $\mathcal{P}(\delta \leq 0 \mid \Theta = \theta) \leq \mathcal{P}(\delta \geq 0 \mid \Theta = \theta)$ since

$$\begin{aligned}
\mathcal{P}(\delta \leq 0 \mid \Theta = \theta) &= \mathcal{P}\left(2{x^*}^T \epsilon^T \left(Cx^* - f(\theta)x^*\right) \leq 0 \mid \Theta = \theta\right) \\
&= \mathbb{E}_{C|\Theta=\theta}\left[\mathcal{P}_{\epsilon|C=c,\Theta=\theta}\left({x^*}^T \epsilon^T \left(c - f(\theta)\right) x^* \leq 0\right)\right] \\
&= \mathbb{E}_{C|\Theta=\theta}\left[0.5\right] \\
&= 0.5
\end{aligned}$$

Therefore, $\mathcal{P}\left(\left\|\widetilde{C} - f(\theta)\right\|_{\text{op}} \leq \widehat{q} \mid \Theta = \theta\right) - \mathcal{P}\left(\|C - f(\theta)\|_{\text{op}} \leq \widehat{q} \mid \Theta = \theta\right) \leq \Delta \leq 0$ ■

C.8.8 LQR C Gradient

We follow the presentation of [Fazel et al., 2018] to provide the derivation of $\nabla_C J(K, C)$. Note that the following derivation is given for the discrete-time setting; the continuous-time derivation follows in a similar fashion with a modification in the Lyapunov equations.

Lemma C.8.9. *Let $J(K, C)$ be the infinite horizon, discrete-time, deterministic analog of that defined in Equation* (1.4)*, i.e. $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top (Q + K^\top RK)x_t)$ for $w = 0$. Then,*

$$\nabla_C J(K, C) = 2P_K CWX_K W^T, \tag{C.11}$$

where X_K and P_K respectively solve the following two Lyapunov equations: $\Delta_K X_K \Delta_K^\top - X_K = -X_0$ and $P_K = \Delta_K^\top P_K \Delta_K + Q + K^\top RK$, where $\Delta_K := A - BK$.

Proof. By the standard reformulation of $J(K, C)$ as described in [Fazel et al., 2018], we can rewrite $J(K, C, x_0) = x_0^\top P_K x_0$, where we now make the notational change to make explicit the dependence on x_0, as it pertains to the derivation below. We then have that

$$\begin{aligned}
J(K, C, x_0) &= x_0^\top \Delta_K^\top P_K \Delta_K x_0 + x_0^\top (Q + K^\top RK)x_0 \\
&= J(K, C, \Delta_K x_0) + x_0^\top (Q + K^\top RK)x_0.
\end{aligned}$$

From here, we have that

$$\begin{aligned}\nabla_C J(K,C,x_0) &= \nabla_C J(K,C,\Delta_K x_0) + \nabla_C(x_0^\top (Q+K^\top RK)x_0)^{\to 0} \\ &= 2P_K CWx_0x_0^TW^T + \nabla_C J(K,C,\Delta_K x_1)|_{x_1:=(A-BK)x_0} \\ &= ... \\ &= 2P_K CW(\sum_{t=0}^{\infty} x_t x_t^T)W^T \\ &= 2P_K CWX_K W^T,\end{aligned}$$

where the final equality follows from the well-known correspondence between this infinite sum and the aforementioned Lyapunov reformulation. ■

C.8.9 Policy Gradient Convergence Guarantee

Lemma C.8.10. *Suppose $f(x,y)$ is $c(y)$-gradient dominated for any $y \in \mathcal{Y}$, i.e. for any fixed y, there is a $c(y)$ such that, for any $x \in \mathcal{X}$ and $x^*(y) := \arg\min_{x\in\mathcal{X}} f(x,y)$:*

$$f(x,y) - f(x^*(y),y) \leq c(y)||\nabla_x f(x,y)||_F^2.$$

Further, let $\phi(x) := \max_{y\in\mathcal{Y}} f(x,y)$ and $x^* := \arg\min_{x\in\mathcal{X}} \phi(x)$. Assume that $\operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y) \neq \emptyset\ \forall x$. Then,

$$\phi(x) - \phi(x^*) \leq c^*(x) \min_{y^*\in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)} ||\nabla_x f(x,y^*)||_F^2,$$

where $c^*(x) := \sup_{y^*\in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)} c(y^*)$.

Proof. Note that, for any maximizer $y^* \in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)$, we have

$$\begin{aligned}\phi(x) - \phi(x^*) &:= \max_{y\in\mathcal{Y}} f(x,y) - \max_{y\in\mathcal{Y}} f(x^*,y) = f(x,y^*) - \max_{y\in\mathcal{Y}} f(x^*,y) \\ &\leq f(x,y^*) - f(x^*,y^*) \leq f(x,y^*) - f(x^*(y^*),y^*) \leq c(y^*)||\nabla_x f(x,y^*)||_F^2\end{aligned}$$

Given that this relationship holds for all such y^*, we immediately get that

$$\phi(x) - \phi(x^*) \leq \min_{y^*\in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)} c(y^*)||\nabla_x f(x,y^*)||_F^2 \leq c^*(x) \min_{y^*\in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)} ||\nabla_x f(x,y^*)||_F^2,$$

where $c^*(x) := \sup_{y^*\in \operatorname{Arg\,max}_{y\in\mathcal{Y}} f(x,y)} c(y^*)$. ■

We now make use of the known fact that $J(K, C)$ is gradient-dominated for any fixed C, in turn satisfying the conditions of Theorem C.8.10, from which we reach the desired conclusion. The former fact was demonstrated in [Bu et al., 2019], which we present below for sake of convenience with modification of notational conventions to match that used herein.

Lemma C.8.11. *(Lemma 5 of [Fazel et al., 2018]) Let $J(K, C)$ be the infinite horizon, discrete-time, deterministic analog of that defined in Equation* (1.4)*, i.e. $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top(Q + K^\top RK)x_t)$ for $w = 0$. Then, if $X_K \succcurlyeq 0$ and $K \in \mathcal{K}(C)$,*

$$J(K, C) - J(K^*(C), C) \leq \frac{||X_{K^*(C)}||}{\sigma_{\min}(X_0)^2\sigma_{\min}(R)}||\nabla_K J(K, C)||_F^2, \tag{C.12}$$

where $X_0 \succ 0$ and X_K and P_K respectively solve the following two equations: $\Delta_K X_K \Delta_K^\top - X_K = -X_0$ and $P_K = \Delta_K^\top P_K \Delta_K + Q + K^\top RK$, where $\Delta_K := A - BK$.

Lemma C.8.12. *Let $\phi(K) := \max_{\widehat{C}\in\mathcal{C}} J(K, \widehat{C})$ and $K^*_{\text{rob}}(\mathcal{C}) := \arg\min_{K\in\mathcal{K}(\mathcal{C})} \phi(K)$, with J the infinite horizon, discrete-time, deterministic analog of that defined in Equation* (1.4)*, i.e. $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top(Q + K^\top RK)x_t)$ for $w = 0$. Then, for $K \in \mathcal{K}(\mathcal{C})$ where $X_K \succcurlyeq 0$ for all $\widehat{C} \in \mathcal{C}$, $\phi(K)$ satisfies*

$$\phi(K) - \phi(K^*_{\text{rob}}(\mathcal{C})) \leq \mu^*(K) \min_{C^*\in\text{Arg}\max_{\widehat{C}\in\mathcal{C}} J(K,\widehat{C})} ||\nabla_K J(K, C^*)||_F^2$$

for $\mu^*(K) := \sup_{C^*\in\text{Arg}\max_{\widehat{C}\in\mathcal{C}} J(K,\widehat{C})} \frac{||X_{K^*(C^*)}||}{\sigma_{\min}(X_0)^2\sigma_{\min}(R)}$, where $X_0 \succ 0$ and X_K and P_K respectively solve the following two equations: $\Delta_K X_K \Delta_K^\top - X_K = -X_0$ and $P_K = \Delta_K^\top P_K \Delta_K + Q + K^\top RK$, where $\Delta_K := A - BK$.

Proof. The proof for this follows immediately by demonstrating the assumption of Theorem C.8.10 is satisfied by Theorem C.8.11. ■

Theorem C.8.13. *Let $\phi(K) := \max_{C\in\mathcal{C}} J(K, C)$ and $K^*_{\text{rob}}(\mathcal{C}) := \arg\min_{K\in\mathcal{K}(\mathcal{C})} \phi(K)$, with J the infinite horizon, discrete-time, deterministic analog of that defined in Equation* (1.4)*, i.e. $J(K, C) := \sum_{t=0}^{\infty}(x_t^\top(Q + K^\top RK)x_t)$ for $w = 0$. Let $K^{(t)}$ be the t-th iterate of Algorithm 6. Assume for each iterate t, the optimization over C converges, i.e. $C^{(T_C)} = C^*(K^{(t)})$, that $K^{(t)} \in \mathcal{K}(\mathcal{C})$, and that $X_K \succcurlyeq 0$ for all $\widehat{C} \in \mathcal{C}$ and $K \in \mathcal{K}(\mathcal{C})$. Denote $\nu := \min_{\widehat{C}\in\mathcal{C}} \min_{K\in\mathcal{K}(\mathcal{C})} \sigma_{\min}(X_K)$. If in Algorithm 6*

$$\eta_K \leq \min_{[\widehat{A},\widehat{B}]:=\widehat{C}\in\mathcal{C}} \frac{1}{16}\min\Big\{\left(\frac{\sigma_{\min}(Q)\nu}{J(K,\widehat{C})}\right)^2 \frac{1}{||\widehat{B}||||\nabla_K J(K,\widehat{C})||(1 + ||\widehat{A} - \widehat{B}K||)}, \frac{\sigma_{\min}(Q)}{2J(K,\widehat{C})||R + \widehat{B}^\top P_K \widehat{B}||}\Big\}. \tag{C.13}$$

then, there exists a $\gamma > 0$ *such that* $\phi(K^{(T)}) - \phi(K^*_{\text{rob}}(\mathcal{C})) \leq (1-\gamma)^T(\phi(K_0) - \phi(K^*_{\text{rob}}(\mathcal{C})))$.

Proof. We follow the proof strategy developed in [Fazel et al., 2018], specifically in their presentation of Lemma 24, in which we leverage the above developed gradient dominance result, namely that in Theorem C.8.12. We first note that Algorithm 6 is equivalent to performing subgradient descent over $\phi(K)$ if we assume convergence of the inner maximization over C, that is if $C^{(T_C)} = C^*(K^{(t)})$ for some $C^* \in \text{Arg}\max_{\widehat{C}\in\mathcal{C}} J(K^{(t)}, \widehat{C})$. It, therefore, suffices to characterize subgradient descent, where $K^{(t+1)} := K^{(t)} - \eta_K g_t$.

To complete this proof, it suffices to demonstrate $\phi(K^{(t)}) - \phi(K^{(t+1)}) \geq \gamma(K^{(t)})||g_t||_F^2$ for some $\gamma(K^{(t)}) > 0$, since this along with subgradient dominance can be used to establish the desired convergence guarantees by first demonstrating this intermediate result:

$$\begin{aligned}\phi(K^{(t+1)}) - \phi(K^*_{\text{rob}}(\mathcal{C})) &= (\phi(K^{(t+1)}) - \phi(K^{(t)})) + (\phi(K^{(t)}) - \phi(K^*_{\text{rob}}(\mathcal{C})))\\ &\leq -\gamma(K^{(t)})||g_t||_F^2 + (\phi(K^{(t)}) - \phi(K^*_{\text{rob}}(\mathcal{C})))\\ &\leq \left(1 - \gamma(K^{(t)})/\mu^*(K^{(t)})\right)(\phi(K^{(t)}) - \phi(K^*_{\text{rob}}(\mathcal{C}))).\end{aligned}$$

To then demonstrate the final convergence, we can simply apply this result inductively as follows:

$$\begin{aligned}&\phi(K^{(t)}) - \phi(K^*_{\text{rob}}(\mathcal{C})) \leq \left(1 - \gamma(K^{(T)})/\mu^*(K^{(T)})\right)(\phi(K^{(T-1)}) - \phi(K^*_{\text{rob}}(\mathcal{C})))\\ &\leq \left(1 - \gamma(K^{(T)})/\mu^*(K^{(T)})\right)\left(1 - \gamma(K^{(T-1)})/\mu^*(K^{(T-1)})\right)(\phi(K^{(T-2)}) - \phi(K^*_{\text{rob}}(\mathcal{C})))\\ &\leq \ldots \leq (\phi(K_0) - \phi(K^*_{\text{rob}}(\mathcal{C})))\prod_{t=1}^{T}\left(1 - \gamma(K^{(t)})/\mu^*(K^{(t)})\right) \leq (1-\gamma)^T(\phi(K_0) - \phi(K^*_{\text{rob}}(\mathcal{C}))),\end{aligned}$$

where we take $\gamma := \min_t \gamma(K^{(t)})/\mu^*(K^{(t)})$. We now prove $\phi(K^{(t)}) - \phi(K^{(t+1)}) \geq \gamma(K^{(t)})||g_t||_F^2$. To do so, we leverage the result of combining Lemmas 3 and 24 from [Fazel et al., 2018], by which it was demonstrated that for any fixed dynamics C, there is a $\beta(C) > 0$ such that $J(K,C) - J(K',C) \geq \beta(C)||\nabla_K J(K,C)||_F^2$ if $K, K' \in \mathcal{K}(C)$, $X_K \succcurlyeq 0$, and if η satisfies:

$$\eta \leq \frac{1}{16}\min\left((\frac{\sigma_{\min}(Q)\nu(C)}{J(K,C)})^2\frac{1}{||B||||\nabla_K J(K,C)||(1+||A-BK||)}, \frac{\sigma_{\min}(Q)}{2J(K,C)||R+B^\top P_K B||}\right),$$

where $\nu(C) := \min_{K\in\mathcal{K}(C)}\sigma_{\min}(X_K)$. The stability assumption is satisfied in assuming all iterates $K^{(t)} \in \mathcal{K}(\mathcal{C})$, as $\mathcal{K}(\mathcal{C}) \subset \mathcal{K}(C)$. $X_K \succcurlyeq 0$ is similarly true under the assumption that this property holds for all optimization iterates. The assumption on the learning rate is guaranteed for any $C \in \mathcal{C}$ under the assumption of Equation (C.13). To leverage this result, we must, therefore, re-express

the quantity of interest into an expression with fixed dynamics:

$$\begin{aligned}\phi(K^{(t)}) - \phi(K^{(t+1)}) &:= J(K^{(t)}, C^*(K^{(t)})) - J(K^{(t+1)}, C^*(K^{(t+1)})) \\ &\geq J(K^{(t)}, C^*(K^{(t)})) - J(K^{(t+1)}, C^*(K^{(t)})) \\ &\geq \beta(C^*(K^{(t)}))||\nabla_K J(K^{(t)}, C^*(K^{(t)}))||_F^2 \\ &= \beta(C^*(K^{(t)}))||g_t||_F^2.\end{aligned}$$

Thus, taking $\gamma(K^{(t)}) := \beta(C^*(K^{(t)}))$ satisfies the desired property and completes the proof. ■

C.8.10 Experimental Controls Setup

As discussed in Section 4.2.3, the standard approach to "robustness via multiplicative noise" is non-data-driven specification of the perturbations anticipated upon deployment. They all, however, share the same standard structure of Equation (4.6), with differences being in the specification of the collection $\{\delta_i\}_{i=1}^p, \{\gamma_i\}_{i=1}^q, \{A_i\}_{i=1}^p$, and $\{B_i\}_{i=1}^q$, where $p = q = 2$ is used across experiments. We consider two strategies for the specification of $(\{A_i\}, \{B_i\})$ and three for that of $(\{\delta_i\}, \{\gamma_i\})$. For the former:

- **Random**
  - $A_i[j,k] \sim \mathcal{N}(0,1)$
  - $B_i[j,k] \sim \mathcal{N}(0,1)$
- **Random Row-Col**
  - $A_i[j,:] = A_i[:,k] = 1$ for $j,k \sim \text{Unif}([n])$
  - $B_i[j,:] = B_i[:,k] = 1$ for $j \sim \text{Unif}([n]), k \sim \text{Unif}([m])$

For the latter, the general strategy is to find those $\{\delta_i\}_{i=1}^p, \{\gamma_i\}_{i=1}^q$ that result in unstable dynamics when paired with the corresponding $(\{A_i\}, \{B_i\})$ for some choice of controller, which varies across the strategies considered. This in turn defines a problem such that, within some radius of misspecified dynamics that retain stability, the controller still performs well. These methods proceed by initializing $\delta_i^{(0)} = \gamma_i^{(0)} = \mathbf{1}$ and iteratively multiplicatively increasing each by some pre-defined factor ρ such that $\delta_i^{(t)} = \rho\delta_i^{(t-1)}$ and similarly for $\gamma^{(t)}$ until

$J(A, B, \{A_i\}, \{B_i\}, \{\delta_i^*\}, \{\gamma_i^*\}, K) = \infty$ in

$$J(A, B, \{A_i\}, \{B_i\}, \{\delta_i\}, \{\gamma_i\}, K) := \int_0^\infty (x^\top Q x + (Kx)^\top R(Kx))dt \tag{C.14}$$
$$\text{s.t.} \quad \dot{x} = \left((A + \sum_{i=1}^{p} \delta_i A_i) - (B + \sum_{i=1}^{q} \gamma_i B_i)K \right) x.$$

The problem specifications, therefore, vary in the K used as the stopping criterion of Equation (C.14) and whether $\{\delta_i^*\}, \{\gamma_i^*\}$ are modified in the final specification as follows:

- **Critical**: Consider $K^{(t)} := \arg\min_K J(A, B, \{A_i\}, \{B_i\}, \{\delta_i^{(t)}\}, \{\gamma_i^{(t)}\}, K)$ in each iterate; Take $\{\delta_i := \delta_i^*\}, \{\gamma_i := \gamma_i^*\}$
- **Open-Loop Mean-Square Stable (Weak)**: Consider $K := \mathbf{0}$; Take $\{\delta_i := \nu\delta_i^*\}, \{\gamma_i := \nu\gamma_i^*\}$ for some $\nu \in (0, 1)$
- **Open-Loop Mean-Square Unstable**: Consider $K := \mathbf{0}$; Take $\{\delta_i := \delta_i^*\}, \{\gamma_i := \gamma_i^*\}$

All prediction models $\widehat{f} : \Theta \to (A, B)$ were multi-layer perceptrons implemented in PyTorch [Paszke et al., 2019] with optimization done using Adam [Kingma and Ba, 2014] with a learning rate of 10^{-3} over 1,000 training steps. Training such models required roughly 10 minutes using an Nvidia RTX 2080 Ti GPU for each experimental setup. Running the robust control optimization algorithm took roughly one hour for 1,000 design trials.

C.8.11 Experimental Dynamical Systems Setup

We consider the following dynamical systems in the experiments. Note that parameters were drawn from normal distributions centered on the nominally reported values from the respective papers these dynamics were considered from.

C.8.11.1 Aircraft Control

We consider the experimental setup studied in [Chrif and Kadda, 2014], in which optimal control is sought on the deflection angles of an aircraft. In particular, we assume the dynamics are given by the following:

$$A = \begin{bmatrix} \gamma_\beta & \gamma_p & \gamma_r & 1 \\ L_\beta & L_p & L_r & 0 \\ N_\beta & N_p & N_r & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} \gamma_{\delta_r} & \gamma_{\delta_a} \\ L_{\delta_r} & L_{\delta_a} \\ N_{\delta_r} & N_{\delta_a} \\ 0 & 0 \end{bmatrix}, \qquad \theta := [\gamma, L, N] \in \mathbb{R}^{15}$$

The parameter sampling distributions are given in Table C.5.

Table C.5: Sampling of parameters for aircraft control task.

Parameter	Symbols	Distribution	Hyperparameter Sampling
γ	$\gamma_\beta, \gamma_p, \gamma_r, \gamma_{\delta_r}, \gamma_{\delta_a}$	$\mathcal{N}(\mu_\gamma, \Sigma_\gamma)$	$\mu_\gamma \sim \mathcal{U}([0,1]^5), \Sigma_\gamma = AA^\top,\ A \sim \mathcal{U}([0,1]^{5\times5})$
L	$L_\beta, L_p, L_r, L_{\delta_r}, L_{\delta_a}$	$\mathcal{N}(\mu_L, \Sigma_L)$	$\mu_L \sim \mathcal{U}([0,1]^5), \Sigma_L = AA^\top,\ A \sim \mathcal{U}([0,1]^{5\times5})$
N	$N_\beta, N_p, N_r, N_{\delta_r}, N_{\delta_a}$	$\mathcal{N}(\mu_N, \Sigma_N)$	$\mu_N \sim \mathcal{U}([0,1]^5), \Sigma_N = AA^\top,\ A \sim \mathcal{U}([0,1]^{5\times5})$

C.8.11.2 Load Positioning Control

We consider the load-positioning system of [Ahmadi et al., 2023, Jiang et al., 2016]. In this case, the dynamics are given by:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{d_L}{m_L} - \frac{d_L}{m_B} & \frac{k_B}{m_B} & \frac{d_B}{m_B} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{d_L}{m_B} & -\frac{k_B}{m_B} & -\frac{d_B}{m_B} \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ \frac{1}{m_L} + \frac{1}{m_B} \\ 0 \\ -\frac{1}{m_B} \end{bmatrix}, \qquad \theta := [m_B, m_L, d_L, k_B, d_B] \in \mathbb{R}^5$$

The parameter sampling distributions are given in Table C.6.

Table C.6: Sampling of parameters for load positioning task.

Parameter	Symbol	Distribution	Hyperparameter Sampling
Mass of body	m_B	$m_B = 1/u$	$u \sim \mathcal{U}(0.04,\ 0.0667)$
Mass of load	m_L	$m_L = 1/u$	$u \sim \mathcal{U}(0.3333,\ 1.0)$
Stiffness of body	k_B	$k_B = u \cdot m_B$	$u \sim \mathcal{U}(0.4,\ 1.3333)$
Damping of body	d_B	$d_B = u \cdot m_B$	$u \sim \mathcal{U}(0.004,\ 0.0667)$

C.8.11.3 Furuta Pendulum

We also consider the Furuta pendulum dynamical system given in [Arulmozhi and Victorie, 2022], in which the system dynamics were specified by

$$A = \frac{1}{J_T}\begin{bmatrix} 0 & 0 & J_T & 0 \\ 0 & \frac{1}{4}M_p L_p^2 L_r g & -(J_p + \frac{1}{4}m_p L_p^2)D_r & \frac{1}{2}m_p L_p L_r D_p \\ 0 & -\frac{1}{2}m_p L_p g(J_r + m_p L_r^2) & \frac{1}{2}m_p L_p L_r D_r & -(J_r + m_p L_r^2)D_p \end{bmatrix}$$

$$B = \frac{1}{J_T}\begin{bmatrix} 0 \\ 0 \\ J_p + \frac{1}{4}m_p L_p^2 \\ -\frac{1}{2}m_p L_p L_r \end{bmatrix} \qquad \theta := [M_p, m_p, L_p, L_r, J_T, J_p, J_r, D_p, D_r] \in \mathbb{R}^9$$

The parameter sampling distributions are given in Table C.7.

Table C.7: Sampling of parameters for Furuta pendulum task.

Parameter	Symbol	Distribution	Hyperparameter Values
Pendulum mass	M_p	$\lvert\mathcal{N}(\mu_{M_p}, \sigma^2_{M_p})\rvert$	$\mu_{M_p} = 0.024,\ \sigma_{M_p} \sim \mathcal{U}(0,1)$
Rotor mass	m_p	$\lvert\mathcal{N}(\mu_{m_p}, \sigma^2_{m_p})\rvert$	$\mu_{m_p} = 0.095,\ \sigma_{m_p} \sim \mathcal{U}(0,1)$
Pendulum length	L_p	$\lvert\mathcal{N}(\mu_{L_p}, \sigma^2_{L_p})\rvert$	$\mu_{L_p} = 0.129,\ \sigma_{L_p} \sim \mathcal{U}(0,1)$
Rotor length	L_r	$\lvert\mathcal{N}(\mu_{L_r}, \sigma^2_{L_r})\rvert$	$\mu_{L_r} = 0.085,\ \sigma_{L_r} \sim \mathcal{U}(0,1)$
Total inertia	J_T	$\lvert\mathcal{N}(\mu_{J_T}, \sigma^2_{J_T})\rvert$	$\mu_{J_T} = f(\mu_{m_p}, \mu_{L_r}, \mu_{J_r}, \mu_{J_p}), \sigma_{J_T} \sim \mathcal{U}(0,1)$
Pendulum inertia	J_p	$\lvert\mathcal{N}(\mu_{J_p}, \sigma^2_{J_p})\rvert$	$\mu_{J_p} = \frac{M_p L_p^2}{12},\ \sigma_{J_p} \sim \mathcal{U}(0,1)$
Rotor inertia	J_r	$\lvert\mathcal{N}(\mu_{J_r}, \sigma^2_{J_r})\rvert$	$\mu_{J_r} = \frac{m_p L_r^2}{12},\ \sigma_{J_r} \sim \mathcal{U}(0,1)$
Pendulum damping	D_p	$\lvert\mathcal{N}(\mu_{D_p}, \sigma^2_{D_p})\rvert$	$\mu_{D_p} = 0.0005,\ \sigma_{D_p} \sim \mathcal{U}(0,1)$
Rotor damping	D_r	$\lvert\mathcal{N}(\mu_{D_r}, \sigma^2_{D_r})\rvert$	$\mu_{D_r} = 0.0015,\ \sigma_{D_r} \sim \mathcal{U}(0,1)$

C.8.11.4 DC Microgrids

We additionally consider the LQR model of DC microgrids given in [Liu et al., 2023a], in which the system dynamics were specified by

$$A = \begin{bmatrix} \frac{2(-u_0 d - NK_2S)}{V_s d} & 0 & \frac{-2NK_4S}{V_s d} & \frac{-4NK_5S}{V_s d} & \frac{2u_0}{V_s} & 0 & 0 & 0 & 0 \\ 0 & \frac{2(-u_0 d - NK_3S)}{V_s d} & \frac{4NK_4S}{V_s d} & \frac{6NK_5S}{V_s d} & 0 & \frac{2u_0}{V_s} & 0 & 0 & 0 \\ \frac{6NK_2S}{V_s d} & \frac{4NK_3S}{V_s d} & \frac{2(-u_0 d - NK_4S)}{V_s d} & 0 & 0 & 0 & \frac{2u_0}{V_s} & 0 & 0 \\ \frac{-4NK_2S}{V_s d} & \frac{-2NK_3S}{V_s d} & 0 & \frac{2(-u_0 d - NK_5S)}{V_s d} & 0 & 0 & 0 & \frac{2u_0}{V_s} & 0 \\ \frac{u_0}{V_t} & 0 & 0 & 0 & \frac{-u_0}{V_t} & 0 & 0 & 0 & 0 \\ 0 & \frac{u_0}{V_t} & 0 & 0 & 0 & \frac{-u_0}{V_t} & 0 & 0 & 0 \\ 0 & 0 & \frac{u_0}{V_t} & 0 & & 0 & \frac{-u_0}{V_t} & 0 & 0 \\ 0 & 0 & 0 & \frac{u_0}{V_t} & 0 & 0 & 0 & \frac{-u_0}{V_t} & 0 \\ \frac{NRT}{FC_2^c} & \frac{-NRT}{FC_3^c} & \frac{NRT}{FC_4^c} & \frac{NRT}{FC_5^c} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{C_2^t - C_2^c}{V_s/2} \\ \frac{C_3^t - C_3^c}{V_s/2} \\ \frac{C_4^t - C_4^c}{V_s/2} \\ \frac{C_5^t - C_5^c}{V_s/2} \\ \frac{C_2^c - C_2^t}{V_t} \\ \frac{C_3^c - C_3^t}{V_t} \\ \frac{C_4^c - C_4^t}{V_t} \\ \frac{C_5^c - C_5^t}{V_t} \\ 0 \end{bmatrix} \qquad \theta := [V_s, V_t, S, d, N, K_2, K_3, K_4, K_5, C_2^c, C_3^c, C_4^c, C_5^c, C_2^t, C_3^t, C_4^t, C_5^t] \in \mathbb{R}^{17}$$

The parameter sampling distributions are given in Table C.8.

Table C.8: Sampling of parameters for DC microgrids task.

Parameter	Symbol	Distribution	Hyperparameter Values
Source voltage	V_s	$\mathcal{N}(\mu_{V_s}, \sigma^2_{V_s})$	$\mu_{V_s} = 40,\ \sigma_{V_s} = 26.67$
Terminal voltage	V_t	$\mathcal{N}(\mu_{V_t}, \sigma^2_{V_t})$	$\mu_{V_t} = 500,\ \sigma_{V_t} = 333.33$
Surface area	S	$\mathcal{N}(\mu_S, \sigma^2_S)$	$\mu_S = 24,\ \sigma_S = 16.00$
Diffusion coefficient	d	$\mathcal{N}(\mu_d, \sigma^2_d)$	$\mu_d = 1.27 \times 10^{-3},\ \sigma_d = 8.47 \times 10^{-4}$
Number of layers	N	$\mathcal{N}(\mu_N, \sigma^2_N)$	$\mu_N = 37,\ \sigma_N = 24.67$
Reaction rate constant 2	K_2	$\mathcal{N}(\mu_{K_2}, \sigma^2_{K_2})$	$\mu_{K_2} = 8.768 \times 10^{-10},\ \sigma_{K_2} = 5.845 \times 10^{-10}$
Reaction rate constant 3	K_3	$\mathcal{N}(\mu_{K_3}, \sigma^2_{K_3})$	$\mu_{K_3} = 3.222 \times 10^{-10},\ \sigma_{K_3} = 2.148 \times 10^{-10}$
Reaction rate constant 4	K_4	$\mathcal{N}(\mu_{K_4}, \sigma^2_{K_4})$	$\mu_{K_4} = 6.825 \times 10^{-10},\ \sigma_{K_4} = 4.550 \times 10^{-10}$
Reaction rate constant 5	K_5	$\mathcal{N}(\mu_{K_5}, \sigma^2_{K_5})$	$\mu_{K_5} = 5.897 \times 10^{-10},\ \sigma_{K_5} = 3.931 \times 10^{-10}$
Capacitance cell 2 (cathode)	C_2^c	$\mathcal{N}(\mu_{C_2^c}, \sigma^2_{C_2^c})$	$\mu_{C_2^c} = 1.0,\ \sigma_{C_2^c} = 0.667$
Capacitance cell 3 (cathode)	C_3^c	$\mathcal{N}(\mu_{C_3^c}, \sigma^2_{C_3^c})$	$\mu_{C_3^c} = 1.0,\ \sigma_{C_3^c} = 0.667$
Capacitance cell 4 (cathode)	C_4^c	$\mathcal{N}(\mu_{C_4^c}, \sigma^2_{C_4^c})$	$\mu_{C_4^c} = 1.0,\ \sigma_{C_4^c} = 0.667$
Capacitance cell 5 (cathode)	C_5^c	$\mathcal{N}(\mu_{C_5^c}, \sigma^2_{C_5^c})$	$\mu_{C_5^c} = 1.0,\ \sigma_{C_5^c} = 0.667$
Capacitance cell 2 (total)	C_2^t	$\mathcal{N}(\mu_{C_2^t}, \sigma^2_{C_2^t})$	$\mu_{C_2^t} = 1.0,\ \sigma_{C_2^t} = 0.667$
Capacitance cell 3 (total)	C_3^t	$\mathcal{N}(\mu_{C_3^t}, \sigma^2_{C_3^t})$	$\mu_{C_3^t} = 1.0,\ \sigma_{C_3^t} = 0.667$
Capacitance cell 4 (total)	C_4^t	$\mathcal{N}(\mu_{C_4^t}, \sigma^2_{C_4^t})$	$\mu_{C_4^t} = 1.0,\ \sigma_{C_4^t} = 0.667$
Capacitance cell 5 (total)	C_5^t	$\mathcal{N}(\mu_{C_5^t}, \sigma^2_{C_5^t})$	$\mu_{C_5^t} = 1.0,\ \sigma_{C_5^t} = 0.667$

C.8.11.5 Nuclear Plant

We finally consider the terminal sliding-mode control of a nuclear plant from [Kirgni and Wang, 2023], given by:

$$A = \begin{bmatrix} -\frac{\beta}{\Lambda} & \frac{\beta_1}{\Lambda} & \frac{\beta_2}{\Lambda} & \frac{\beta_3}{\Lambda} & \frac{\alpha_f\theta}{\Lambda} & \frac{\alpha_c\theta}{2\Lambda} & -\frac{\sigma_X\theta}{\nu\Sigma_f\Lambda} & 0 \\ \lambda_1 & -\lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \lambda_2 & 0 & -\lambda_2 & 0 & 0 & 0 & 0 & 0 \\ \lambda_3 & 0 & 0 & -\lambda_3 & 0 & 0 & 0 & 0 \\ \frac{\epsilon_f P_0}{\mu_f} & 0 & 0 & 0 & -\frac{\Omega}{\mu_f} & \frac{\Omega}{\mu_f} & 0 & 0 \\ \frac{(1-\epsilon_f)P_0}{\mu_c} & 0 & 0 & 0 & \frac{\Omega}{\mu_c} & \frac{2M+\Omega}{2\mu_c} & 0 & 0 \\ (\gamma_X\Sigma_f - \sigma_X X_0)\phi_0 P_0 & 0 & 0 & 0 & 0 & 0 & -(\lambda_X + \phi_0 P_0\theta) & \lambda_I \\ \gamma_I\Sigma_f\phi_0 P_0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda_I \end{bmatrix}$$

$$B = \begin{bmatrix} -\frac{\theta}{\Lambda} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \theta := \begin{bmatrix} \alpha_c, \alpha_f, \beta, \beta_1, \beta_2, \beta_3, \Lambda, \lambda_I, \lambda_X, \lambda_1, \lambda_2, \lambda_3, \mu_f, \mu_c, \gamma_X, \gamma_I, \sigma_X, \Sigma_f, \\ \nu, \epsilon_f, \Omega, M, \theta, P_0, \phi_0, X_0 \end{bmatrix} \in \mathbb{R}^{26}$$

The parameter sampling distributions are given in Table C.9.

Table C.9: Sampling of parameters for nuclear plant task. Note that some of the parameters, specifically θ, μ_c, ν, Ω, M, ϕ_0, X_0, were not ascribed values in the paper from which the dynamics were provided. These were assumed to be in normalized units for the simulation.

Parameter	Symbol	Distribution	Hyperparameter Values
Coolant reactivity coefficient	α_c	$\mathcal{N}(\mu_{\alpha_c}, \sigma^2_{\alpha_c})$	$\mu = -2.0,\ \sigma = 2.0$
Fuel reactivity coefficient	α_f	$\mathcal{N}(\mu_{\alpha_f}, \sigma^2_{\alpha_f})$	$\mu = -14.0,\ \sigma = 14.0$
Total delayed neutron fraction	β	$\mathcal{N}(\mu_{\beta}, \sigma^2_{\beta})$	$\mu = 0.0065,\ \sigma = 0.0065$
Delayed neutron precursor (group 1)	β_1	$\mathcal{N}(\mu_{\beta_1}, \sigma^2_{\beta_1})$	$\mu = 0.00021,\ \sigma = 0.00021$
Delayed neutron precursor (group 2)	β_2	$\mathcal{N}(\mu_{\beta_2}, \sigma^2_{\beta_2})$	$\mu = 0.00225,\ \sigma = 0.00225$
Delayed neutron precursor (group 3)	β_3	$\mathcal{N}(\mu_{\beta_3}, \sigma^2_{\beta_3})$	$\mu = 0.00404,\ \sigma = 0.00404$
Prompt neutron lifetime	Λ	$\mathcal{N}(\mu_{\Lambda}, \sigma^2_{\Lambda})$	$\mu = 2.1,\ \sigma = 2.1$
Iodine decay constant	λ_I	$\mathcal{N}(\mu_{\lambda_I}, \sigma^2_{\lambda_I})$	$\mu = 10.0,\ \sigma = 10.0$
Xenon decay constant	λ_X	$\mathcal{N}(\mu_{\lambda_X}, \sigma^2_{\lambda_X})$	$\mu = 2.9,\ \sigma = 2.9$
Decay const. neutron prec. group 1	λ_1	$\mathcal{N}(\mu_{\lambda_1}, \sigma^2_{\lambda_1})$	$\mu = 0.0124,\ \sigma = 0.0124$
Decay const. neutron prec. group 2	λ_2	$\mathcal{N}(\mu_{\lambda_2}, \sigma^2_{\lambda_2})$	$\mu = 0.0369,\ \sigma = 0.0369$
Decay const. neutron prec. group 3	λ_3	$\mathcal{N}(\mu_{\lambda_3}, \sigma^2_{\lambda_3})$	$\mu = 0.632,\ \sigma = 0.632$
Fuel heat capacity	μ_f	$\mathcal{N}(\mu_{\mu_f}, \sigma^2_{\mu_f})$	$\mu = 0.0263,\ \sigma = 0.0263$
Coolant heat capacity	μ_c	$\mathcal{N}(\mu_{\mu_c}, \sigma^2_{\mu_c})$	$\mu = 1.0,\ \sigma = 1.0$
Fission yield (xenon)	γ_X	$\mathcal{N}(\mu_{\gamma_X}, \sigma^2_{\gamma_X})$	$\mu = 0.003,\ \sigma = 0.003$
Fission yield (iodine)	γ_I	$\mathcal{N}(\mu_{\gamma_I}, \sigma^2_{\gamma_I})$	$\mu = 0.059,\ \sigma = 0.059$
Xenon absorption cross-section	σ_X	$\mathcal{N}(\mu_{\sigma_X}, \sigma^2_{\sigma_X})$	$\mu = 3.5 \times 10^{-18},\ \sigma = 3.5 \times 10^{-18}$
Fission cross-section	Σ_f	$\mathcal{N}(\mu_{\Sigma_f}, \sigma^2_{\Sigma_f})$	$\mu = 0.3358,\ \sigma = 0.3358$
Neutrons per fission	ν	$\mathcal{N}(\mu_{\nu}, \sigma^2_{\nu})$	$\mu = 1.0,\ \sigma = 1.0$
Power deposition fraction in fuel	ϵ_f	$\mathcal{N}(\mu_{\epsilon_f}, \sigma^2_{\epsilon_f})$	$\mu = 0.92,\ \sigma = 0.92$
Heat transfer coefficient	Ω	$\mathcal{N}(\mu_{\Omega}, \sigma^2_{\Omega})$	$\mu = 1.0,\ \sigma = 1.0$
Coolant mass	M	$\mathcal{N}(\mu_{M}, \sigma^2_{M})$	$\mu = 1.0,\ \sigma = 1.0$
Control reactivity	θ	$\mathcal{N}(\mu_{\theta}, \sigma^2_{\theta})$	$\mu = 1.0,\ \sigma = 1.0$
Nominal reactor power	P_0	$\mathcal{N}(\mu_{P_0}, \sigma^2_{P_0})$	$\mu = 3.0,\ \sigma = \sqrt{3.0}$
Neutron flux	ϕ_0	$\mathcal{N}(\mu_{\phi_0}, \sigma^2_{\phi_0})$	$\mu = 1.0,\ \sigma = 1.0$
Nominal xenon conc.	X_0	$\mathcal{N}(\mu_{X_0}, \sigma^2_{X_0})$	$\mu = 1.0,\ \sigma = 1.0$

C.8.12 Additional Experimental Results

C.8.12.1 Raw Results

We here provide the raw regrets from Section 4.2.4.1 in Table C.10 and the proportion of stabilized dynamics in Table C.11.

Table C.10: Each of the results below are median normalized regrets over 1,000 i.i.d. test samples with median absolute deviations in parentheses. For clarity, we have not reported any cases with $> 80\%$ unstable cases (see Table C.11 for respective percentages).

	Airfoil	Load Positioning	Furuta Pendulum	DC Microgrids	Fusion Plant
Random Critical	—	—	—	—	—
Random OL MSS (Weak)	0.091 (0.045)	—	—	—	—
Random OL MSUS	—	—	—	—	—
Row-Col Critical	—	—	—	—	—
Row-Col OL MSS (Weak)	0.101 (0.063)	—	—	—	—
Row-Col OL MSUS	0.104 (0.066)	—	—	—	—
CPC	**0.085 (0.058)**	**0.033 (0.023)**	**0.002 (0.002)**	**0.000 (0.000)**	**0.011 (0.011)**
Shared Lyapunov	0.349 (0.221)	0.358 (0.255)	0.055 (0.039)	0.000 (0.000)	0.030 (0.027)
Auxiliary Stabilizer	0.322 (0.202)	0.343 (0.256)	0.048 (0.036)	0.000 (0.000)	0.029 (0.027)
$\mathcal{H}_\infty$	0.288 (0.188)	0.087 (0.060)	0.063 (0.045)	0.012 (0.010)	0.035 (0.032)

Table C.11: Percentages of cases with unstable robust control over 1,000 i.i.d. test samples.

	Airfoil	Load Positioning	Furuta Pendulum	DC Microgrids	Fusion Plant
Random Critical	1.000	1.000	1.000	1.000	1.000
Random OL MSS (Weak)	0.783	1.000	0.920	1.000	0.987
Random OL MSUS	0.825	1.000	0.961	1.000	0.990
Row-Col Critical	0.998	1.000	1.000	1.000	1.000
Row-Col OL MSS (Weak)	0.200	1.000	0.948	1.000	0.960
Row-Col OL MSUS	0.210	1.000	0.951	1.000	0.963
CPC	0.088	0.251	0.174	0.009	0.643
Shared Lyapunov	0.093	0.229	0.141	0.008	0.561
Auxiliary Stabilizer	0.087	0.223	0.142	0.007	0.556
$\mathcal{H}_\infty$	0.081	0.236	0.142	0.007	0.570

C.8.12.2 Method Timings

CPC is more computationally expensive than alternatives. This pairs well with the anticipated use cases, namely in engineering design workflows involving UCCD, i.e. where the control problem is solved *offline*.

Table C.12: Comparison of average method timing (to convergence) across tasks as measured over 10 trials for each experimental setup.

Method	**Airfoil**	**Load Position**	**Pendulum**	**Battery**	**Fusion**
$\mathcal{H}_\infty$	0.17	0.14	0.16	0.19	0.23
Shared Lyapunov	2.68	2.63	2.53	1.13	2.12
Auxiliary Stabilizer	1.70	1.75	1.85	1.01	1.41
Random Critical	7.50	1.74	6.78	7.02	10.50
Random OL MSS (Weak)	6.72	1.34	4.94	7.36	6.39
Random OL MSUS	6.63	2.18	7.25	15.21	7.11
Row-Col Critical	5.37	2.32	5.51	5.48	5.44
Row-Col OL MSS (Weak)	4.35	2.34	4.13	1.00	2.88
Row-Col OL MSUS	3.43	1.84	4.77	4.54	3.18
CRC	13.03	13.44	12.47	12.19	10.88

APPENDIX D

Preliminary Results for Future Directions

D.1 Non-Parameteric Conformal Distributionally Robust Optimization

Proof. We consider the event of interest conditionally on a pair $(x, \mathcal{P}_C)$ where $\mathcal{P}_C \in \mathcal{U}(x)$:

$$
\begin{aligned}
&|\inf_{w\in\mathcal{W}} \sup_{q\in\mathcal{U}(x)} \mathbb{E}_q[f(w,C)] - \inf_{w\in\mathcal{W}} \mathbb{E}_{\mathcal{P}_C}[f(w,C)]| \\
&\leq \sup_{w\in\mathcal{W}} |\sup_{q\in\mathcal{U}(x)} \mathbb{E}_q[f(w,C)] - \mathbb{E}_{\mathcal{P}_C}[f(w,C)]| \\
&\leq \sup_{w\in\mathcal{W}} \sup_{q\in\mathcal{U}(x)} |\mathbb{E}_q[f(w,C)] - \mathbb{E}_{\mathcal{P}_C}[f(w,C)]| \\
&\leq \sup_{w\in\mathcal{W}} \sup_{q\in\mathcal{U}(x)} L\mathcal{W}_1(q,\mathcal{P}_C) = L\mathrm{diam}(\mathcal{U}(x)).
\end{aligned}
$$

Since we have the assumption that $\mathcal{P}(C \in \mathcal{U}(X)) \geq 1-\alpha$, the result immediately follows. ■

D.2 Conformally Robust Engineering Design

D.2.1 Background

D.2.1.1 Sobolev Spaces

The study of numerical simulation of PDEs is a mature field. We, therefore, only provide a brief introduction to the topic, referring readers to the book [Brezis and Brézis, 2011] for an excellent treatment of the relevant materials. Differential problems are posited in the form

$$
Du(x) = f(x) \quad x \in \Omega \qquad u(x) = 0 \quad x \in \partial\Omega, \tag{D.1}
$$

where $\Omega \subset \mathbb{R}^d$ is a compact domain, $f, u : \mathbb{R}^d \to \mathbb{R}$ are scalar fields, and D is a differential operator. The existence and uniqueness of a solution u to such a posited PDE can then often be established over a Sobolev space, defined as those functions with bounded Sobolev norm, i.e. $\mathcal{W}^{s,p}(\Omega) := \{u(x) : ||u(x)||_{\mathcal{W}^{s,p}(\Omega)} < \infty\}$, where

$$||u(x)||_{\mathcal{W}^{s,p}(\Omega)} := \sum_{\alpha \in \Lambda_{\leq s}} ||\partial_x^\alpha u(x)||^p_{\mathcal{L}^p(\Omega)} \qquad \text{where} \qquad \Lambda_{\leq s} := \{\alpha \in \mathbb{N}_0^d : ||\alpha||_1 \leq s\}. \tag{D.2}$$

Note that we employ the common condensed notation $\partial_x^\alpha u := \partial_{x_1}^{\alpha_1} ... \partial_{x_d}^{\alpha_d} u$ for $\alpha := (\alpha_1, ..., \alpha_d)$. Since all partials are with respect to x in this manuscript, we condense the notation further and simply denote this operator as ∂^α. Notably, this space assumes a Hilbert structure in the special case of $p = 2$, which we denote as $\mathcal{H}^s(\Omega) := \mathcal{W}^{s,2}(\Omega)$. In the further specialized case of $\Omega = \mathbb{T}^d$, the space $\mathcal{H}^s(\mathbb{T}^d)$ can be defined by the equivalent norm over the function's Fourier spectrum arising from Parseval's identity, namely

$$||u||^2_{\mathcal{H}^s(\mathbb{T}^d)} := \sum_{n \in \mathbb{Z}^d} (1 + ||n||_2^2)^s \langle u, \varphi_n \rangle^2 \quad \text{where} \quad \varphi_n := e^{2\pi i n \cdot x}, \tag{D.3}$$

The notion of "equivalent norms" is the standard definition, where $||\cdot||_a$ and $||\cdot||_b$ are called equivalent if $\exists\, c$ and C such that for any x, $c||x||_b \leq ||x||_a \leq C||x||_b$.

D.2.1.2 Neural Operators

Data-driven approaches to modeling have classically focused on learning maps between finite-dimensional spaces. With the increasing interest in leveraging machine learning in domains such as solving PDEs, however, approaches that learn maps between infinite-dimensional function spaces have emerged. So-called "operator-learning" methods, therefore, seek to learn a map $\mathcal{G} : \mathcal{A} \to \mathcal{U}$ between two function spaces $\mathcal{A}$ and $\mathcal{U}$, where observations $\mathcal{D} := \{(a_i, u_i)\}$ have been made. We assume that there exists some true, deterministic operator $\mathcal{G}$ such that $u = \mathcal{G}(a)$. While many different learning-based approaches have been proposed to solve this learning problem, they all can be abstractly framed as seeking to recover this true map, formally

$$\min_{\widehat{\mathcal{G}}} ||\widehat{\mathcal{G}} - \mathcal{G}||^2_{\mathcal{L}^2(\mathcal{A},\mathcal{U})} := \int_{\mathcal{A}} ||\widehat{\mathcal{G}}(a) - \mathcal{G}(a)||^2_{\mathcal{U}} da. \tag{D.4}$$

While the operator learning task can be framed in this general light, most works studying operator learning methods have focused on the setting of PDEs, where it is of interest to learn the solution operator of a given PDE [Li et al., 2020b,c,a, Bonev et al., 2023].

One such class of approaches are "spectral neural operators" [Fanaskov and Oseledets, 2023,

Wu et al., 2023, Du et al., 2023, Liu et al., 2023b]. Such methods amortize classical spectral methods, which posit that the solution function can be decomposed as $u(x) = \sum_n u_n \varphi_n(x)$ for some pre-specified $\{\varphi_n\}$ basis and solve for the corresponding $\{u_n\}$ coefficients. In this setting, therefore, the generally posed neural operator objective of Equation (D.4) reduces to a simple vector-to-vector regression loss over the spectral representations of a_i and u_i. That is, we assume the existence of spectral decompositions for each, which we denote $\vec{a}^{(i)}, \vec{u}^{(i)} \in \mathbb{R}^{N_{\max}^d}$, and train a map $\mathcal{G} : \mathbb{R}^{N_{\max}^d} \to \mathbb{R}^{N_{\max}^d}$ on the resulting dataset.

D.2.2 Method

We now discuss our proposed methodology for performing robust engineering design under infinite-dimensional conformal prediction regions. In particular, we introduce the calibration procedure in Section D.2.2.1.

D.2.2.1 Spectral Operator Calibration

Throughout this exposition, we assume the PDE surrogate map is learned as a spectral neural operator, as described in Section D.2.1.2. For notational ease, in the discussions that follow, we adopt the convention that the *full* sequence of spectral coefficients of a function u is denoted $\{u_n\}$. This implicitly indexes over $n \in \mathbb{Z}^d$. We distinguish this from truncated spectra with the same vector notation as before, i.e. $\vec{u} := \{u_n\}_{n:|n|_\infty \leq N_{\max}}$. Formally, therefore, the spectral operator is learned as a finite-dimensional map $\mathcal{G} : \mathbb{R}^{N_{\max}^d} \to \mathbb{R}^{N_{\max}^d}$ on a dataset of function spectra $\vec{a}^{(i)}, \vec{u}^{(i)} \in \mathbb{R}^{N_{\max}^d}$.

As discussed, one crucial detail that distinguishes conformal guarantees in functional settings versus those in typical settings is that the outputs in this setting are fundamentally only *partially* observable. That is, functions are not directly observable: only discrete samplings, either as evaluations in the spatial domain or as truncated spectral representations, can be observed. We, however, seek to provide coverage guarantees on the *full* function, i.e. where the spectral expansion is not truncated. To achieve this, we define the following family of score functions

$$s_{N;\nu}(\vec{a}, \vec{u}) := \sum_{n \in \mathbb{Z}^d : |n|_\infty \leq N} (1 + ||n||_2^2)^{s-\nu} ([\mathcal{G}(\vec{a})]_n - \vec{u}_n)^2 \quad \text{where} \quad \nu \in \{1, ..., s\}. \tag{D.5}$$

This choice of score is motivated by its equivalence via Parseval's theorem to the $(s - \nu)$-Sobolev norm residual. Notably, the score function itself is parameterized by two values: the spectral truncation N and the Sobolev norm parameter ν, where $N \leq N_{\max}$. As discussed extensively over later sections, the scoring over variable truncations is leveraged for multi-resolution optimization; importantly, the quantiles over all truncations can be computed efficiently with a single vectorized

operation.

Critical to note is that the score was defined using the $(s-\nu)$-Sobolev norm rather than the more natural choice of the s-Sobolev norm. The lower bound of $\nu \geq 1$ is necessary to guarantee asymptotic coverage of the true function, as we formalize below. Note again that the statement of Equation (D.6) is made directly on the *full* spectrum $\{u'_n\}_{n\in\mathbb{Z}^d}$, not on its finite spectral projection, i.e. not on $\vec{u}'$. Intuitively, the probabilistic bound is achieved by leveraging the conformal quantile to control the behavior of the observed lower-order modes and the smoothness of functions in $\mathcal{H}^s(\mathbb{T}^d)$ to ensure the decay of higher-order modes.

Theorem D.2.1. *Let* $(\vec{a}^{(i)}, \{u_n^{(i)}\}) \cup (\vec{a}', \{u'_n\}) \sim \mathcal{P}(\vec{A}, \{U_n\})$ *and* $\mathcal{P}(||\{U_n\}||_s^2 \leq B) = 1$ *for some* $B \geq 0$. *Further, let* $\mathcal{D}_C := \{(\vec{a}^{(i)}, \vec{u}^{(i)})\}$. *Let* $\alpha \in (0,1)$ *and* $\widehat{q}_{N;\nu}$ *be the* $\lceil (N_C+1)(1-\alpha)\rceil / N_C$ *quantile of Equation* (D.5) *over* $\mathcal{D}_C$ *for a fixed* $\mathcal{G}$ *and* $\nu \in \{1, ..., s\}$. *Then*

$$\mathcal{P}_{\vec{A}', \{U'_n\}} \left(\sum_{n\in\mathbb{Z}^d} (1+||n||_2^2)^{s-\nu} ([\mathcal{G}(\vec{A}')]_n - U'_n)^2 \leq \widehat{q}_{N;\nu} + BN^{-2\nu} \right) \geq 1-\alpha. \tag{D.6}$$

Proof. We consider the event of interest conditionally on $s_{N;\nu}(\vec{a}', \{u'_n\}) \leq \widehat{q}_{N;\nu}$. Then:

$$\begin{aligned}
&\sum_{n\in\mathbb{Z}^d} (1+||n||_2^2)^{s-\nu} ([\mathcal{G}(\vec{a}')]_n - u'_n)^2 \\
&= \underbrace{\sum_{n\in\mathbb{Z}^d : |n|_\infty \leq N} (1+||n||_2^2)^{s-\nu} ([\mathcal{G}(\vec{a}')]_n - u'_n)^2}_{\mathcal{E}_{\leq N}} + \underbrace{\sum_{n\in\mathbb{Z}^d : |n|_\infty > N} (1+||n||_2^2)^{s-\nu} (u'_n)^2}_{\mathcal{E}_{>N}}
\end{aligned}$$

We now demonstrate how each of the above two terms can be bounded. For the former, the result immediately follows in noting that $\mathcal{E}_{\leq N}$ is precisely $s_{N;\nu}(\vec{a}', \vec{u}')$, from which we have $\mathcal{E}_{\leq N} \leq \widehat{q}_{N;\nu}$ directly by assumption on $(\vec{a}', \vec{u}')$. For the latter, we appeal to standard techniques for Fourier truncation analysis as follows

$$\begin{aligned}
\sum_{n\in\mathbb{Z}^d : |n|_\infty > N} (1+||n||_2^2)^{s-\nu} (u'_n)^2 &= \sum_{n\in\mathbb{Z}^d : |n|_\infty > N} \frac{(1+||n||_2^2)^s}{(1+||n||_2^2)^\nu} (u'_n)^2 \\
&\leq \frac{1}{(1+N^2)^\nu} \sum_{n\in\mathbb{Z}^d : |n|_\infty > N} (1+||n||_2^2)^s (u'_n)^2 \leq BN^{-2\nu}.
\end{aligned}$$

We conclude by noting that $\mathcal{P}_{\vec{A}', \{U'_n\}}(s_{N;\nu}(\vec{A}', \{U'_n\}) \leq \widehat{q}_{N;\nu}) \geq 1-\alpha$ from standard results of conformal prediction, completing the proof. ■

The implication of such statements is that the typical conformal quantile retains coverage guarantees on the underlying function if augmented with an additional, finite margin, which decays

with a "more complete" observation of the function, i.e. with larger N. Intuitively, with observation of the full function, i.e. when $N \to \infty$, this margin becomes zero. While such a result is natural, an interesting note is that such a property does *not* hold if we took $\nu = 0$ to define the score in Theorem D.2.1. This result also motivates the choice of $\nu = s$ to achieve a faster decay rate of the margin; however, this needs to be balanced against the increased conservatism of the prediction regions that results from using higher values of ν.

From here, we can immediately make similar coverage claims on a commonly encountered class of elliptic PDEs by appealing to results from classical results of elliptic regularity theory. Briefly, an elliptic PDE operator L is given by

$$L = -\sum_{i,j=1}^{n} a^{ij}(x)\partial_{x_i x_j} + \sum_{i=1}^{n} b^i(x)\partial_{x_i} + c(x), \tag{D.7}$$

and is one for which $\exists\, \theta > 0$ such that $\sum_{i,j=1}^{n} a^{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2$ for a.e. $x \in \mathbb{T}^d$ and $\xi \in \mathbb{R}^n$. The "order" of such an operator specifies its highest total derivative. In particular, we can establish the following corollary; we provide the statement of the relevant result of classical regularity theory from which this result follows. We present below a special case of the statement from [Sturm, 2017] for scalar fields on $\mathbb{T}^d$.

Theorem D.2.2. *(Theorem 6 of [Sturm, 2017]) Let L be an elliptic operator of order ℓ on $\mathbb{T}^d$ such that $Lu = f$. Then, there exists $C(L, s)$ such that for $u \in \mathcal{H}^{s+\ell}(\mathbb{T}^d) \cap \mathrm{Ker}(L)^\perp$, $f \in \mathcal{H}^s(\mathbb{T}^d)$ and $\|u\|_{\mathcal{H}^{s+\ell}(\mathbb{T}^d)} \leq C(L, s)\|f\|_{\mathcal{H}^s(\mathbb{T}^d)}$.*

Corollary D.2.3. *Let L be an elliptic operator of order ℓ on $\mathbb{T}^d$ such that $Lu = f$. Let $s \in \mathbb{N}$ such that $s \geq \ell$. Let $\mathcal{D}_C := \{(\vec{f}^{(i)}, \vec{u}^{(i)})\}$ and $(\vec{f}^{(i)}, \{u_n\}) \cup (\vec{f}', \{u'_n\}) \sim \mathcal{P}(\vec{F}, \{U_n\})$, where $Lu_i = f_i$, and $\alpha \in (0, 1)$, $\mathcal{G}$, N, $\nu \in \{1, ..., s\}$, and $\widehat{q}_{N;\nu}$ be as defined in Theorem D.2.1 with respect to such $\mathcal{D}_C$. Then, there exists $C(L, s)$ such that for $u \in \mathcal{H}^s(\mathbb{T}^d) \cap \mathrm{Ker}(L)^\perp$,*

$$\mathcal{P}_{\vec{F}', \{U'_n\}}\left(\sum_{n \in \mathbb{Z}^d} (1 + ||n||_2^2)^{s-\nu}([\mathcal{G}(\vec{F}')]_n - \widehat{U}'_n)^2 \leq \widehat{q}_{N;\nu} + C(L, s)\|\vec{F}'\|^2_{\mathcal{H}^{s-\ell}(\mathbb{T}^d)} N^{-2\nu}\right) \geq 1 - \alpha$$

Notably, the requirement that $u \in \mathrm{Ker}(L)^\perp$ is to ensure u is a unique solution to the posited PDE. An important special case of Theorem D.2.3 is when $L = \Delta$. In this case, $\ell = 2$, and the requirement that u be a unique solution can be naturally enforced by restricting u to zero-mean fields, i.e. restricting solutions to satisfy $\int_{\mathbb{T}^d} u(x)\, dx = 0$. We denote this margin-padded quantile as $\widehat{q}^*_{N;\nu} := \widehat{q}_{N;\nu} + BN^{-2\nu}$, for which the corresponding prediction region $\mathcal{C}^*_{N;\nu}(\vec{a}) := \mathcal{B}^{||\cdot||_{s-\nu}}_{\widehat{q}^*_{N;\nu}}(\mathcal{G}(\vec{a}))$ has coverage guarantees per Equation (D.6).

BIBLIOGRAPHY

B Abareshi, J Aguilar, S Ahlen, Shadab Alam, David M Alexander, R Alfarsy, L Allen, C Allende Prieto, O Alves, J Ameel, et al. Overview of the instrumentation for the dark energy spectroscopic instrument. *The Astronomical Journal*, 164(5):207, 2022.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Abhinav Agrawal, Daniel R Sheldon, and Justin Domke. Advances in black-box vi: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33:17358–17369, 2020.

Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *arXiv preprint arXiv:1912.12132*, 2019.

Peyman Ahmadi, Mehdi Rahmani, and Aref Shahmansoorian. Lqr based optimal co-design for linear control systems with input and state constraints. *International Journal of Systems Science*, 54(5):1136–1149, 2023.

Christopher T Aksland, Daniel L Clark, Christopher A Lupp, and Andrew G Alleyne. An approach to robust co-design of plant and closed-loop controller. In *2023 IEEE Conference on Control Technology and Applications (CCTA)*, pages 918–925. IEEE, 2023.

Mary B Alatise and Gerhard P Hancke. A review on challenges of autonomous mobile robot and sensor fusion methods. *IEEE Access*, 8:39830–39846, 2020.

James T Allison, Tinghao Guo, and Zhi Han. Co-design of an active suspension using simultaneous dynamic optimization. *Journal of Mechanical Design*, 136(8):081003, 2014.

Luca Ambrogioni, Umut Güçlü, Julia Berezutskaya, Eva Borne, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel Gerven. Forward amortized inference for likelihood-free variational marginalization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 777–786. PMLR, 2019.

Enrico Angelelli, Valentina Morandi, Martin Savelsbergh, and Maria Grazia Speranza. System optimal routing of traffic flows with user constraints using linear programming. *European journal of operational research*, 293(3):863–879, 2021.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.

N Arulmozhi and T Victorie. Kalman filter and h∞ filter based linear quadratic regulator for furuta pendulum. *Computer Systems Science & Engineering*, 43(2), 2022.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Georgy Ayzel, Tobias Scheffer, and Maik Heistermann. Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6): 2631–2644, 2020.

Saeed Azad and Michael J Alexander-Ramos. Robust mdsdo for co-design of stochastic dynamic systems. *Journal of Mechanical design*, 142(1):011403, 2020a.

Saeed Azad and Michael J Alexander-Ramos. A single-loop reliability-based mdsdo formulation for combined design and control optimization of stochastic dynamic systems. *Journal of Mechanical Design*, 143(2):021703, 2020b.

Saeed Azad and Michael J Alexander-Ramos. Robust combined design and control optimization of hybrid-electric vehicles using mdsdo. *IEEE Transactions on Vehicular Technology*, 70(5): 4139–4152, 2021.

Saeed Azad and Daniel R Herber. Control co-design under uncertainties: formulations. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 86229, page V03AT03A008. American Society of Mechanical Engineers, 2022.

Saeed Azad and Daniel R Herber. Concurrent probabilistic control co-design and layout optimization of wave energy converter farms using surrogate modeling. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 87318, page V03BT03A035. American Society of Mechanical Engineers, 2023a.

Saeed Azad and Daniel R Herber. An overview of uncertain control co-design formulations. *Journal of Mechanical Design*, 145(9):091709, 2023b.

Saeed Azad, Mohammad Behtash, Arian Houshmand, and Michael Alexander-Ramos. Comprehensive phev powertrain co-design performance studies using mdsdo. In *Advances in Structural and Multidisciplinary Optimization: Proceedings of the 12th World Congress of Structural and Multidisciplinary Optimization (WCSMO12) 12*, pages 83–97. Springer, 2018.

Saeed Azad, Mohammad Behtash, Arian Houshmand, and Michael J Alexander-Ramos. Phev powertrain co-design with vehicle performance considerations using mdsdo. *Structural and Multidisciplinary Optimization*, 60:1155–1169, 2019.

Saeed Azad, Daniel R Herber, Suraj Khanal, and Gaofeng Jia. Site-dependent solutions of wave energy converter farms with surrogate models, control co-design, and layout optimization. *arXiv preprint arXiv:2405.06794*, 2024.

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Mohammad Behtash and Michael J Alexander-Ramos. Decomposition-based mdsdo for co-design of large-scale dynamic systems. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 51753, page V02AT03A003. American Society of Mechanical Engineers, 2018.

Mohammad Behtash and Michael J Alexander-Ramos. A comparative study between the generalized polynomial chaos expansion-and first-order reliability method-based formulations of simulation-based control co-design. *Journal of Mechanical Design*, pages 1–17, 2024.

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *arXiv preprint arXiv:2305.10124*, 2023.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

Petar Bevanda, Max Beier, Shahab Heshmati-Alamdari, Stefan Sosnowski, and Sandra Hirche. Towards data-driven lqr with koopmanizing flows. *IFAC-PapersOnLine*, 55(15):13–18, 2022.

Hassan Bevrani, Mohammad Ramin Feizi, and Sirwan Ataee. Robust frequency control in an islanded microgrid: h_∞ and μ-synthesis approaches. *IEEE transactions on smart grid*, 7(2): 706–717, 2015.

Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

Trevor J Bird, Jacob A Siefert, Herschel C Pangborn, and Neera Jain. A set-based approach for robust control co-design. *arXiv preprint arXiv:2310.11658*, 2023.

Erik Blasch, Tien Pham, Chee-Yee Chong, Wolfgang Koch, Henry Leung, Dave Braines, and Tarek Abdelzaher. Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges. *IEEE Aerospace and Electronic Systems Magazine*, 36(7):80–93, 2021.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017.

Jan Boelts, Jan-Matthis Lueckmann, Richard Gao, and Jakob H Macke. Flexible and efficient simulation-based inference for models of decision-making. *Elife*, 11:e77220, 2022.

William M Bolstad and James M Curran. *Introduction to Bayesian statistics*. John Wiley & Sons, 2016.

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. *arXiv preprint arXiv:2306.03838*, 2023.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Ramon F Brena, Antonio A Aguileta, Luis A Trejo, Erik Molino-Minero-Re, and Oscar Mayora. Choosing the best sensor fusion method: A machine-learning approach. *Sensors*, 20(8):2350, 2020.

Haim Brezis and Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. Lqr through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Christian Bucher. Asymptotic sampling for high-dimensional reliability analysis. *Probabilistic Engineering Mechanics*, 24(4):504–510, 2009.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Clement Caron, Philippe Lauret, and Alain Bastide. Machine learning to speed up computational fluid dynamics engineering simulations for built environments: A review. *Building and Environment*, 267:112229, 2025.

Jesús Carrete, Hadrián Montes-Campos, Ralf Wanzenböck, Esther Heid, and Georg KH Madsen. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *The Journal of Chemical Physics*, 158(20), 2023.

Diah Chaerani, Cornelis Roos, and A Aman. The robust shortest path problem by means of robust linear optimization. In *Operations Research Proceedings 2004: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR). Jointly Organized with the Netherlands Society for Operations Research (NGB) Tilburg, September 1–3, 2004*, pages 335–342. Springer, 2005.

Vivien J Challis and James K Guest. Level set topology optimization of fluids in stokes flow. *International journal for numerical methods in engineering*, 79(10):1284–1308, 2009.

Timothy M Chan. A (slightly) faster algorithm for klee's measure problem. In *Proceedings of the twenty-fourth annual symposium on Computational geometry*, pages 94–100, 2008.

Prasad Vilas Chanekar, Nikhil Chopra, and Shapour Azarm. Co-design of linear systems using generalized benders decomposition. *Automatica*, 89:180–193, 2018.

Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional gaussian mixture model for characterizing variability in cryo-em. *Nature methods*, 18(8):930–936, 2021.

Zheng Chen, Bin Yao, and Qingfeng Wang. μ-synthesis-based adaptive robust control of linear motor driven stages with high-frequency dynamics: A case study. *IEEE/ASME Transactions on Mechatronics*, 20(3):1482–1490, 2014.

Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *Operations Research*, 2022.

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, 35:9525–9537, 2022.

Meysam Cheramin, Richard Li-Yang Chen, Jianqiang Cheng, and Ali Pinar. Data-driven robust optimization using scenario-induced uncertainty sets. *arXiv preprint arXiv:2107.04977*, 2021.

Labane Chrif and Zemalache Meguenni Kadda. Aircraft control system using lqg and lqr controller with optimal estimation-kalman filter design. *Procedia Engineering*, 80:245–257, 2014.

Santiago Cortes-Gomez, Carlos Patino, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. *arXiv preprint arXiv:2410.01767*, 2024.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

Tonghui Cui, James T Allison, and Pingfeng Wang. A comparative study of formulations and algorithms for reliability-based co-design problems. *Journal of Mechanical Design*, 142(3): 031104, 2020a.

Tonghui Cui, James T Allison, and Pingfeng Wang. Reliability-based co-design of state-constrained stochastic dynamical systems. In *AIAA Scitech 2020 Forum*, page 0413, 2020b.

Tonghui Cui, Zhuoyuan Zheng, and Pingfeng Wang. Control co-design of lithium-ion batteries for enhanced fast-charging and cycle life performances. *Journal of Electrochemical Energy Conversion and Storage*, 19(3):031001, 2022.

Tore Dalenius. The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1950 (3-4):203–213, 1950.

Tore Dalenius and Margaret Gurney. The problem of optimum stratification. ii. *Scandinavian Actuarial Journal*, 1951(1-2):133–148, 1951.

Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. *arXiv preprint arXiv:2210.04815*, 2022.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. Towards reliable simulation-based inference with balanced neural ratio estimation. *arXiv preprint arXiv:2208.13624*, 2022.

Arnaud Delaunoy, Benjamin Kurt Miller, Patrick Forré, Christoph Weniger, and Gilles Louppe. Balancing simulation-based inference for conservative posteriors. *arXiv preprint arXiv:2304.10978*, 2023.

John Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3, 1990.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Yiheng Du, Nithin Chalapathi, and Aditi Krishnapriyan. Neural spectral methods: Self-supervised learning in the spectral domain. *arXiv preprint arXiv:2312.05225*, 2023.

Peter D Dunning and H Alicia Kim. Introducing the sequential linear programming level-set method for topology optimization. *Structural and Multidisciplinary Optimization*, 51:631–643, 2015.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in Neural Information Processing Systems*, 32, 2019.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: normalizing flows in PyTorch, November 2020a. URL https://doi.org/10.5281/zenodo.4296287.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International conference on machine learning*, pages 2771–2781. PMLR, 2020b.

H. Edelsbrunner. The union of balls and its dual shape. *Discrete & Computational Geometry*, 13 (3):415–440, Jun 1995. ISSN 1432-0444. doi: 10.1007/BF02574053. URL https://doi.org/10.1007/BF02574053.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022.

Adam N Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, pages 2858–2867. PMLR, 2020.

Robert Falck, Justin S Gray, Kaushik Ponnapalli, and Ted Wright. dymos: A python package for optimal control of multidisciplinary systems. *Journal of Open Source Software*, 6(59):2809, 2021.

Maciej Falkiewicz, Naoya Takeishi, Imahn Shekhzadeh, Antoine Wehenkel, Arnaud Delaunoy, Gilles Louppe, and Alexandros Kalousis. Calibrating neural simulation-based inference with differentiable coverage probability. *arXiv preprint arXiv:2310.13402*, 2023.

Vladimir Sergeevich Fanaskov and Ivan V Oseledets. Spectral neural operators. In *Doklady Mathematics*, volume 108, pages S226–S232. Springer, 2023.

Kai-Tai Fang and Jianxin Pan. A review of representative points of statistical distributions and their applications. *Mathematics*, 11(13):2930, 2023.

Hosam K Fathy, Julie A Reyer, Panos Y Papalambros, and AG Ulsov. On the coupling between the plant and controller optimization problems. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 3, pages 1864–1869. IEEE, 2001.

Hosam K Fathy, Panos Y Papalambros, A Galip Ulsoy, and Davor Hrovat. Nested plant/controller optimization with application to combined passive/active automotive suspensions. In *Proceedings of the 2003 American Control Conference, 2003.*, volume 4, pages 3375–3380. IEEE, 2003.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023a.

Shai Feldman, Bat-Sheva Einbinder, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pages 624–626. PMLR, 2023b.

Sebastian Felix Fischer, Matthias Feurer, and Bernd Bischl. Openml-ctr23–a curated tabular regression benchmarking suite. In *AutoML Conference 2023 (Workshop)*, 2023.

George Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer Science & Business Media, 2013.

Bernard D Flury and Thaddeus Tarpey. Representing a large collection of curves: A case for principal points. *The American Statistician*, 47(4):304–306, 1993.

Gabriele Franch, Daniele Nerini, Marta Pendesini, Luca Coviello, Giuseppe Jurman, and Cesare Furlanello. Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events. *Atmosphere*, 11(3):267, 2020.

Bernard Friedland. *Advanced control system design*. Prentice-Hall, Inc., 1995.

Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3):471–483, 2014.

Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Mario Garcia-Sanz. Control co-design: an engineering game changer. *Advanced Control for Applications: Engineering and Industrial Systems*, 1(1):e18, 2019.

Matteo Gasparin and Aaditya Ramdas. Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*, 2024a.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*, 2024b.

Mohamad Gharib and Andrea Bondavalli. On the evaluation measures for machine learning algorithms for safety-critical systems. In *2019 15th European Dependable Computing Conference (EDCC)*, pages 141–144. IEEE, 2019.

Daniel Gilman, Jo Bovy, Tommaso Treu, Anna Nierenberg, Simon Birrer, Andrew Benson, and Omid Sameie. Strong lensing signatures of self-interacting dark matter in low-mass haloes. *Monthly Notices of the Royal Astronomical Society*, 507(2):2432–2447, 2021.

Lily Goli, Cody Reading, Silvia Selllán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. *arXiv preprint arXiv:2309.03185*, 2023.

Vignesh Gopakumar, Joel Oskarrson, Ander Gray, Lorenzo Zanisi, Stanislas Pamela, Daniel Giles, Matt Kusner, and Marc Deisenroth. Valid error bars for neural weather models using conformal prediction. *arXiv preprint arXiv:2406.14483*, 2024.

Benjamin Gravell and Tyler Summers. Robust learning-based control via bootstrapped multiplicative noise. In *Learning for Dynamics and Control*, pages 599–607. PMLR, 2020.

Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020a.

Benjamin Gravell, Iman Shames, and Tyler Summers. Robust data-driven output feedback control via bootstrapped multiplicative noise. In *Learning for Dynamics and Control Conference*, pages 650–662. PMLR, 2022.

Benjamin J Gravell, Peyman Mohajerin Esfahani, and Tyler H Summers. Robust control design for linear systems via multiplicative noise. *IFAC-PapersOnLine*, 53(2):7392–7399, 2020b.

Ander Gray, Vignesh Gopakumar, Sylvain Rousseau, and Sebastien Destercke. Guaranteed prediction sets for functional surrogate models. In *The 41st Conference on Uncertainty in Artificial Intelligence*.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

Harshit Gupta, Michael T McCann, Laurene Donati, and Michael Unser. Cryogan: A new reconstruction paradigm for single-particle cryo-em via deep adversarial learning. *IEEE Transactions on Computational Imaging*, 7:759–774, 2021.

ChangHoon Hahn, KJ Kwon, Rita Tojeiro, Malgorzata Siudek, Rebecca EA Canning, Mar Mezcua, Jeremy L Tinker, David Brooks, Peter Doel, Kevin Fanning, et al. The desi probabilistic value-added bright galaxy survey (provabgs) mock challenge. *The Astrophysical Journal*, 945(1):16, 2023.

Marc Hallin, Davy Paindaveine, and Marianna Šiman. Multivariate quantiles and multiple-output regression quantiles: From ℓ_1 optimization to halfspace depth. *Annals of Statistics*, 38:635–669, 2010.

Kanza Hamid, Amina Asif, Wajid Abbasi, Durre Sabih, and Fayyaz-ul-Amir Afsar Minhas. Machine learning with abstention for automated liver disease diagnosis. In *2017 international conference on frontiers of information technology (FIT)*, pages 356–361. IEEE, 2017.

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020.

Daniel R Herber and James T Allison. Nested and simultaneous solution strategies for general combined plant and control design problems. *Journal of Mechanical Design*, 141(1):011402, 2019.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.

Joeri Hermans, Nilanjan Banik, Christoph Weniger, Gianfranco Bertone, and Gilles Louppe. Towards constraining warm dark matter with stellar streams through neural simulation-based inference. *Monthly Notices of the Royal Astronomical Society*, 507(2):1999–2011, 2021a.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021b.

Yashar D Hezaveh, Neal Dalal, Daniel P Marrone, Yao-Yuan Mao, Warren Morningstar, Di Wen, Roger D Blandford, John E Carlstrom, Christopher D Fassnacht, Gilbert P Holder, et al. Detection of lensing substructure using ALMA observations of the dusty galaxy SDP.81. *The Astrophysical Journal*, 823(1):37, 2016.

Matthew D Hoffman, Tuan Anh Le, Pavel Sountsov, Christopher Suter, Ben Lee, Vikash K Mansinghka, and Rif A Saurous. Probnerf: Uncertainty-aware inference of 3d shapes from 2d images. In *International Conference on Artificial Intelligence and Statistics*, pages 10425–10444. PMLR, 2023.

David W Hogg and RD Blandford. The gravitational lens system B1422+231: dark matter, superluminal expansion and the Hubble constant. *Monthly Notices of the Royal Astronomical Society*, 268(4):889–893, 1994.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Yeh-Liang Hsu. A review of structural shape optimization. *Computers in Industry*, 25(1):3–13, 1994.

Yuge Hu, Joseph Musielewicz, Zachary W Ulissi, and Andrew J Medford. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 3(4):045028, 2022.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *The Journal of Machine Learning Research*, 23(1):3772–3803, 2022.

Eric Jenn, Alexandre Albore, Franck Mamalet, Grégory Flandin, Christophe Gabreau, Hervé Delseny, Adrien Gauffriau, Hugues Bonnin, Lucian Alecu, Jérémy Pirard, et al. Identifying challenges to the certification of machine learning for safety critical systems. In *European congress on embedded real time systems (ERTS 2020)*, volume 1, 2020.

Yu Jiang, Yebin Wang, Scott A Bortoff, and Zhong-Ping Jiang. An iterative approach to the optimal co-design of linear control systems. *International Journal of Control*, 89(4):680–690, 2016.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Conformal and Probabilistic Prediction and Applications*, pages 72–90. PMLR, 2021.

Benoît Jubin. Intrinsic volumes of sublevel sets. *arXiv preprint arXiv:1903.01592*, 2019.

Abdullah Kamadan, Gullu Kiziltas, and Volkan Patoglu. Co-design strategies for optimal variable stiffness actuation. *IEEE/ASME Transactions on Mechatronics*, 22(6):2768–2779, 2017.

Margot E Kaminski. The right to explanation, explained. *Berkeley Technology Law Journal*, 34 (1):189–218, 2019.

Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165, 2021.

Taylor Killian, George Konidaris, and Finale Doshi-Velez. Transfer learning across patient variations with hidden parameter markov decision processes. *arXiv preprint arXiv:1612.00475*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hamza Boubacar Kirgni and Junling Wang. Lqr-based adaptive tsmc for nuclear reactor in load following operation. *Progress in Nuclear Energy*, 156:104560, 2023.

Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*, 2025.

Jonas Köhler, Andreas Krämer, and Frank Noé. Smooth normalizing flows. *Advances in Neural Information Processing Systems*, 34:2796–2809, 2021.

Linglong Kong and Ivan Mizera. Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica*, pages 1589–1610, 2012.

Václav Kořenář. Vehicle routing problem with stochastic demands. *ALLOCATION FRAGMENTS OF THE DISTRIBUTED DATABASE*, page 24, 2003.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

John M Lee and John M Lee. *Smooth manifolds*. Springer, 2012.

Sang Hoon Lee and Wei Chen. A comparative study of uncertainty propagation methods for black-box type functions. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 48078, pages 1275–1284, 2007.

Sang Hoon Lee and Wei Chen. A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and multidisciplinary optimization*, 37(3):239–253, 2009.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11668–11675. IEEE, 2024.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. *arXiv preprint arXiv:2302.03026*, 2023.

Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022a.

Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in Neural Information Processing Systems*, 35:13038–13049, 2022b.

Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

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*, 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, Andrew Stuart, Kaushik Bhattacharya, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems*, 33:6755–6766, 2020c.

Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212, 2022.

Mingxing Liu, Junfeng Wang, Tao Lin, Quan Ma, Zhiyang Fang, and Yanqun Wu. An empirical study of the code generation of safety-critical software using llms. *Applied Sciences*, 14(3): 1046, 2024.

Yulin Liu, Tianhao Qie, Xinan Zhang, Hao Wang, Zhongbao Wei, Herbert HC Iu, and Tyrone Fernando. A novel online learning-based linear quadratic regulator for vanadium redox flow battery in dc microgrids. *Journal of Power Sources*, 587:233672, 2023a.

Ziyuan Liu, Yuhang Wu, Daniel Zhengyu Huang, Hong Zhang, Xu Qian, and Songhe Song. Spfno: Spectral operator learning for pdes with dirichlet and neumann boundary conditions. *arXiv preprint arXiv:2312.06980*, 2023b.

Lennart Ljung et al. Theory for the user. *System identification*, 1987.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in neural information processing systems*, 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

Rui Luo and Zhixin Zhou. Weighted aggregation of conformity scores for classification. *arXiv preprint arXiv:2407.10230*, 2024.

Ziqi Ma, Kamyar Azizzadenesheli, and Anima Anandkumar. Calibrated uncertainty quantification for operator learning via conformal prediction. *arXiv preprint arXiv:2402.01960*, 2024.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Giorgos Mamakoukas, Maria Castano, Xiaobo Tan, and Todd Murphey. Local koopman operators for data-driven control of robotic systems. In *Robotics: science and systems*, 2019.

Huiying Mao, Ryan Martin, and Brian J Reich. Valid model-free spatial prediction. *Journal of the American Statistical Association*, pages 1–11, 2022.

Simon Martina-Perez, Matthew J Simpson, and Ruth E Baker. Bayesian uncertainty quantification for data-driven equation learning. *Proceedings of the Royal Society A*, 477(2254):20210426, 2021.

Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1): 7–12, 1960.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Benjamin Kurt Miller, Christoph Weniger, and Patrick Forré. Contrastive neural ratio estimation. *arXiv preprint arXiv:2210.06170*, 2022.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Amirhossein Mollaali, Izzet Sahin, Iqrar Raza, Christian Moya, Guillermo Paniagua, and Guang Lin. A physics-guided bi-fidelity fourier-featured operator learning framework for predicting time evolution of drag and lift coefficients. *arXiv preprint arXiv:2311.03639*, 2023.

Andrea Mor and Maria Grazia Speranza. Vehicle routing problems over time: a survey. *Annals of Operations Research*, 314(1):255–275, 2022.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

Mervin E Muller. A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM*, 2(4):19–20, 1959.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL probml.ai.

Austin L Nash, Herschel C Pangborn, and Neera Jain. Robust control co-design with receding-horizon mpc. In *2021 American Control Conference (ACC)*, pages 373–379. IEEE, 2021.

Youssef SG Nashed, Frédéric Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. Cryoposenet: end-to-end simultaneous learning of single-particle orientation and 3d map reconstruction from cryo-electron microscopy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4076, 2021.

Vu Linh Nguyen, Chin-Hsing Kuo, and Po Ting Lin. Reliability-based analysis and optimization of the gravity balancing performance of spring-articulated serial robots with uncertainties. *Journal of Mechanisms and Robotics*, 14(3):031016, 2022.

Shunichi Ohmori. A predictive prescription using minimum volume k-nearest neighbor enclosing ellipsoid and robust optimization. *Mathematics*, 9(2):119, 2021.

Michał Okulewicz and Jacek Mańdziuk. A metaheuristic approach to solve dynamic vehicle routing problem in continuous search space. *Swarm and Evolutionary Computation*, 48:44–61, 2019.

Davy Paindaveine and Miroslav Šiman. On directional multiple-output quantile regression. *Journal of Multivariate Analysis*, 102(2):193–212, 2011.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

Paraskevas N Paraskevopoulos. *Modern control engineering*. CRC Press, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Yash Patel, Declan McNamara, Jackson Loper, Jeffrey Regier, and Ambuj Tewari. Variational inference with coverage guarantees. *arXiv preprint arXiv:2305.14275*, 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Egon Peršak and Miguel F Anjos. Contextual robust optimisation with uncertainty quantification. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 124–132. Springer, 2023.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Apostolos F Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, 477:111902, 2023.

Seppo Pulkkinen, Daniele Nerini, Andrés A Pérez Hortal, Carlos Velasco-Forero, Alan Seed, Urs Germann, and Loris Foresti. Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development*, 12(10):4185–4219, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Julie A Reyer, Hosam K Fathy, Panos Y Papalambros, and A Galip Ulsoy. Comparison of combined embodiment design and control optimization strategies using optimality conditions. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 80234, pages 1023–1032. American Society of Mechanical Engineers, 2001.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Hippolyt Ritter and Theofanis Karaletsos. Tyxe: Pyro-based bayesian neural nets for pytorch. *arXiv preprint arXiv:2110.00276*, 2021.

Eduardo Ochoa Rivera, Yash Patel, and Ambuj Tewari. Conformal prediction for ensembles: Improving efficiency via score-based aggregation. *arXiv preprint arXiv:2405.16246*, 2024.

Peter J Rousseeuw and Anja Struyf. Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8:193–203, 1998.

Meead Saberi and İ Ömer Verbas. Continuous approximation model for the vehicle routing problem for emissions minimization at the strategic level. *Journal of Transportation Engineering*, 138(11):1368–1376, 2012.

Utsav Sadana, Abhilash Chenreddy, Erick Delage, Alexandre Forel, Emma Frejinger, and Thibaut Vidal. A survey of contextual optimization methods for decision making under uncertainty. *arXiv preprint arXiv:2306.10374*, 2023.

Izzet Sahin, Christian Moya, Amirhossein Mollaali, Guang Lin, and Guillermo Paniagua. Deep operator learning-based surrogate models with uncertainty quantification for optimizing internal cooling channel rib profiles. *International Journal of Heat and Mass Transfer*, 219:124813, 2024.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Robert E Schapire et al. A brief introduction to boosting. In *Ijcai*, volume 99, pages 1401–1406. Citeseer, 1999.

Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

Daniela Schuster. Abstaining machine learning: philosophical considerations. *AI & SOCIETY*, pages 1–21, 2025.

Peter Seiler, Andrew Packard, and Pascal Gahinet. An introduction to disk margins [lecture notes]. *IEEE Control Systems Magazine*, 40(5):78–95, 2020.

Robert Serfling. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56(2):214–232, 2002.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Chao Shang and Fengqi You. A data-driven robust optimization approach to scenario-based stochastic model predictive control. *Journal of Process Control*, 75:24–39, 2019.

Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *2021 International Conference on 3D Vision (3DV)*, pages 972–981. IEEE, 2021.

Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 540–557. Springer, 2022.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Jan Sokolowski, Jean-Paul Zolésio, Jan Sokolowski, and Jean-Paul Zolesio. *Introduction to shape optimization*. Springer, 1992.

Jacob Sturm. Elliptic partial differential equations. https://sites.rutgers.edu/jacob-sturm/wp-content/uploads/sites/553/2021/11/Elliptic-PDE-112717.pdf, 2017. Lecture notes, Rutgers University. Version dated 27 Nov 2017.

Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6301–6310, 2019.

Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual lp. *arXiv preprint arXiv:2305.15686*, 2023.

Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23: 2031–2038, 2013.

Yue Sun and Maryam Fazel. Learning optimal controllers by policy gradient: Global optimality via convex parameterization. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4576–4581. IEEE, 2021.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. *arXiv preprint arXiv:2209.08718*, 2022.

Florian Tambon, Gabriel Laberge, Le An, Amin Nikanjam, Paulina Stevia Nouwou Mindom, Yann Pequignot, Foutse Khomh, Giulio Antoniol, Ettore Merlo, and Francois Laviolette. How to certify machine learning based safety-critical systems? a systematic literature review. *Automated Software Engineering*, 29(2):38, 2022.

Russ Tedrake. *Underactuated Robotics*. 2023. URL https://underactuated.csail.mit.edu.

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. Likelihood-Free Inference by Ratio Estimation. *Bayesian Analysis*, 17(1):1 – 31, 2022. doi: 10.1214/20-BA1238. URL https://doi.org/10.1214/20-BA1238.

Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE, 2020.

Rohit K Tripathy and Ilias Bilionis. Deep uq: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of computational physics*, 375:565–588, 2018.

Vladimir G Trunov and Vladimir V V'yugin. Online aggregation of conformal predictive systems. In *Conformal and Probabilistic Prediction with Applications*, pages 430–449. PMLR, 2023.

Renukanandan Tumu, Lars Lindemann, Truong Nghiem, and Rahul Mangharam. Physics constrained motion prediction with uncertainty quantification. *arXiv preprint arXiv:2302.01060*, 2023.

Karen Ullrich, Rianne van den Berg, Marcus Brubaker, David Fleet, and Max Welling. Differentiable probabilistic models of scientific imaging with the fourier slice theorem. *arXiv preprint arXiv:1906.07582*, 2019.

S Vegetti, LVE Koopmans, A Bolton, T Treu, and R Gavazzi. Detection of a dark substructure through gravitational imaging. *Monthly Notices of the Royal Astronomical Society*, 408(4): 1969–1981, 2010.

Simona Vegetti and LVE Koopmans. Statistics of mass substructure from strong gravitational lensing: quantifying the mass fraction and mass function. *Monthly Notices of the Royal Astronomical Society*, 400(3):1583–1592, 2009.

VV V'yugin and VG Trunov. Online aggregation of conformal forecasting systems. *Journal of Communications Technology and Electronics*, 68(Suppl 2):S239–S253, 2023.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David M Blei. Probabilistic conformal prediction using conditional random samples. *arXiv preprint arXiv:2206.06584*, 2022.

Gege Wen, Zongyi Li, Qirui Long, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. Accelerating carbon capture and storage modeling using fourier neural operators. *arXiv*, 2022.

Jan Willems. Least squares stationary optimal control and the algebraic riccati equation. *IEEE Transactions on automatic control*, 16(6):621–634, 1971.

Haixu Wu, Tengge Hu, Huakun Luo, Jianmin Wang, and Mingsheng Long. Solving high-dimensional pdes with latent spectral models. *arXiv preprint arXiv:2301.12664*, 2023.

Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, and Yuandong Tian. Learning and planning with a semantic model. *arXiv preprint arXiv:1809.10842*, 2018.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, pages 1–13, 2024.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. A multi-view deep learning method for epileptic seizure detection using short-time fourier transform. In *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*, pages 213–222, 2017.

Ye Yuan, Guangxu Xun, Kebin Jia, and Aidong Zhang. A multi-view deep learning framework for eeg seizure detection. *IEEE journal of biomedical and health informatics*, 23(1):83–94, 2018.

Jerzy Zabczyk. *Mathematical control theory*. Springer, 2020.

Cha Zhang and Yunqian Ma. *Ensemble machine learning*, volume 144. Springer, 2012.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024.

Zhi-rong Zhang, Liu Hui, ZHAO Feng, et al. Application of cfd in ship engineering design practice and ship hydrodynamics. *Journal of Hydrodynamics, Ser. B*, 18(3):315–322, 2006.

Dongdong Zhao, Xiaodi Yang, Yichang Li, Li Xu, Jinhua She, and Shi Yan. A kalman–koopman lqr control approach to robotic systems. *IEEE Transactions on Industrial Electronics*, 2024.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021a.

Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021b.

Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394:56–81, 2019.

Zongren Zou and George Em Karniadakis. L-hydra: multi-head physics-informed neural networks. *arXiv preprint arXiv:2301.02152*, 2023.