Data-Driven Models for Uncertainty and Behavior

by

Vishal Gupta

B.A., Yale University (2004)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author .....
Sloan School of Management
May 7, 2014

Certified by .....
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-Director, Operations Research Center
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor, Department of Electrical Engineering
and Computer Science
Co-Director, Operations Research Center

Data-Driven Models for Uncertainty and Behavior

by

Vishal Gupta

Submitted to the Sloan School of Management
on May 7, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The last decade has seen an explosion in the availability of data. In this thesis, we propose new techniques to leverage these data to tractably model uncertainty and behavior. Specifically, this thesis consists of three parts:

In the first part, we propose a novel schema for utilizing data to design uncertainty sets for robust optimization using hypothesis testing. The approach is flexible and widely applicable, and robust optimization problems built from our new data-driven sets are computationally tractable, both theoretically and practically. Optimal solutions to these problems enjoy a strong, finite-sample probabilistic guarantee. Computational evidence from classical applications of robust optimization – queuing and portfolio management – confirm that our new data-driven sets significantly outperform traditional robust optimization techniques whenever data is available.

In the second part, we examine in detail an application of the above technique to the unit commitment problem. Unit commitment is a large-scale, multistage optimization problem under uncertainty that is critical to power system operations. Using real data from the New England market, we illustrate how our proposed data-driven uncertainty sets can be used to build high-fidelity models of the demand for electricity, and that the resulting large-scale, mixed-integer adaptive optimization problems can be solved efficiently. With respect to this second contribution, we propose new data-driven solution techniques for this class of problems inspired by ideas from machine learning. Extensive historical back-testing confirms that our proposed approach generates high quality solutions that compare with state-of-the-art methods.

In the third part, we focus on behavioral modeling. Utility maximization (single-agent case) and equilibrium modeling (multi-agent case) are by far the most common behavioral models in operations research. By combining ideas from inverse optimization with the theory of variational inequalities, we develop an efficient, data-driven technique for estimating the primitives of these models. Our approach supports both parametric and nonparametric estimation through kernel learning. We prove that our estimators enjoy a strong generalization guarantee even when the model is misspecified. Finally, we present computational evidence from applications in economics and transportation science illustrating the effectiveness of our approach and its scalability

to large-scale instances.

Thesis Supervisor: Dimitris Bertsimas
Title: Boeing Professor of Operations Research
Co-Director, Operations Research Center

Acknowledgments

First, and foremost, I would like to thank my supervisor, Dimitris. I've had many teachers who taught me how to solve problems; Dimitris, you taught me to be a researcher. Your indefatigable positivity and unshakable belief that good research can have lasting impact have shaped me profoundly.

I would also like to thank my committee members, Yannis and Vivek. Your guidance and reassuring support throughout my PhD were unwavering, and I am grateful. In addition, I would like to thank Georgia and Retsef for their mentorship. Know that as I continue my career as an academic, I will strive to be as generous, thoughtful and encouraging with students as you four have been with me.

I am incredibly fortunate to have so many friends and colleagues, without whom I may never have completed this thesis. Andre, you've been a brother to me. Thank you for always reminding me to work on my "life resume" in addition to my academic CV. Allison, Adam and Ross: late nights, endless simulations, broken proofs and the peer-review process are only manageable when you have friends with whom to commiserate. Thank you for being there for me. Nathan, our collaboration always reminds me how much fun good research can be.

Although they graduated many years ago, thank you to Nikos Trichakis, Dan Iancu, David Goldberg, Adrian Becker, Phil Keller, Eric Zarybnsky and Chaitanya Bandi. Your theses were inspirational, setting a high-standard that motivated me throughout my PhD. Thank you also to all of the other ORC students, especially Velibor, Paul, Nataly, Kris, Angie, Fernanda, Maxime, John, and, although she is not a student, Phebe. Together we created a community that supports intellectual exploration, discussion and work-life balance. I will miss it, dearly, when I leave.

Sasha, Hugh, Kevin Bonham, Ilana, Alison Hill, Travis, Jon, Nick Howard, Heyman and Erica Nelson: when over-specialization and hours in the office threatened to silo my perspective on the world, you would remind me of all the rest of science, art, beauty and society around us. Thank you for pulling me out of the rabbit hole. Erica, in particular, thank you for reminding me to be "all of myself," and that true

integrity compels us to question if we're doing our best and making a difference every single day.

Thank you to the ORC Staff, Andrew Carvalho and Laura Rose, for all the things you've done to help me navigate the bureaucracy of the institute.

Thank you also to Andy Sun and Jinye Zhao for providing the data used in Chapter 3 of this thesis and for helpful conversations on current state-of-practice at NE ISO. Special thanks also to Iain Dunning and Miles Lubin, two ORC students and dear friends that have developed an amazing suite of optimization tools critical to the experiments in Chapter 3. Thank you for taking the time to teach me, implement feature requests and promptly fix the incredibly rare bug. This thesis is massively improved by your help.

Finally, I'd like to thank my family. Throughout my life you've supported and encouraged me, even when you weren't always sure what I was doing – "what is operations research, exactly?" In particular, I'd like to thank my sister, Srishti, my brother-in-law Vas, and their two children, Kabir and Sai. In my hardest times, I could always depend on you for a warm meal, an ear to bend, and a child to chase around the house. I would not have made it this far without you.

Cambridge, May 2014 Vishal Gupta

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has witnessed an unprecedented explosion in the availability of data. Massive quantities of data are now routinely collected in many application domains. Retailers archive terabytes of transaction data. Suppliers track order patterns across their supply chains. Energy markets can access global weather data, historical demand profiles, and, in some cases, real-time power consumption information. These data have motivated a shift in thinking – away from a priori assumptions and reasoning and towards a new data-centered paradigm.

As part of this paradigm shift, there has been renewed emphasis on using these data to model uncertainty and human behavior. Much of this work in the statistics and machine learning communities has focused on leveraging data to create models with good predictive power, i.e., models capable of making accurate predictions that can be validated with new, yet to be seen, data.

On the other hand, in operations research and operations management applications, models for uncertainty and behavior are frequently used in conjunction with optimization techniques to build larger decision-making frameworks. These applications, then, require models that not only demonstrate good predictive power, but also integrate tractably with existing optimization methods. Unfortunately, many state-of-the-art techniques in statistics and machine learning do not exhibit this property. As an example, artificial neural networks, which have been shown to have exceptional predictive power in certain contexts, are inherently difficult to incorporate into

large-scale optimization problems because of their underlying non-convex structure.

Consequently, in what follows, we propose general purpose techniques to model uncertainty and behavior that dovetail naturally and tractably with existing optimization methods. Specifically, this thesis consists of three parts. In each part, we treat a distinct optimization method or modeling paradigm which has proven successful in practice. Given this paradigm, we seek a data-driven methodology to model uncertainty (or behavior) that 1) can be tractably incorporated into that paradigm and 2) faithfully represents the relevant uncertainty. There are various ways of making this second goal precise. We will focus on showing both theoretically and empirically that solutions to optimizations problems built from our models perform well against new, yet to be seen, data.

We next summarize the three parts of this thesis with respect to the above goals. Detailed contributions and a literature review can be found in the introduction to each chapter.

1.1 Constructing Uncertainty Sets from Data

In the first part of this thesis, we consider robust optimization, an increasingly popular approach to optimization under uncertainty. Robust optimization has proven practically successful in a range of applications, including inventory management, dynamic pricing, assortment optimization, portfolio allocation, and optimization in energy markets. (See, e.g., [3, 31, 35, 65, 89, 101, 103].) The crux of the approach is to define an uncertainty set of possible realizations of the uncertain parameters and then optimize against worst-case realizations within this set. Computational experience suggests that with well-chosen uncertainty sets, robust models yield tractable optimization problems whose solutions perform as well or better than other approaches. With poorly chosen uncertainty sets, however, robust models may be overly-conservative or even computationally intractable. Choosing a good set is crucial.

Consequently, we consider the problem of constructing uncertainty sets from data. We propose a novel schema based upon hypothesis testing. The approach is versatile,

applying to a number of common modeling situations, and can be tailored to yield a variety of distinct sets, each with their own geometric and computational properties. (The interested reader may want to skip ahead to Table 2.1 at this point to see some examples.)

With respect to our first goal (tractability), we show that robust optimization problems built from our sets are tractable both theoretically and practically. With respect to our second goal (fidelity), we first prove that solutions to these problems satisfy a strong, finite-sample, probabilistic performance guarantee. We further illustrate computationally that sets built from our schema are able to learn features of the underlying data distribution including skewness, modality, and eccentricity. Moreover, as the amount of data increases, our sets learn these features more accurately, and, consequently, shrink in size. Thus, with even modest amounts of data, they are generally much smaller (with respect to subset containment) than conventional uncertainty sets from the literature. Simulations in classical applications from robust optimization – queuing theory and portfolio management – confirm that solutions built from our data-driven sets outperform conventional approaches whenever data is available.

1.2 Data-Driven Approaches to the Unit Commitment Problem

In the second part of this thesis, we consider adaptive optimization, a generalization of robust optimization to a sequential, multistage context. (See [15] and references therein.) Like robust optimization, adaptive optimization assumes that all uncertainties belong to a pre-specified uncertainty set and then optimizes against worst-case outcomes within this set. Unlike robust optimization, however, in adaptive optimization, decisions in later stages are allowed to depend on previously observed uncertainties. This additional flexibility translates into higher quality solutions with better practical performance [16, 26].

By virtue of the sequential nature of the application, historical data for adaptive optimization often constitute a time series. To develop a model of uncertainty suitable for adaptive optimization, we propose a simple modification of our previous data-driven uncertainty sets for time series data. By construction, many of their aforementioned favorable properties are retained.

We then explore in detail an application of the above technique to the unit commitment problem in the New England electricity market. Unit commitment (UC) is a fundamental problem in electricity market operations with substantial scope. For example, the New England Independent System Operator (NE ISO) uses a variant of UC as the basis of its real-time market-clearing mechanism to determine the price of electricity. In 2012, this energy market cleared $5 billion dollars in revenue. (See Sec. 3.1 for additional context.)

We consider a two-stage, adaptive optimization problem to model UC as a large-scale, mixed-integer linear optimization problem similar in spirit to [36, 116]. Using real-data from NE ISO, we first illustrate the use of the above adaptation of our uncertainty set construction to build a high-fidelity model of the demand uncertainty for electricity. We then demonstrate that a near-optimal, affinely adaptive policy can be computed efficiently for our formulation. With respect to this contribution, we propose several novel data-driven solution techniques inspired by ideas from machine learning. These techniques are generically applicable to adaptive optimization problems and may merit interest in their own right. Finally, through extensive historical backtesting we confirm that our proposed approach generates high quality solutions that compare with state-of-the-art methods.

1.3 Inverse Variational Inequalities and Modeling Behavior

In the third part of this thesis, we switch our focus to modeling behavior. The most common approaches to behavioral modeling in operations research are, by far, utility

maximization for a single agent and equilibrium models such as Nash equilibrium for multiple agents. Unfortunately, in real applications the primitives of such models, namely the underlying utility functions, are generally unobservable. Consequently, we consider the problem of calibrating these functions from observable data, i.e., the historical actions of agents. Our calibrated functions can naturally then be incorporated in any of a number of existing utility based models for behavior.

By combining ideas from inverse optimization with the theory of variational inequalities, we develop an efficient, data-driven technique for estimating the primitives of these models. Specifically, we first recast both of the above models and several others in the general language of variational inequalities. VIs are a natural tool for describing equilibria with examples spanning economics, transportation science, physics, differential equations, and optimization. (See Section 4.2.1 or [70] for detailed examples.) In the language of variational inequalities, calibrating the above behavioral models amount to determining the function which describes the variational inequality. In other words, we seek to solve an *inverse variational inequality problem*: given data that we believe are equilibria, i.e., solutions to some VI, estimate the function which describes this VI, i.e., the model primitives.

We propose a general purpose, tractable approach to solving this problem. Our approach supports both the parametric case, where the underlying function is known to have some specific parametric form, and the nonparametric case, where nothing is known about its underlying structure. This second nonparametric technique draws heavily on kernel methods from statistical learning, providing a new, interesting application of that idea.

We also prove that our estimators enjoy a strong generalization guarantee, even if the underlying model is completely misspecified. We consider this an important strength distinguishing our approach from classical methods in econometrics. Indeed, recent studies in behavioral economics, most notably the nobel prize-winning work of Kahneman and Traversky, suggest that the actions of real agents rarely accord with utility theory or Nash theory exactly. Consequently, there may not even exist a true utility function which would reconcile observed behavior. In this context our

generalization guarantees prove that if our estimators fit the observed data well, they will continue to fit new data well, *even if the agents do not abide by utility maximization or Nash theory.* (See Sec. 4.1 for complete discussion of this point.)

We present computational evidence with examples drawn from economics and transportation science illustrating the scalability of our approach and its potential use in nonparametric estimation and inference.

1.4 Notational Conventions

Throughout the thesis, we use boldfaced capital letters (e.g. $\mathbf{A}, \mathbf{W}$) to denote matrices, boldfaced lowercase letters and greek symbols (e.g., $\mathbf{x}, (\cdot), \boldsymbol{\mu}$) to denote vectors and vector-valued functions, and ordinary lowercase letters and greek symbols to denote scalars. We will use calligraphic capital letters (e.g. $\mathcal{S}$) to denote sets. Finally, variables or vectors superscripted with a "tilde" (e.g. $\tilde{\mathbf{u}}$) refer to random variables or uncertain quantities, while the same variables superscripted with a "hat" (e.g., $\hat{\mathbf{u}}$) refer to realizations of those variables in a particular data set. Any other notation specific to a chapter is introduced in that chapter.

Chapter 2

Constructing Uncertainty Sets From Data

2.1 Introduction

In this chapter, we propose a general schema for designing uncertainty sets for robust optimization problems. We focus on modeling a single, uncertain linear constraint $\tilde{\mathbf{u}}^T\mathbf{x} \leq b$, with $\tilde{\mathbf{u}}$ uncertain, by a corresponding robust constraint

$$\mathbf{u}^T\mathbf{x} \leq b, \quad \forall \mathbf{u} \in \mathcal{U}. \tag{2.1}$$

Traditional approaches [9, 12, 14, 33, 47] typically assume $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known exactly. These approaches combine a priori reasoning with mild assumptions on $\mathbb{P}^*$. For example, they may assume that $\mathbb{P}^*$ has independent components, but typically do not assume that its marginal distributions are known precisely. These approaches then seek $\mathcal{U}$ that satisfy two key properties:

- The robust linear constraint $\mathbf{u}^T\mathbf{x} \leq b \; \forall \mathbf{u} \in \mathcal{U}$ is *computationally tractable.*
- For any desired $\epsilon > 0$, the set $\mathcal{U}$ can be tuned so that it *implies a probabilistic guarantee for* $\mathbb{P}^*$ *at level* ϵ. By this we mean that for any $\mathbf{x}^* \in \mathbb{R}^d$ and $b \in \mathbb{R}$, we have the following implication:

$$\text{If } \mathbf{u}^T\mathbf{x}^* \leq b \;\; \forall \mathbf{u} \in \mathcal{U}, \text{ then } \mathbb{P}^*(\tilde{\mathbf{u}}^T\mathbf{x}^* \leq b) \geq 1 - \epsilon.$$

Observe that this property is *less* stringent than asking that $\mathbb{P}^*(\tilde{\mathbf{u}} \in \mathcal{U}) \geq 1 - \epsilon$ [cf. pg. 32-33 12]. Rather, the property ensures that a feasible solution to the robust constraint will also be feasible with probability ϵ with respect to $\mathbb{P}^*$. Existing proposals for uncertainty sets achieve this second property – despite not knowing $\mathbb{P}^*$ exactly – by leveraging the a priori structural features of $\mathbb{P}^*$.

Like previous proposals, we, too, will assume the uncertainty $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known exactly, and we will seek sets $\mathcal{U}$ that satisfy the above two properties. Unlike previous proposals – and this is critical – we additionally assume that we have data $\mathcal{S}$ drawn i.i.d. according to $\mathbb{P}^*$. These data $\mathcal{S}$ contain more detailed information about $\mathbb{P}^*$ than the original a priori structural features. By leveraging this additional information, we can design new sets that imply similar probabilistic guarantees, but which are much smaller (with respect to subset containment) than their traditional counterparts. Consequently, robust models built from our new sets yield less conservative solutions than traditional counterparts, while retaining their robustness properties.

The key to our schema is to use the confidence region of a statistical hypothesis test to quantify our knowledge about $\mathbb{P}^*$ from the data. Specifically, our set constructions will depend on three ingredients: the a priori assumption on $\mathbb{P}^*$, the data, and the choice of hypothesis test. By pairing different a priori assumptions with different hypothesis tests, we obtain different data-driven uncertainty sets, each with its own geometric properties, computational burden and modeling power. These sets are capable of capturing a variety of features of $\mathbb{P}^*$ – skewness, heavy-tails or correlations. In this sense, our approach is very flexible.

In principle, there is a multitude of possible pairings of a priori assumptions and tests, yielding a multitude of different sets. In this paper, we focus on those pairings we believe are most relevant to applied robust modeling. Specifically, we consider a priori assumptions that are common in practice and hypothesis tests that lead to

tractable uncertainty sets. Our list is non-exhaustive; there may exist other pairings that yield effective sets. Nonetheless, we feel these pairings cover a broad range of realistic scenarios. Specifically, we consider situations where:

- $\mathbb{P}^*$ has known, finite discrete support (Sec. 2.3).
- $\mathbb{P}^*$ *may* have continuous support, and the components of $\tilde{\mathbf{u}}$ are independent (Sec. 2.4).
- We observe data drawn from the marginal distributions of $\mathbb{P}^*$ separately, and these marginals *may* have continuous support (Sec. 2.5). This situation may occur, e.g., when the components of the data are sampled asynchronously, or there are many missing values.
- $\mathbb{P}^*$ *may* have continuous support, and the data are sampled from the joint distribution (Sec. 2.6). This is the general case.

Table 2.1 summarizes the a priori structural assumptions we consider, the corresponding hypothesis test and the resulting uncertainty set. Each set is convex and admits a tractable, compact description; see the referenced equations. One can use general purpose techniques to represent robust linear constraints over each of our sets tractably. By exploiting their specific structure, however, we propose specialized algorithms with improved performance. The column "Computational Burden" in Table 2.1 roughly describes the bottleneck in separating over a robust linear constraint for each set.

We are not the first to consider using hypothesis tests in robust optimization. [65] have proposed an uncertainty set for the mean and covariance of a random variable based on hypothesis tests in a specific linear-regression context. Similarly, [81] have proposed a method for a specific inventory control problem based on a distributionally robust dynamic program that is motivated from Pearson's χ^2 test. They, however, do not connect their work to designing uncertainty sets, and it is unclear how to extend their approach to other contexts such as distributions with continuous support. Finally, [17] have characterized the tractability of robust linear constraints over sets

Assumptions on $\mathbb{P}^*$	Hypothesis Test	Description	Eqs.	Support Function
Discrete support	χ^2-test	SOCP	(2.11) (2.13)	
Discrete support	G-test	LP*	(2.8) (2.12)	SOCP / GP
Independent marginals	KS Test	LP*	(2.15) (2.20)	Line search
Independent marginals	K Test	LP*	(A.9) (A.13)	Line search
Independent marginals	CvM Test	SOCP*	(A.13) (A.11) (A.6)	
Independent marginals	W Test	SOCP*	(A.13) (A.11) (A.7)	
Independent marginals	AD Test	GP	(2.20) (A.11) (A.8)	
Independent marginals	Forward/Backward Deviations	SOCP	(2.27)	Closed form
I.I.D. marginals	KS 2 Sample Test	LP	(2.28) (A.18)	Sorting
None	Marginal Samples	LP	(2.32)	Closed form
None	Calafiore & El Ghaoui, 2006	SOCP	(2.34)	Closed form
None	Delage & Ye, 2010	SDP	(2.35)	

Table 2.1: Summary of data-driven uncertainty sets proposed in this paper. We use LP, SOCP, GP, and SDP to denote linear, second-order cone, geometric, and semidefinite optimization problems respectively. The additional "*" notation indicates a problem of the above type with one additional, convex, nonlinear constraint. We use KS, K, CvM, W, and AD to abbreviate the Kolmogorov-Smirnov, Kuiper, Cramer-von Mises, Watson and Anderson-Darling goodness of fit tests, respectively.

describing uncertain probability distributions motivated by phi-divergences. Phi-divergences are related to certain hypothesis tests when $\mathbb{P}^*$ has discrete support. It is unclear how to extend their method when the uncertain parameters are not probability distributions.

By contrast, we offer a comprehensive study of the connection between hypothesis testing and uncertainty set design. Importantly, this connection allows us to apply state-of-the-art methods from statistics directly in designing our sets. We illustrate this idea by showing how the bootstrap and Gaussian approximations can be used to further refine many of the constructions in Table 2.1.

Moreover, we argue this connection to hypothesis testing provides a unified view of many other data-driven methods from the literature. For example, [42] and [52] have proposed alternative data-driven methods for chance-constrained and distributionally robust problems, respectively. We build on these works by showing, first, how they, too, can be interpreted through the lens of hypothesis testing, and, second, how they can also be used to design data-driven uncertainty sets. Thus, hypothesis testing provides a common ground from which to compare and contrast these methods.

We summarize our contributions:

1. We propose a new, unified schema for constructing uncertainty sets from data using statistical hypothesis tests. Under the assumption that the data is drawn i.i.d. from an unknown distribution $\mathbb{P}^*$, sets built from our schema imply a probabilistic guarantee for $\mathbb{P}^*$ at any desired level ϵ.

2. We illustrate our schema by constructing a multitude of uncertainty sets. Each set is applicable under slightly different a priori assumptions on $\mathbb{P}^*$ as described in Table 2.1.

3. We show that robust linear optimization problems over each of our sets can be solved in polynomial time with practically efficient algorithms suitable for large-scale instances.

4. Through applications in portfolio allocation and queueing, we compare the

strengths and weaknesses of our sets in relation to one another and their traditional counterparts. Overall, we find that our new sets outperform their traditional variants whenever data is available with minimal additional computational overhead.

We stress that our aspiration in this paper is to influence the *practice* of robust optimization. We intend practitioners to actually use the constructions referenced in Table 2.1 in real-life applications of robust modeling. To this end, throughout the paper we focus on the construction of each set, its modeling power, and the computational complexity of solving robust constraints over these sets. Whenever possible, we have deferred technical proofs to the online e-companion.

The remainder of the paper is structured as follows. Sec. 2.2 introduces our general schema for constructing uncertainty sets from the confidence regions of hypothesis tests. Sec. 2.3-2.6 describe the various constructions in Table 2.1. Sec. 2.7 outlines the usage of the bootstrap and Gaussian approximations to refine the above constructions. Sec. 2.8 provides guidelines for practitioners. Sec. 2.9 presents applications and Sec. 2.10 concludes.

2.1.1 Additional Notation

In this chapter and the next, we will always use $\tilde{\mathbf{u}} \in \mathbb{R}^d$ to denote a *random* vector and $\tilde{u}_i$ to denote its components. $\mathbb{P}^*$ denotes its true (unknown) measure. Moreover, for a generic probability measure $\mathbb{P}$, $\mathbb{P}_i$ denotes the marginal measure of $\tilde{u}_i$. Finally, we let $\mathcal{S} = \{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_N\}$ be a sample of N data points drawn i.i.d. according to $\mathbb{P}^*$ and let $\mathbb{P}^*_{\mathcal{S}}$ denote the measure of the sample $\mathcal{S}$, i.e., the N-fold product distribution of $\mathbb{P}^*$.

2.2 General Schema

2.2.1 Background on Hypothesis Tests

A typical hypothesis test compares two hypotheses, a null-hypothesis H_0 and an alternative hypothesis H_A, each of which make a claim about an unknown distribution $\mathbb{P}^*$. For a given significance level δ with $0 < \delta < 1$ and some data $\mathcal{S}$ drawn from $\mathbb{P}^*$, the test prescribes a threshold depending on δ and a statistic depending on $\mathcal{S}$. If this statistic exceeds the threshold, we reject the null-hypothesis in favor of the alternative. Otherwise, we deem there is insignificant evidence to reject the null-hypothesis. Since the test statistic depends on $\mathcal{S}$, it is random. The threshold is chosen so that, under additional assumptions on $\mathbb{P}^*$ which depend on the particular hypothesis test, the probability (with respect to the sample) of *incorrectly* rejecting the null-hypothesis is at most δ. For most tests, lower values of δ correspond to a lower probability of *correctly* rejecting the null-hypothesis. Thus, the significance level δ implicitly controls this tradeoff. Choosing an appropriate significance level is an application specific task, though values of $\delta = 1\%, 5\%$ or 10% are common [cf. 83, Chapt. 3.1].

As an example, consider Student's t-test [83, Chapt. 5]. Given some value $\mu_0 \in \mathbb{R}$, the t-test compares the hypothesis $H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{u}] = \mu_0$ with $H_A : \mathbb{E}^{\mathbb{P}^*}[\tilde{u}] \neq \mu_0$, using the test statistic $t \equiv \frac{\hat{\mu}-\mu_0}{\hat{\sigma}\sqrt{N}}$. Here $\hat{\mu}, \hat{\sigma}$ are the sample mean and sample standard deviation, respectively. It rejects H_0 at level δ if $\left|\frac{\hat{\mu}-\mu_0}{\hat{\sigma}\sqrt{N}}\right| > t_{N-1,1-\delta/2}$ where $t_{N-d,1-\delta}$ is the $1-\delta$ quantile of the Student t distribution with $N-1$ degrees of freedom. Under the assumption that $\mathbb{P}^*$ is Gaussian, the test guarantees that we will incorrectly reject H_0 with probability at most δ.

Given data $\mathcal{S}$ and any hypothesis test, the $1-\delta$ confidence region of the test is defined as the set of parameters that would pass the test at level δ for that data. For example, the confidence region of the t test is $\left\{\mu \in \mathbb{R} : \left|\frac{\hat{\mu}-\mu}{\hat{\sigma}\sqrt{N}}\right| \leq t_{N-1,1-\delta/2}\right\}$. (Notice here that both $\hat{\mu}$ and $\hat{\sigma}$ depend on $\mathcal{S}$.) In what follows, however, we will commit a slight abuse of nomenclature and instead use the term confidence region to refer to the set of all measures which are consistent with the assumptions of the hypothesis test and also pass the test given the data. With this definition, the confidence region

of the t-test becomes

$$\mathcal{P}^t \equiv \left\{ \mathbb{P} \in \Theta : \mathbb{P} \text{ is Gaussian with mean } \mu, \text{ and } \left| \frac{\hat{\mu} - \mu}{\hat{\sigma}\sqrt{N}} \right| \leq t_{N-1,1-\delta/2} \right\}, \tag{2.2}$$

where Θ is the set of Borel probability measures on $\mathbb{R}$.

Our interest in confidence regions stems from the following fact: when the assumptions of a hypothesis test hold, the probability (with respect to the sampling procedure) that $\mathbb{P}^*$ is a member its confidence region is at least $1 - \delta$. For example, with the t-test, when $\mathbb{P}^*$ is truly Gaussian, we have that $\mathbb{P}^*_S(\mathbb{P}^* \in \mathcal{P}^t) \geq 1 - \delta$. (Recall, $\mathbb{P}^*_S$ denotes probability with respect to the sample $\mathcal{S}$, i.e., the N-fold product distribution of $\mathbb{P}^*$.)

This is a critical observation. Despite not knowing $\mathbb{P}^*$, we can use a hypothesis test to create a set of distributions from the data which will contain $\mathbb{P}^*$ for any specified probability. These confidence regions will play a pivotal role in our schema for designing uncertainty sets.

2.2.2 Designing Uncertainty Sets from Confidence Regions

For any set $\mathcal{U}$, the support function of $\mathcal{U}$, denoted $\phi_{\mathcal{U}}$, is defined by

$$\phi_{\mathcal{U}}(\mathbf{x}) \equiv \max_{\mathbf{u} \in \mathcal{U}} \mathbf{u}^T \mathbf{x}.$$

By construction, support functions of convex sets are convex and positively homogenous. (We say $\phi(\mathbf{x})$ is positively homogenous if $\phi(\lambda \mathbf{x}) = \lambda \phi(\mathbf{x})$ for all $\lambda > 0$.) Moreover, for any convex, positively homogenous function ϕ, there exists a convex set $\mathcal{U}$ such that $\phi = \phi_{\mathcal{U}}$ [23]).

For any $\mathbf{x} \in \mathbb{R}^d$ and measure $\mathbb{P}$, the Value at Risk at level ϵ with respect to $\mathbf{x}$ is

$$\text{VaR}^{\mathbb{P}}_{\epsilon}(\mathbf{x}) \equiv \inf \left\{ t : \mathbb{P}(\tilde{\mathbf{u}}^T \mathbf{x} \leq t) \geq 1 - \epsilon \right\}. \tag{2.3}$$

Value at Risk is positively homogenous (in $\mathbf{x}$), but typically non-convex.

There is a close relationship between $\text{VaR}^{\mathbb{P}}_{\epsilon}$ and sets $\mathcal{U}$ which imply a probabilistic

guarantee for $\mathbb{P}$ at level ϵ. The following proposition is implicitly used by a number of authors in the literature when designing uncertainty sets [12, 48], but, to the best of our knowledge, has never been stated explicitly. For concreteness, we include a short proof.

Proposition 2.1. *A set $\mathcal{U}$ implies a probabilistic guarantee for $\mathbb{P}$ at level ϵ if and only if*

$$\phi_{\mathcal{U}}(\mathbf{x}) \geq VaR_{\epsilon}^{\mathbb{P}}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d. \tag{2.4}$$

Proof. Suppose (2.4) is true and that $\mathbf{x}^*$ is feasible in Eq. (2.1). Then,

$$b \geq \max_{\mathbf{u} \in \mathcal{U}} \mathbf{u}^T \mathbf{x}^* = \phi_{\mathcal{U}}(\mathbf{x}^*) \geq \mathrm{VaR}_{\epsilon}^{\mathbb{P}}(\mathbf{x}).$$

From Eq. (2.3), this last inequality implies $\mathbb{P}(\tilde{\mathbf{u}}^T \mathbf{x}^* \leq b) \geq 1 - \epsilon$. Hence, $\mathcal{U}$ implies a probabilistic guarantee.

Next, suppose $\mathcal{U}$ implies a probabilistic guarantee, but $\exists \mathbf{x}^* \in \mathbb{R}^d$ such that, $\phi_{\mathcal{U}}(\mathbf{x}^*) < \mathrm{VaR}_{\epsilon}^{\mathbb{P}}(\mathbf{x}^*)$. Choose b such that $\phi_{\mathcal{U}}(\mathbf{x}^*) \leq b < \mathrm{VaR}_{\epsilon}^{\mathbb{P}}(\mathbf{x}^*)$. Since $\phi_{\mathcal{U}}(\mathbf{x}^*) \leq b$ and $\mathcal{U}$ implies a probabilistic guarantee, it must be that $\mathbb{P}(\tilde{\mathbf{u}}^T \mathbf{x} \leq b) \geq 1 - \epsilon$. This contradicts Eq. (2.3) since $b < \mathrm{VaR}_{\epsilon}^{\mathbb{P}}(\mathbf{x}^*)$. □

Proposition 2.1 suggests that an *ideal* set $\mathcal{U}$ would satisfy $\phi_{\mathcal{U}}(\mathbf{x}) = \mathrm{VaR}_{\epsilon}^{\mathbb{P}^*}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. Indeed, if we could find such a $\mathcal{U}$, then $\mathcal{U}$ would be the smallest set with respect to subset containment which implies a probabilistic guarantee. (Specifically, for any other set $\mathcal{U}'$ that implies a probabilistic guarantee $\phi_{\mathcal{U}'} \geq \phi_{\mathcal{U}}$, which implies $\mathcal{U} \subseteq \mathcal{U}'$.)

There are at least two challenges with finding such an ideal set. First, if $\mathrm{VaR}_{\epsilon}^{\mathbb{P}^*}(\mathbf{x})$ is non-convex, such a $\mathcal{U}$ may not exist. Second, $\mathbb{P}^*$ is not known precisely. Using the confidence regions of the previous section an the data $\mathcal{S}$, however, we can identify a set of measures $\mathcal{P}$, which contain $\mathbb{P}^*$ with probability $1 - \delta$. This motivates the following schema: Fix δ with $0 < \delta < 1$ and ϵ with $0 < \epsilon < 1$.

1. Let $\mathcal{P}(\mathcal{S})$ be the confidence region of a hypothesis test at level δ.

2. Construct a convex, positively homogenous function $\phi(\mathbf{x}, \mathcal{S})$ such that

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S})} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \phi(\mathbf{x}, \mathcal{S}) \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

3. Identify the set $\mathcal{U}(\mathcal{S}, \delta)$ whose support function coincides with $\phi(\mathbf{x}, \mathcal{S})$.

We have

Theorem 2.1. *$\mathcal{U}(\mathcal{S}, \delta)$ implies a probabilistic guarantee for every $\mathbb{P} \in \mathcal{P}(\mathcal{S})$. In particular, with probability $1 - \delta$ with respect to the sampling distribution, $\mathcal{U}(\mathcal{S}, \delta)$ implies a probabilistic guarantee for $\mathbb{P}^*$.*

Proof. The first part of the theorem follows directly from Proposition 2.1. The second follows by the construction of $\mathcal{P}(\mathcal{S})$. □

In what follows, δ and $\mathcal{S}$ are typically fixed, and ϵ is typically clear from context. Consequently, we may suppress some or all of them in the notation.

Of course, one could use any family $\mathcal{P}$ such that $\mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^* \in \mathcal{P}) \geq 1 - \delta$ in Step 1, above, and the resulting $\mathcal{U}$ would satisfy the same guarantee. In Sec. 2.6.1, however, we show that any such family can be interpreted as the confidence region of an appropriate hypothesis test. Thus, the above schema unifies a variety of approaches.

All of the sets we derive in this chapter follow from applying this schema with different hypothesis tests. In this sense, this schema is the key idea of the chapter. The challenge in each case is in constructing a suitable upper bound ϕ in Step 2. Constructing such upper bounds in the case when $\mathbb{P}^*$ is a known, fixed measure is a well-studied problem [55, 91]. Many of our proofs involve generalizing these bounds to the case $\mathbb{P} \in \mathcal{P}(\mathcal{S})$.

2.2.3 Relationship with Other Optimization Approaches

Before proceeding to applications of the above schema, however, we relate it three other research streams in optimization: chance-constrained optimization, coherent risk measures and distributionally robust optimization.

Chance-Constrained Optimization

In the chance-constrained paradigm, one models an uncertain constraint such as $\tilde{\mathbf{u}}^T\mathbf{x} \leq b$ by its chance-constrained counterpart

$$\mathbb{P}^*(\tilde{\mathbf{u}}^T\mathbf{x} > b) \leq \epsilon. \tag{2.5}$$

for some, user-specified ϵ. When $\mathbb{P}^*$ is not known exactly, we replace eq. (2.5) by

$$\sup_{\mathbb{P}\in\mathcal{P}^{Chance}} \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > b) \leq \epsilon. \tag{2.6}$$

where $\mathcal{P}^{Chance}$ is a family of measures known to contain $\mathbb{P}^*$ with certainty. Most authors consider families $\mathcal{P}^{Chance}$ motivated by a priori, structural assumptions, e.g., all measures with a specified mean and covariance. (A recent exception is [79] which considers a data-driven family motivated by kernel density estimation.) Depending on $\mathbb{P}^*$ (resp. $\mathcal{P}^{Chance}$), eq. (2.5) (resp. eq. (2.6)) may be non-convex. Most authors, consequently, focus on convex, conservative approximations of these constraints, typically called *safe approximations.*

In the case of linear constraints, there is, loosely speaking, a one-to-one correspondence between safe approximations and uncertainty sets that imply a probabilistic guarantee. More precisely, taking $\mathcal{P}(\mathcal{S}) = \mathcal{P}^{Chance}$ in Step 1 of our schema, the robust constraint Eq. (2.1) is a safe approximation to the original chance constraint. Similarly, given any safe approximation to a chance constraint, denoted $\phi(\mathbf{x}) \leq b$, its homogenization $\bar{\phi}(\mathbf{x}) \leq b$ is a (possibly stronger) safe approximation, where $\bar{\phi}(\mathbf{x}) \equiv \inf_{\lambda>0} \lambda\phi(\mathbf{x}/\lambda)$. Furthermore, $\bar{\phi}$ is positively homogenous by construction. Consequently, there exists $\mathcal{U}$ such that $\phi_{\mathcal{U}} = \bar{\phi}$, and such a $\mathcal{U}$ will imply a probabilistic guarantee for all $\mathbb{P} \in \mathcal{P}^{Chance}$ by Proposition 2.1

In other words, all of the results in this chapter could be recast in the language of chance constrained optimization and would constitute novel results in that literature. In which paradigm one chooses to express results is largely a matter of taste. We prefer to work with uncertainty sets.

Connection to Coherent Risk Measures

Coherent risk measures were first introduced by [7] as a convex alternative to $\text{VaR}_\epsilon^{\mathbb{P}^*}$. The most common example of a coherent risk measure is $\text{CVaR}_\epsilon^{\mathbb{P}^*}$ (cf. Eq. (2.9)).

[24, 90] prove a one-to-one correspondence between coherent risk measures and robust linear constraints. Namely, for any compact, convex $\mathcal{U}$, there exists a coherent risk measure ρ such that the $\mathbf{u}^T\mathbf{x} \leq b \ \forall \mathbf{u} \in \mathcal{U} \iff \rho(\tilde{\mathbf{u}}^T\mathbf{x}) \leq b$. Similarly, for any coherent risk measure ρ, there exists compact, convex $\mathcal{U}$ such that the same implication holds.

Consequently, to each of our data-driven uncertainty sets in the following sections, there exists a corresponding data-driven coherent risk measure. Although we do not pursue this idea in this chapter, it is sometimes possible to describe this risk measure explicitly. More importantly, however, this risk measure differs from classical risk measures like $\text{CVaR}_\epsilon^{\mathbb{P}^*}$ because, instead of being defined with respect to the unknown distribution $\mathbb{P}^*$, it is defined with respect to the data, and, thus, entirely data-driven. Nonetheless, it still implies various guarantees with respect to this true distribution; for example, it is an upper bound to $\text{VaR}_\epsilon^{\mathbb{P}^*}$. This features distinguishes these measures from approaches like [24] which implicitly assume $\mathbb{P}^*$ is equal to the empirical probability distribution of the data.

Distributionally Robust Optimization

For a given function $g(\mathbf{x}, \mathbf{u})$, distributionally robust optimization proxies the uncertain constraint $g(\mathbf{x}, \tilde{\mathbf{u}}) \leq 0$ by $\sup_{\mathbb{P} \in \mathcal{P}^{DRO}} \mathbb{E}^{\mathbb{P}}[g(\mathbf{x}, \tilde{\mathbf{u}})] \leq 0$, where $\mathcal{P}^{DRO}$ is a family probability measures containing $\mathbb{P}^*$. Clearly, we can substitute the confidence region of a hypothesis test for $\mathcal{P}^{DRO}$ in the above constraint to yield a new, data-driven approach to distributionally robust optimization. Characterizing the complexity of the resulting constraint is well-beyond the scope of this chapter. See our related work [28] for details on when the resulting problem is tractable, comparison to the method of [52], and the resulting probabilistic guarantees.

2.3 Uncertainty Sets for Discrete Distributions

In this section, we assume a priori that $\mathbb{P}^*$ has known, finite support, i.e., $\text{supp}(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$. We consider two hypothesis tests for this setup: Pearson's χ^2 test and the G test [99]. Both tests compare the hypotheses

$$H_0 : \mathbb{P}^* = \mathbb{P}_0 \quad \text{vs.} \quad H_A : \mathbb{P}^* \neq \mathbb{P}_0, \tag{2.7}$$

where $\mathbb{P}_0$ is some specified measure. Specifically, let $p_i = \mathbb{P}_0(\tilde{\mathbf{u}} = a_i)$ be the specified null-hypothesis, and let $\hat{\mathbf{p}}$ denote the empirical probability distribution over the sample, i.e.,

$$\hat{p}_i \equiv \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}(\hat{\mathbf{u}}_j = a_i) \quad i = 0, \ldots, n-1.$$

Pearson's χ^2 test rejects H_0 at level δ if $N \sum_{i=0}^{n-1} \frac{(p_i - \hat{p}_i)^2}{p_i} > \chi^2_{n-1,1-\delta}$, where $\chi^2_{n-1,1-\delta}$ is the $1-\delta$ quantile of a χ^2 distribution with $n-1$ degrees of freedom. Similarly, the G test rejects the null hypothesis at level δ if $D(\hat{\mathbf{p}}, \mathbf{p}) > \frac{1}{2N}\chi^2_{n-1,1-\delta}$ where $D(\mathbf{p}, \mathbf{q}) \equiv \sum_{i=1}^{n} p_i \log(p_i/q_i)$ is the relative entropy between $\mathbf{p}$ and $\mathbf{q}$.

The confidence regions for Pearson's χ^2 test and the G test are, respectively,

$$\begin{aligned} \mathcal{P}^{\chi^2} &= \left\{ \mathbf{p} \in \Delta_n : \sum_{i=0}^{n-1} \frac{(p_i - \hat{p}_i)^2}{2p_i} \leq \frac{1}{2N}\chi^2_{n-1,1-\delta} \right\}, \\ \mathcal{P}^{G} &= \left\{ \mathbf{p} \in \Delta_n : D(\hat{\mathbf{p}}, \mathbf{p}) \leq \frac{1}{2N}\chi^2_{n-1,1-\delta} \right\}. \end{aligned} \tag{2.8}$$

Here $\Delta_n = \left\{ (p_0, \ldots, p_{n-1})^T : \mathbf{e}^T\mathbf{p} = 1, \;\; p_i \geq 0 \;\; i = 0, \ldots, n-1 \right\}$ denotes the probability simplex. We will use these two confidence regions in Step 1 of our schema.

For any measure $\mathbb{P}$, vector $\mathbf{x} \in \mathbb{R}^d$ and threshold ϵ, define the Conditional Value at Risk by

$$\text{CVaR}^{\mathbb{P}}_{\epsilon}(\mathbf{x}) \equiv \min_{v} \left\{ v + \frac{1}{\epsilon}\mathbb{E}^{\mathbb{P}}[(\tilde{\mathbf{u}}^T\mathbf{x} - v)^+] \right\}. \tag{2.9}$$

Conditional Value at Risk has been widely studied as a convex upper bound to Value at Risk [2, 100]. We will use this bound in Step 2 of our schema. We first require the

following well-known result, which, to the best of our knowledge, is first due to [100].

Theorem 2.2 (Rockafellar and Ursayev, 2000)**.** *Suppose supp*$(\mathbb{P}) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$ *and let* $\mathbb{P}(\tilde{\mathbf{u}} = \mathbf{a}_j) = p_j$*. Define the set*

$$\mathcal{U}^{CVaR_\epsilon^{\mathbb{P}}} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j,\ \mathbf{q} \in \Delta_n,\ \mathbf{q} \leq \frac{1}{\epsilon}\mathbf{p} \right\}. \tag{2.10}$$

Then, the support function of $\mathcal{U}^{CVaR_\epsilon^{\mathbb{P}}}$ *is* $\phi_{\mathcal{U}^{CVaR_\epsilon^{\mathbb{P}}}}(\mathbf{x}) = CVaR^{\mathbb{P}}(\mathbf{x})$*.*

We then have the following theorem:

Theorem 2.3. *Suppose supp*$(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$*. With probability* $1 - \delta$ *over the sample, the sets*

$$\mathcal{U}^{\chi^2} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j,\ \mathbf{q} \in \Delta_n,\ \mathbf{q} \leq \frac{1}{\epsilon}\mathbf{p},\ \mathbf{p} \in \mathcal{P}^{\chi^2} \right\}, \tag{2.11}$$

$$\mathcal{U}^{G} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j,\ \mathbf{q} \in \Delta_n,\ \mathbf{q} \leq \frac{1}{\epsilon}\mathbf{p},\ \mathbf{p} \in \mathcal{P}^{G} \right\}, \tag{2.12}$$

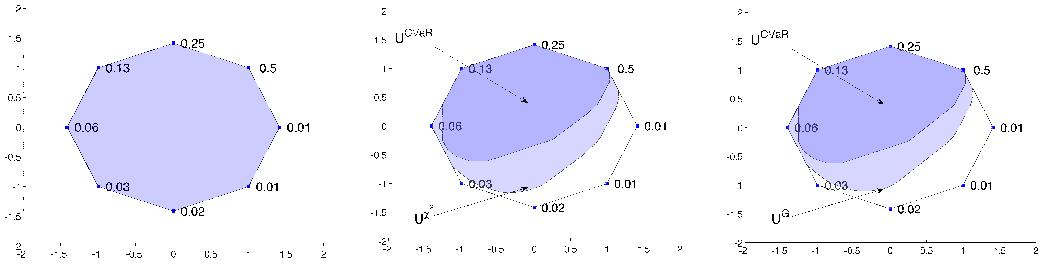
each imply a probabilistic guarantee at level ϵ *for* $\mathbb{P}^*$*.*

Proof. We prove the theorem for $\mathcal{U}^{\chi^2}$. The proof for $\mathcal{U}^G$ is identical. From our schema, it suffices to show that $\max_{\mathbf{u} \in \mathcal{U}^{\chi^2}} \mathbf{u}^T\mathbf{x}$ is an upper bound to $\sup_{\mathbb{P} \in \mathcal{P}^{\chi^2}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x})$:

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{P}^{\chi^2}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) &\leq \sup_{\mathbb{P} \in \mathcal{P}^{\chi^2}} \mathrm{CVaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) && \text{(CVaR is an upper bound to VaR)} \\
&= \sup_{\mathbb{P} \in \mathcal{P}^{\chi^2}} \max_{\mathbf{u} \in \mathcal{U}^{\mathrm{CVaR}_\epsilon^{\mathbb{P}}}} \mathbf{u}^T\mathbf{x} && \text{(Thm. 2.2)} \\
&= \max_{\mathbf{u} \in \mathcal{U}^{\chi^2}} \mathbf{u}^T\mathbf{x} && \text{(Combining Eqs. (2.11) and (2.8)).}
\end{aligned}$$

□

Observe that the sets $\mathcal{U}^{\chi^2}$, $\mathcal{U}^G$ bear a strong resemblance to a popular CVaR heuristic for constructing uncertainty sets from data. In this heuristic, one uses the set $\mathcal{U}^{\mathrm{CVaR}^{\hat{\mathbb{P}}}}$ (formed by replacing $\mathbf{p}$ with $\hat{\mathbf{p}}$ in Eq. (2.10)). In fact, as $N \to \infty$, all



(a) No data (b) $\mathcal{U}^{\chi^2}$ (c) $\mathcal{U}^{G}$

Figure 2-1: The best set in the absence of data, $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^{G}$.

three of these sets converge almost surely to $\mathcal{U}^{\mathrm{CVaR}^{\mathbb{P}^*}}$ (formed by replacing $\mathbf{p}$ by $\mathbf{p}^*$ in Eq. (2.10)). The key difference is that for finite N, $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^{G}$ imply a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ, while $\mathcal{U}^{\mathrm{CVaR}^{\hat{\mathbb{P}}}}$ does not.

2.3.1 Example: $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^{G}$

Figure 2-1 illustrates the sets $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^{G}$ with a particular numerical example. The true distribution is supported on the vertices of the given octagon. Each vertex is labeled with its true probability. In the absence of data when the support of $\mathbb{P}^*$ is known, the only uncertainty set $\mathcal{U}$ which implies a probabilistic guarantee for $\mathbb{P}^*$ is the convex hull of these points. We sample $N = 500$ data points from this distribution and construct the sets $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^{G}$ (lightly shaded regions). For reference, we also plot $\mathcal{U}^{\mathrm{CVaR}_\epsilon^{\mathbb{P}^*}}$, which is the limit of both sets as $N \to \infty$. Notice that our data-driven sets are considerably smaller than the "No data" set. Moreover, both sets are very similarly shaped. We discuss this similarity in more detail in Sec. 2.3.2.

2.3.2 Solving Robust Optimization Problems over $\mathcal{U}^{\chi^2}$ and $\mathcal{U}^G$

By using the second-order cone representation of the hyperbolic constraint $2t_i p_i \geq (p_i - \hat{p}_i)^2$ [85], one can show that

$$\mathcal{P}^{\chi^2} = \Bigg\{ \mathbf{p} \in \Delta_n : \exists \mathbf{t} \in \mathbb{R}^n \quad \sum_{i=1}^{n} t_i \leq \frac{1}{2N} \chi^2_{n-1,1-\delta}, \\ \left\| \begin{pmatrix} 2(p_i - \hat{p}_i) \\ p_i - 2t_i \end{pmatrix} \right\|_2 \leq p_i + 2t_i, \text{ for } i = 1, \ldots, n \Bigg\}. \tag{2.13}$$

Consequently, $\mathcal{U}^{\chi^2}$ is second order cone representable. Using standard techniques from robust optimization, one can reformulate a robust linear constraint over $\mathcal{U}^{\chi^2}$ as a series of linear and second order cone constraints [12].

Robust linear constraints over $\mathcal{U}^G$ are less straightforward. One possibility is to use techniques from [18] to rewrite these robust constraints as a set of convex constraints which admit an explicit, self-concordant barrier. One can then optimize over these convex constraints using a custom barrier algorithm, or reformulate them as geometric programming problem and invoke an off-the-shelf-solver. Alternatively, [28] shows that $\mathcal{P}^G$ is second order cone representable using $O(N)$ variables and inequalities. This implies that $\mathcal{U}^G$ is also second order cone representable, and, consequently, robust linear optimization problems over $\mathcal{U}^G$ can be reformulated as second order cone problems.

Neither approach is entirely satisfactory. Customizing a barrier implementation requires care and programming expertise, while geometric programming problems can be numerically challenging. Finally, the $O(N)$ formulation of [28] may be impractical for large N.

Consequently, robust optimization problems over $\mathcal{U}^{\chi^2}$ are somewhat simpler than problems over $\mathcal{U}^G$. Fortunately, for large N, the difference between these two sets is negligible. We observed this feature numerically in Figure 2-1; it holds generally.

Proposition 2.2. *With arbitrarily high probability, for any* $\mathbf{p} \in \mathcal{P}^G$, $|D(\hat{\mathbf{p}}, \mathbf{p}) -$

$\sum_{j=0}^{n-1} \frac{(\hat{p}_j - p_j)^2}{2p_j}| = O(nN^{-3})$.

For a proof, see A.1.1. Thus, for large N, the constraint defining $\mathcal{P}^G$ is approximately equal to the constraint defining $\mathcal{P}^{\chi^2}$, whereby $\mathcal{U}^G$ is approximately equal to $\mathcal{U}^{\chi^2}$. From a modeling perspective, then, $\mathcal{U}^{\chi^2}$ should be preferred for its computational tractability.

2.4 Uncertainty Sets for Independent Marginal Distributions

We would like to extend these results to cases when $\mathbb{P}^*$ may have continuous support. Unfortunately, multivariate goodness-of-fit testing for distributions with continuous support is still an active area of research. Sophisticated techniques based on kernel methods and universal learning machines have been proposed [59, 67]. Few of these proposals, however, have been widely adopted in practice. In our opinion, the two most common approaches are either to group the data into a small number of bins and apply one of the tests of the previous section, or else to assume a specific dependence structure between the marginal distributions of $\mathbb{P}^*$ and then separately test the goodness-of-fit of each marginal of $\mathbb{P}^*$ and $\mathbb{P}_0$. We have already treated the first approach.

In this section, we consider the second approach, assuming the marginal distributions are independent.

2.4.1 Confidence Region the Kolmogorov-Smirnov Test

In this section, we will develop a confidence region for the i-th marginal distribution $\mathbb{P}_i$. This region will be instrumental in later constructing our sets. To simplify notation, we will drop the index i for the remainder of the subsection. We also assume that $\hat{u}^{(0)} \leq \tilde{u} \leq \hat{u}^{(N+1)}$ almost surely, and denote by $\hat{u}^{(j)}$ the j^{th} largest element of $\hat{u}^{(0)}, \hat{u}_1, \ldots, \hat{u}_N, \hat{u}^{(N+1)}$.

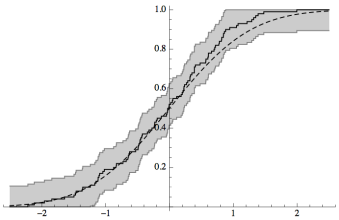


Figure 2-2: The empirical distribution function and confidence region corresponding to the KS test.

The Kolmogorov-Smirnov (KS) goodness-of-fit test is a popular, goodness-of-fit test for univariate data [109]. Given some (potentially continuous) measure $\mathbb{P}_0$, the KS test, like Pearson's χ^2 test, compares the hypotheses in Eq. (2.7). It rejects the null hypothesis at level δ if

$$\max_{j=1,\ldots,N} \max\left(\frac{j}{N} - \mathbb{P}(\tilde{u} \leq \hat{u}^{(j)}), \mathbb{P}(\tilde{u} < \hat{u}^{(j)}) - \frac{j-1}{N}\right) > \Gamma^{KS}.$$

Here $\Gamma^{KS} = \Gamma^{KS}(\delta, N)$ is the $1-\delta$ quantile of an appropriate null-distribution. Tables of such quantiles are widely available [106, 109]. The confidence region of the test is

$$\overline{\mathcal{P}}^{KS} = \left\{ \mathbb{P} \in \Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}] : \; \mathbb{P}(\tilde{u} \leq \hat{u}_j) \geq \frac{j}{N} - \Gamma^{KS}, \right.$$
$$\left. \mathbb{P}(\tilde{u} < \hat{u}_j) \leq \frac{j-1}{N} + \Gamma^{KS}, \quad j = 1, \ldots, N \right\},$$

where $\Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}]$ is the set of all Borel probability measures on $[\hat{u}^{(0)}, \hat{u}^{(N+1)}]$. Unlike $\mathcal{P}^{\chi^2}$ and $\mathcal{P}^G$, this confidence region is infinite dimensional. (We use the overline to emphasize this infinite dimensionality.) It can be visualized graphically (see Figure 2-2). Using the data, we plot the empirical distribution function $\hat{\mathbb{P}}(\tilde{u} \leq t) \equiv \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(\hat{u}_j \leq t)$ (solid line in figure.) The confidence region of the test is the band of distribution functions no more than Γ^{KS} above or below this line (grey region).

In Secs. 2.4.2 and Sec. 2.4.4, we will show that we can design data-driven uncertainty sets using this confidence region provided that we can evaluate worst-case expectations of the form $\sup_{\mathbb{P}\in\overline{\mathcal{P}}^{KS}} \mathbb{E}[e^{x\tilde{u}}]$ and $\sup_{\mathbb{P}\in\overline{\mathcal{P}}^{KS}} \mathbb{E}[x\tilde{u}]$ efficiently. Here x is

a decision variable from an outer optimization problem. In the remainder of this subsection, we will show that we can in fact evaluate this supremum in closed form.

Define

$$q_j^L(\Gamma) = \begin{cases} \Gamma & \text{if } j = 0, \\ \frac{1}{N} & \text{if } 1 \leq j \leq \lfloor N(1-\Gamma) \rfloor, \\ 1 - \Gamma - \frac{\lfloor N(1-\Gamma) \rfloor}{N} & \text{if } j = \lfloor N(1-\Gamma) \rfloor + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Similarly, define $q_j^R(\Gamma) = q_{N+1-j}^L(\Gamma), \quad j = 0, \ldots, N+1$. Observe that $\mathbf{q}^L(\Gamma), \mathbf{q}^R(\Gamma) \in \Delta_{N+2}$ so that each vector can be interpreted as a discrete probability distribution on the points $\hat{u}^{(0)}, \ldots, \hat{u}^{(N+1)}$. Finally, define

$$\mathcal{P}^{KS} = \left\{ \mathbf{p} \in \Delta_{N+2} : \exists \theta \in [0,1],\ \mathbf{p} = \theta \mathbf{q}^R(\Gamma^{KS}) + (1-\theta)\mathbf{q}^L(\Gamma^{KS}) \right\} \tag{2.15}$$

We call this set the *finite dimensional analogue* of $\overline{\mathcal{P}}^{KS}$.

Theorem 2.4.

i) Suppose $g(u)$ is monotonic. Then,

$$\sup_{\mathbb{P} \in \overline{\mathcal{P}}^{KS}} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})] = \max_{\mathbf{p} \in \mathcal{P}^{KS}} \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}), \tag{2.16}$$

ii) If $g(u)$ is non-decreasing (resp. non-increasing), then the optimum in Eq. (2.16) is attained when $\theta = 1$ (resp. $\theta = 0$).

The theorem can be proven intuitively from Fig. 2-2. We provide an explicit proof that will generalize to other results later in the paper.

Proof. For any $\theta \in [0,1]$, the discrete distribution which assigns mass $\theta q_j^R(\Gamma^{KS}) + (1-\theta) q_j^L(\Gamma^{KS})$ to the point $\hat{u}^{(j)}$ is an element of $\overline{\mathcal{P}}^{KS}$. It follows that Eq. (2.16) holds with "=" replaced by "$\geq$".

We next prove the reverse inequality. Let $\mathbb{P} \in \overline{\mathcal{P}}$. We have two cases. Suppose

first that $g(u)$ is non-decreasing. Define $\mathbb{Q} \in \overline{\mathcal{P}}$ by

$$\begin{aligned} &\mathbb{Q}(\tilde{u} = \hat{u}^{(0)}) \equiv 0, \quad \mathbb{Q}(\tilde{u} = \hat{u}^{(1)}) \equiv \mathbb{P}(\hat{u}^{(0)} \leq \tilde{u} \leq \hat{u}^{(1)}), \\ &\mathbb{Q}(\tilde{u} = \hat{u}^{(j)}) \equiv \mathbb{P}(\hat{u}^{(j-1)} < \tilde{u} \leq \hat{u}^{(j)}), \quad j = 2, \ldots, N+1. \end{aligned} \tag{2.17}$$

Then, $\mathbb{Q} \in \overline{\mathcal{P}}$. Furthermore, since $g(u)$ is non-decreasing, $\mathbb{E}^{\mathbb{P}}[g(\tilde{u})] \leq \mathbb{E}^{\mathbb{Q}}[g(\tilde{u})]$. Thus, the measure attaining the supremum on the left-hand side of Eq. (2.16) has discrete support $\{\hat{u}^{(0)}, \ldots, \hat{u}^{(N+1)}\}$, and the supremum is equivalent to the linear optimization problem:

$$\begin{aligned} \max_{\mathbf{p}} \quad & \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}) \\ \text{s.t.} \quad & \mathbf{p} \geq \mathbf{0}, \quad \mathbf{e}^T \mathbf{p} = 1, \\ & \sum_{k=0}^{j} p_k \geq \frac{j}{N} - \Gamma^{KS}, \quad \sum_{k=j}^{N+1} p_k \geq \frac{N-j+1}{N} - \Gamma^{KS}, \quad j = 1, \ldots, N, \end{aligned}$$

(We have used the fact that $\mathbb{P}(\tilde{u} < \hat{u}^{(j)}) = 1 - \mathbb{P}(\tilde{u} \geq \hat{u}^{(j)})$.) Its dual is:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y},t} \quad & \sum_{j=1}^{N} x_j \left(\Gamma^{KS} - \frac{j}{N} \right) + \sum_{j=1}^{N} y_j \left(\Gamma^{KS} - \frac{N-j+1}{N} \right) + t \\ \text{s.t.} \quad & t - \sum_{k \leq j \leq N} x_j - \sum_{1 \leq j \leq k} y_j \geq g(\hat{u}^{(k)}), \quad k = 0, \ldots, N+1, \\ & \mathbf{x}, \mathbf{y} \geq \mathbf{0}. \end{aligned}$$

Observe that the primal solution $\mathbf{q}^R(\Gamma^{KS})$ and dual solution $\mathbf{y} = \mathbf{0}$, $t = g(\hat{u}^{(N+1)})$ and

$$x_j = \begin{cases} g(\hat{u}^{(j+1)}) - g(\hat{u}^{(j)}) & \text{for } N - j^* \leq j \leq N, \\ 0 & \text{otherwise,} \end{cases}$$

constitute a primal-dual optimal pair. This proves ii) when g is non-decreasing.

Next assume that $g(u)$ is non-increasing. Given $\mathbb{P} \in \overline{\mathcal{P}}$, define $\mathbb{Q}$ by

$$\mathbb{Q}(\tilde{u} = \hat{u}^{(j)}) \equiv \mathbb{P}(\hat{u}^{(j)} \leq \tilde{u} < \hat{u}^{(j+1)}), \quad j = 0, \ldots, N-1,$$
$$\mathbb{Q}(\tilde{u} = \hat{u}^{(N)}) \equiv \mathbb{P}(\hat{u}^{(N)} \leq \tilde{u} \leq \hat{u}^{(N+1)}), \quad \mathbb{Q}(\tilde{u} = \hat{u}^{(N+1)}) \equiv 0.$$

Again, $\mathbb{Q} \in \overline{\mathcal{P}}$ and $\mathbb{E}^{\mathbb{P}}[g(\tilde{u})] \leq \mathbb{E}^{\mathbb{Q}}[g(\tilde{u})]$. Consequently, the measure attaining the supremum in Eq. (2.16) is discrete and can be found by solving the same linear optimization problem as above. Moreover, $\mathbf{q}^L(\Gamma^{KS})$ and the solution $\mathbf{x} = \mathbf{0}$, $t = g(\hat{u}^{(0)})$ and

$$y_j = \begin{cases} g(\hat{u}^{(j-1)}) - g(\hat{u}^{(j)}), & \text{for } 1 \leq j \leq j^* + 1, \\ 0 & \text{otherwise,} \end{cases}$$

constitute a primal-dual optimal pair. This proves ii) when g is non-increasing. Combining both cases proves the "$\leq$" inequality in Eq. (2.16). □

Note

$$\max_{\mathbf{p} \in \mathcal{P}^{KS}} \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}) = \max \left(\sum_{j=0}^{N+1} q_j^R(\Gamma^{KS}) g(\hat{u}^{(j)}), \sum_{j=0}^{N+1} q_j^L(\Gamma^{KS}) g(\hat{u}^{(j)}) \right). \tag{2.18}$$

Thus, from Thm. 2.4, we can evaluate worst-case expectations over $\overline{\mathcal{P}}^{KS}$ in closed-form for monotonic g.

2.4.2 Uncertainty Sets Built from the Kolmogorov-Smirnov Test

In this section, we use $\overline{\mathcal{P}}^{KS}$ to construct uncertainty sets when the components of $\tilde{\mathbf{u}}$ are independent. Under this assumption, the following is a valid goodness-of-fit test: Reject the null-hypothesis if $\mathbb{P}_i$ fails the KS test at level $\delta' = 1 - \sqrt[d]{1-\delta}$ for any i. Indeed,

$$\mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^*_i \text{ is accepted by KS at level } \delta' \text{ for all } i = 1, \ldots, d) = \prod_{i=1}^{d} \sqrt[d]{1-\delta'} = 1 - \delta,$$

by independence.

The confidence region of this test is

$$\overline{\mathcal{P}}^I = \Big\{ \mathbb{P} \in \Theta[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] : \mathbb{P} = \prod_{i=1}^{d} \mathbb{P}_i, \quad \mathbb{P}_i \in \overline{\mathcal{P}}_i^{KS} \quad i = 1, \ldots, d \Big\},$$

where $\Theta[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$ is the set of Borel probability measures supported on $[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. (The superscript "I" is to emphasize independence). We use this confidence region in Step 1 of our schema.

[91] prove the following upper bound to $\mathrm{VaR}_\epsilon^{\mathbb{P}}$ when the marginals are known to be independent: $\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \inf_{\lambda \geq 0} \left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log \mathbb{E}^{\mathbb{P}_i}[e^{x_i \tilde{u}_i / \lambda}] \right)$. We use this bound in Step 2 of our schema. Namely, by passing the supremum through the infimum and logarithm and invoking Thm. 2.4, we obtain

$$\sup_{\mathbb{P} \in \overline{\mathcal{P}}^I} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \inf_{\lambda \geq 0} \left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log \left(\max_{\mathbf{p}_i \in \mathcal{P}_i^{KS}} \sum_{j=0}^{N+1} p_{ij} e^{x_i \hat{u}_i^{(j)} / \lambda} \right) \right). \tag{2.19}$$

This upper bound is convex and positively homogenous. Thm 2.5 describes its uncertainty set.

Theorem 2.5. *Suppose $\mathbb{P}^*$ is known to have independent components, with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Let $\mathcal{P}_i^{KS}$ be the finite dimensional analogue of a $1 - \sqrt[d]{1-\delta}$ confidence region for the i^{th} marginal distribution for the KS test. Then, with probability $1 - \delta$ over the sample, the set*

$$\mathcal{U}^I = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \ \exists \mathbf{q}_i \in \Delta_{N+2}, \quad \sum_{j=0}^{N+1} \hat{u}_i^{(j)} q_{ij} = u_i, \quad \mathbf{p}_i \in \mathcal{P}_i^{KS}, \quad i = 1, \ldots, d, \\ \sum_{i=1}^{d} D(\mathbf{q}_i, \mathbf{p}_i) \leq \log(1/\epsilon) \Bigg\} \tag{2.20}$$

implies a probabilistic guarantee at level ϵ for $\mathbb{P}^*$.

Proof. We prove that the support function of $\mathcal{U}^I$ is equal to the right-hand side

Eq. (2.19). By Lagrangian duality,

$$\max_{\mathbf{u}\in\mathcal{U}^I} \mathbf{u}^T\mathbf{x} = \inf_{\lambda\geq 0} \left(\begin{array}{ll} \lambda \log(1/\epsilon) + \max\limits_{\mathbf{u}} & \sum\limits_{i=1}^{d} x_i \sum\limits_{j=0}^{N+1} \hat{u}_i^{(j)} q_{ij} - \lambda \sum\limits_{i=1}^{d} D(\mathbf{q}_i, \mathbf{p}_i) \\ & \text{s.t.} \quad \mathbf{q}_i \in \Delta_{N+2}, \quad \mathbf{p}_i \in \mathcal{P}_i^{KS}, \quad i = 1, \ldots, d. \end{array} \right)$$

The inner maximization decouples in the variables indexed by i. The i^{th} subproblem is

$$\max_{\mathbf{p}_i\in\mathcal{P}_i^{KS}} \lambda \left\{ \max_{\mathbf{q}_i\in\Delta_{N+2}} \left\{ \sum_{j=0}^{N+1} \frac{x_i \hat{u}_i^{(j)}}{\lambda} q_{ij} - D(\mathbf{q}_i, \mathbf{p}_i) \right\} \right\}.$$

The inner maximization can be solved analytically [38, pg. 93], yielding:

$$q_{ij} = \frac{p_{ij} e^{x_i \hat{u}_i^{(j)}/\lambda}}{\sum_{j=0}^{N+1} p_{ij} e^{x_i \hat{u}_i^{(j)}/\lambda}}. \tag{2.21}$$

Substituting in this solution and recombining subproblems yields Eq. (2.19). □

2.4.3 Solving Robust Problems over $\mathcal{U}^I$

Using the exponential cone, robust linear constraints over $\mathcal{U}^I$ can be reformulated as conic optimization problems (see [38] for background on the exponential cone). Although polynomial time solvable, these problems can be numerically challenging.

Instead, we utilize a cutting plane algorithm to generate valid cuts for the robust constraint $\mathbf{u}^T\mathbf{x} \leq b \ \forall \mathbf{u} \in \mathcal{U}^I$. Specifically, for any $\mathbf{x}_0 \in \mathbb{R}^d$, we evaluate Eq. (2.19) at $\mathbf{x}_0$. If the optimum value is at most b, then $\mathbf{x}_0$ is valid for the robust constraint. Else, we use the corresponding λ and $\mathbf{p}$ in Eq. (2.21) to generate a valid cut that is violated at $\mathbf{x}_0$. The key bottleneck, then, is evaluating Eq. (2.19), which by Eq. (2.18) reduces to solving a one dimensional convex optimization over λ and can be done via a line search.

In some applications, iterative cut generation may not be viable. In the next subsection, we introduce a relaxation of the set $\mathcal{U}^I$ that may be more computationally

appealing.

2.4.4 Uncertainty Sets Motivated by Forward and Backward Deviations

Previous authors have suggested upper bounding the worst case moment generating function in Eq. (2.19) by simpler functions [12, Chapt 2.]. Although this provides a worse bound for the VaR_ϵ, the resulting uncertainty set may be more computationally tractable.

In particular, given a distribution $\mathbb{P}_i^*$ with known mean μ_i, define its forward and backward deviations by

$$\sigma_{fi}^* = \sup_{x>0} \sqrt{-\frac{2\mu_i}{x} + \frac{2}{x^2}\log(\mathbb{E}^{\mathbb{P}_i^*}[e^{x\tilde{u}_i}])}, \quad \sigma_{bi}^* = \sup_{x>0} \sqrt{\frac{2\mu_i}{x} + \frac{2}{x^2}\log(\mathbb{E}^{\mathbb{P}_i^*}[e^{-x\tilde{u}_i}])}. \tag{2.22}$$

[48] suggest the upper bound

$$\mathbb{E}^{\mathbb{P}_i^*}[e^{x_i\tilde{u}_i}] \leq \begin{cases} e^{x_i\mu_i + x_i^2\sigma_f^{*2}/2} \text{ if } x_i \geq 0, \\ e^{x_i\mu_i + x_i^2\sigma_b^{*2}/2} \text{ if } x_i < 0. \end{cases}$$

The validity of the bound follows directly from the definitions of $\sigma_{fi}^*, \sigma_{bi}^*$.

We propose the following adaptation of this bound when the mean and forward and backward deviations are unknown, but we have access to data:

$$\sup_{\mathbb{P}_i \in \overline{\mathcal{P}}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}[e^{x_i\tilde{u}_i}] \leq \begin{cases} e^{m_{fi}x_i + x_i^2\sigma_{fi}^2/2} & \text{if } x_i \geq 0, \\ e^{m_{bi}x_i + x_i^2\sigma_{bi}^2/2} & \text{if } x_i < 0, \end{cases} \tag{2.23}$$

where

$$m_{fi} = \sup_{\mathbb{P}_i \in \overline{\mathcal{P}}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}[\tilde{u}_i]. \quad \sigma_{fi}^2 = \sup_{x>0} -\frac{2m_{fi}}{x} + \frac{2}{x^2}\log\left(\sup_{\mathbb{P}_i \in \overline{\mathcal{P}}^{KS}} \mathbb{E}^{\mathbb{P}_i}[e^{x_i\tilde{u}_i}]\right),$$
$$m_{bi} = \inf_{\mathbb{P}_i \in \overline{\mathcal{P}}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}[\tilde{u}_i]. \quad \sigma_{bi}^2 = \sup_{x>0} \frac{2m_{bi}}{x} + \frac{2}{x^2}\log\left(\sup_{\mathbb{P}_i \in \overline{\mathcal{P}}^{KS}} \mathbb{E}^{\mathbb{P}_i}[e^{-x_i\tilde{u}_i}]\right).$$

Again, the validity of the bound follows directly from the definitions of σ_{fi}^2 and σ_{bi}^2. Using Thm. 2.4 we can solve each of the inner optimizations in closed-form, yielding:

$$\begin{aligned} m_{fi} &= \sum_{j=0}^{N+1} q_j^R(\Gamma^{KS})\hat{u}_i^{(j)}, \quad \sigma_{fi}^2 = \sup_{x>0} -\frac{2m_{fi}}{x} + \frac{2}{x^2}\log\left(\sum_{j=0}^{N+1} q_j^R(\Gamma^{KS})e^{x_i\hat{u}_i^{(j)}}\right), \\ m_{bi} &= \sum_{j=0}^{N+1} q_j^L(\Gamma^{KS})\hat{u}_i^{(j)}, \quad \sigma_{bi}^2 = \sup_{x>0} \frac{2m_{bi}}{x} + \frac{2}{x^2}\log\left(\sum_{j=0}^{N+1} q_j^L(\Gamma^{KS})e^{-x_i\hat{u}_i^{(j)}}\right). \end{aligned} \tag{2.24}$$

Notice the optimizations defining σ_{fi}, σ_{bi} are one dimensional, convex problems which can be solved by a line search.

Substituting Eq. (2.23) into Eq. (2.19) yields

$$\begin{aligned} \sup_{\mathbb{P}\in\overline{\mathcal{P}}^I} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq & \sum_{i:x_i\geq 0} m_{fi}x_i + \sum_{i:x_i<0} m_{bi}x_i \\ & + \inf_{\lambda\geq 0} \lambda\log(1/\epsilon) + \frac{1}{2\lambda}\left(\sum_{i:x_i\geq 0}\sigma_{fi}^2x_i^2 + \sum_{i:x_i<0}\sigma_{bi}^2x_i^2\right). \end{aligned} \tag{2.25}$$

The optimization in λ can be computed in closed-form:

$$\begin{aligned} \sup_{\mathbb{P}\in\overline{\mathcal{P}}^I} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq & \sum_{i:x_i\geq 0} m_{fi}x_i + \sum_{i:x_i<0} m_{bi}x_i \\ & + \sqrt{2\log(1/\epsilon)\left(\sum_{i:x_i\geq 0}\sigma_{fi}^2x_i^2 + \sum_{i:x_i<0}\sigma_{bi}^2x_i^2\right)}. \end{aligned} \tag{2.26}$$

We use this upper bound in Step 2 of our schema. We obtain

Theorem 2.6. *Suppose $\mathbb{P}^*$ is known to have independent components with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$ bounded. Then with probability $1-\delta$ with respect to the sample, the set*

$$\begin{aligned} \mathcal{U}^{FB} = \Bigg\{ & \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_3 : \mathbf{y}_2, \mathbf{y}_2 \in \mathbb{R}_+^d, \quad \sum_{i=1}^d \frac{y_{2i}^2}{2\sigma_{fi}^2} + \frac{y_{3i}^2}{2\sigma_{bi}^2} \leq \log(1/\epsilon), \\ & m_{bi} \leq y_{1i} \leq m_{fi}, \quad i = 1,\ldots,d \Bigg\} \end{aligned} \tag{2.27}$$

implies a probabilistic guarantee for $\mathbb{P}^*$ *at level* ϵ.

The proof is straightforward and requires showing that support function of $\mathcal{U}^{FB}$ is given by Eq. (2.26). We defer it until Appendix A.1.2.

2.4.5 Solving Robust Optimization Problems over $\mathcal{U}^{FB}$

$\mathcal{U}^{FB}$ is second order cone representable. We can reformulate robust linear constraints over $\mathcal{U}^{FB}$ as a second order cone constraints using standard techniques.

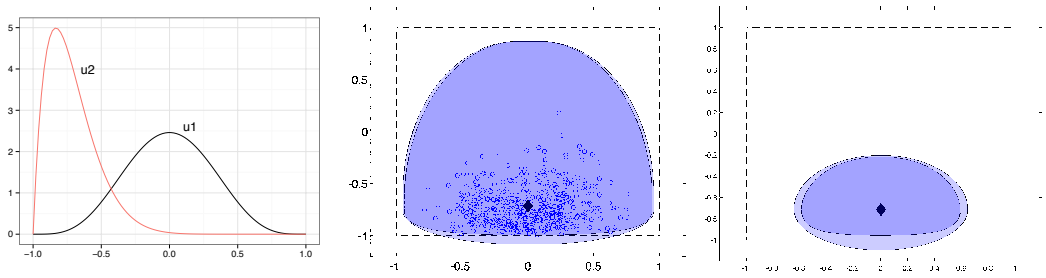
We note in passing that in some cases, $\mathcal{U}^{FB} \not\subseteq \text{supp}(\mathbb{P}^*)$. In these cases, the smaller set $\mathcal{U}^{FB} \cap \text{supp}(\mathbb{P}^*)$ also implies a probabilistic guarantee and may be preferred. The complexity of solving robust optimization problems over this intersection depends on the shape of $\text{supp}(\mathbb{P}^*)$. In the most common case when $\text{supp}(\mathbb{P}^*)$ is a box, this can still be accomplished as a second order cone problem.

2.4.6 Example: $\mathcal{U}^I$ and $\mathcal{U}^{FB}$

Figure 2-3 illustrates the sets $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ numerically. The marginal distributions of $\mathbb{P}^*$ are independent and their densities are given in Fig. 2-3a. Notice that the first marginal is symmetric while the second is highly skewed. In the absence of any data and knowing only $\text{supp}(\mathbb{P}^*)$ and that it has independent components, the only uncertainty set which implies a probabilistic guarantee is the unit square.

We then sample $N = 500$ data points from this distribution and plot the sets $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ using the KS test with $\epsilon = \delta = 10\%$ in Fig. 2-3b. For reference we also include the true mean (black diamond), sampled data points (blue circles) and $\text{supp}(\mathbb{P}^*)$ (dotted unit square). In Fig. 2-3c, we plot the limiting shape of these two sets as $N \to \infty$.

Several features are evident from the plots: First, both sets are considerably smaller than the corresponding "No Data" set. Second, in this example, there is only a small difference between $\mathcal{U}^I$, and $\mathcal{U}^{FB}$; it is barely discernible with 500 data points. Third, both sets are able to learn that the true distribution is skewed downwards in the u_2 direction (the sets taper towards the top) and that the true distribution



(a) Densities (b) $N = 500$ (c) $N \to \infty$

Figure 2-3: The left panel shows the marginal densities. The middle panel shows $\mathcal{U}^I$ (darker blue) and $\mathcal{U}^{FB}$ (paler blue) built from 500 data points (blue circles). The right panel shows the same two sets as $N \to \infty$. In both panels, we plot the true mean (black diamond) and support (dotted square) of $\mathbb{P}^*$.

is symmetric in the u_1 direction (the sets display vertical symmetry.) This last feature highlights the ability of these methods to learn features of the underlying the distribution directly from the data. Finally, $\mathcal{U}^{FB}$ is not contained contained within $\text{supp}(\mathbb{P}^*)$.

2.4.7 Extensions to Other Empirical Distribution Tests

The KS is one of many goodness-of-fit tests based on the empirical distribution function (EDF), including the Kuiper (K), Cramer von-Mises (CvM), Watson (W) and Andersen-Darling (AD) tests [109, Chapt. 5]. We can, in fact, define analogues of $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ for each of these tests. These analogues each differ slightly in shape. Although optimizing over robust linear constraints for each of these versions can be done in polynomial time, they require significantly more computational effort than the KS versions.

Through simulation studies with a variety of different distributions – normal, exponential, bernoulli, gumbel, beta, and mixtures of these choices – we have found that the versions of $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ based on the KS test often perform as well as or better than the other EDF tests. Consequently, we recommend practitioners use the sets $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ as described. For completeness, we present the constructions for the analogous tests in Appendix A.2.

2.4.8 Uncertainty Sets for Independent, Identically Distributed Marginals

When the marginals of $\mathbb{P}^*$ are i.i.d., we can construct an uncertainty set based on the 2-sample Kolmogorov-Smirnov goodness-of-fit test. Define the functions

$$\overline{k}(i) = \min\left(N+1, \left\lceil \frac{N}{d}(i + d\Gamma^{2KS}) \right\rceil\right), \quad \underline{k}(i) = \max\left(0, \left\lceil \frac{N}{d}(i + 1 - d\Gamma^{2KS}) \right\rceil\right).$$

where Γ^{2KS} is the $1-\epsilon$ threshold of the 2-sample KS test with N and d samples, and define the set,

$$\mathcal{U}^{2KS} = \{\mathbf{u} \in \mathbb{R}^d : \hat{u}^{(\underline{k}(i))} \leq u^{(i)} \leq \hat{u}^{(\overline{k}(i))}, \quad i = 1, \ldots, d\}. \tag{2.28}$$

Theorem 2.7.

i) The set conv($\mathcal{U}^{2KS}$) implies a probabilistic guarantee for $\mathbb{P}^*$ with at level ϵ.

ii) Let σ be the permutation which orders the components of $\mathbf{x}$, i.e., such that $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(d)}$. Then,

$$\phi_{2KS}(\mathbf{x}) \equiv \max_{\mathbf{u} \in conv(\mathcal{U}^{2KS})} \mathbf{u}^T\mathbf{x} = \sum_{i=1}^{d} \max(x_{\sigma(i)}\hat{u}^{(\underline{k}(i))}, x_{\sigma(i)}\hat{u}^{(\overline{k}(i))}).$$

For a proof see Appendix A.3. From the second part of the theorem, we can separate over robust linear constraints over $conv(\mathcal{U}^{2KS})$ by sorting the components of $\mathbf{x}$.

2.5 Uncertainty Sets for Asynchronously Drawn Data

In this section, we assume we observe samples from the marginal distributions of $\mathbb{P}^*$ separately. This may happen, e.g., if the samples are drawn asynchronously, or if there are many missing values. In these cases, it is impossible to learn about the joint

distribution of $\mathbb{P}^*$ from the data. To streamline the exposition, we assume that we observe exactly N samples of each marginal distribution. The results generalize to the case of different numbers of samples at the expense of more notation.

We first develop a hypothesis test for the Value at Risk of each marginal. Define the index s by

$$s = \min\left\{k \in \mathbb{N} : \sum_{j=k}^{N} \binom{N}{j} (\epsilon/d)^{N-j}(1-\epsilon/d)^j \leq \frac{\delta}{2d}\right\}, \tag{2.29}$$

and let $s = N+1$ if the corresponding set is empty. It can be shown that $\frac{s}{N} \downarrow \frac{1-\epsilon}{2d}$ as $N \to \infty$. Moreover, in the typical case when ϵ/d is small, $N - s + 1 < s$. We have the following:

Proposition 2.3. *Consider the hypotheses:*

$$H_0 : \mathit{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) \geq \overline{q}_i \text{ and } \mathit{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\mathbf{e}_i) \geq \underline{q}_i \text{ for all } i = 1, \ldots, d,$$
$$H_A : \mathit{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) < \overline{q}_i \text{ or } \mathit{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\mathbf{e}_i) < \underline{q}_i \text{ for all } i = 1, \ldots, d,$$

where $\overline{q}_i, \underline{q}_i$ are some fixed constants. Then, the following is a valid hypothesis test at level δ:

$$\textit{Reject if for any } i,\ \hat{u}_i^{(s)} < \overline{q}_i \text{ or } -\hat{u}_i^{(N-s+1)} < \underline{q}_i.$$

The proof is a multivariate generalization of a common nonparametric test of quantiles for univariate data. See [51, Sec. 7.1] for the univariate result and Appendix A.1.3 for the multivariate proof.

The confidence region corresponding to the above test is

$$\overline{\mathcal{P}}^M = \left\{\mathbb{P} \in \Theta[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] : \ \mathrm{VaR}^{\mathbb{P}_i}_{\epsilon/d} \leq \hat{u}_i^{(s)}, \ \ \mathrm{VaR}^{\mathbb{P}_i}_{\epsilon/d} \geq \hat{u}_i^{(N-s+1)}, \ \ i = 1, \ldots, d\right\}.$$

Here "M" is to emphasize "marginals." We use this set in Step 1 of our schema.

Next, we bound $\mathrm{VaR}^{\mathbb{P}^*}_{\epsilon}(\mathbf{x})$ given only information about the marginal distributions. This problem has been well-studied in finance. Indeed, when the marginal

distributions of $\mathbb{P}^*$ are known exactly, we have

$$\mathrm{VaR}_\epsilon^{\mathbb{P}^*}(\mathbf{x}) \leq \min_{\boldsymbol{\alpha}:\mathbf{e}^T\boldsymbol{\alpha}=\epsilon} \sum_{i=1}^{d} \mathrm{VaR}_{\alpha}^{\mathbb{P}^*}(x_i\mathbf{e}_i). \tag{2.30}$$

[e.g., 56]. Moreover, this bound is tight in the sense that for any $\mathbf{x}$, ϵ and $\mathbb{P}^*$, there exists a measure $\mathbb{P}_0$ with the same marginal distributions as $\mathbb{P}^*$ for which it is tight. The minimization on the right-hand side can be difficult. We use the weaker bound $\mathrm{VaR}_\epsilon^{\mathbb{P}^*}(\mathbf{x}) \leq \sum_{i=1}^{d} \mathrm{VaR}_{\epsilon/d}^{\mathbb{P}^*}(x_i\mathbf{e}_i)$ obtained by letting $\alpha_i = \epsilon/d$ for all i.

We compute the worst case value of this bound over $\overline{\mathcal{P}}^M$. Assuming $N-s+1 < s$,

$$\sup_{\mathbb{P}\in\overline{\mathcal{P}}^M} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) \leq \sup_{\mathbb{P}\in\overline{\mathcal{P}}^M} \sum_{i=1}^{d} \mathrm{VaR}_{\epsilon/d}^{\mathbb{P}}(x_i\mathbf{e}_i) = \sum_{i=1}^{d} \max(x_i\hat{u}_i^{(N-s+1)}, x_i\hat{u}_i^{(s)}), \tag{2.31}$$

where the last equality follows from the definition of $\overline{\mathcal{P}}^M$ and the positive homogeneity of VaR. It is straightforward to check that Eq. (2.31) is convex and positively homogenous.

Theorem 2.8. *If s defined by Eq.* (2.29) *satisfies $N-s+1 < s$, then, with probability at least $1-\delta$ over the sample, the set*

$$\mathcal{U}^M = \left\{\mathbf{u} \in \mathbb{R}^d : \hat{u}_i^{(N-s+1)} \leq u_i \leq \hat{u}_i^{(s)} \quad i = 1,\ldots,d\right\}. \tag{2.32}$$

implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ.

Proof. By inspection, the support function of $\mathcal{U}^M$ is given by Eq. (2.31). □

2.6 Uncertainty Sets for General, Joint Distributions

In this section, we assume we observe samples drawn from the joint distribution of $\mathbb{P}^*$ and know a bound on the support of $\mathbb{P}^*$.

2.6.1 Uncertainty Set Motivated by Calafiore and El Ghaoui, 2006

[42] upper bound $\text{VaR}_\epsilon^{\mathbb{P}^*}$ under the assumption that the mean and covariance of $\tilde{\mathbf{u}}$ under $\mathbb{P}^*$ belong to a given convex set $\mathcal{C}$. They adapt their approach to data-driven settings by using the following result from [104]. Let $\|\cdot\|_F$ denote the Frobenius norm between matrices.

Theorem 2.9 (Cristianini and Shawe-Taylor, 2003)**.** *Suppose that supp*$(\mathbb{P}^*)$ *is contained within a ball of radius* R *and that* $N > (2 + 2\log(2/\delta))^2$*. With probability at least* $1 - \delta$ *with respect to the sampling,*

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \Gamma_1(\delta/2, N) \text{ and } \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \leq \Gamma_2(\delta/2, N),$$

where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ *denote the sample mean and covariance,* $\Gamma_1(\delta, N) = \frac{R}{\sqrt{N}}\left(2 + \sqrt{2\log 1/\delta}\right)$ *and* $\Gamma_2(\delta, N) = \frac{2R^2}{\sqrt{N}}\left(2 + \sqrt{2\log 2/\delta}\right)$.

Calafiore and El Ghaoui then apply their bound to the set

$$\mathcal{C} = \{\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d} : \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \Gamma_1(\delta/2, N), \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \leq \Gamma_2(\delta/2, N), \boldsymbol{\Sigma} \succeq \mathbf{0}\}$$

to prove:

Theorem 2.10 (Calafiore and El Ghaoui, 2006)**.** *Suppose the conditions in Thm. 2.9 hold. With probability* $1 - \delta$ *with respect to the sampling,*

$$VaR_\epsilon^{\mathbb{P}^*}(\mathbf{x}) \leq \hat{\boldsymbol{\mu}}^T\mathbf{x} + \Gamma_1(\delta/2, N)\|\mathbf{x}\|_2 + \sqrt{\frac{1-\epsilon}{\epsilon}}\sqrt{\mathbf{x}^T(\hat{\boldsymbol{\Sigma}} + \Gamma_2(\delta/2, N)\mathbf{I})\mathbf{x}}. \tag{2.33}$$

To the best of our knowledge, the authors do not connect their work to developing uncertainty sets. Creating the corresponding uncertainty set, however, is straightforward.

Theorem 2.11. *Define*

$$\mathcal{U}^{CS} = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \exists \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^d \text{ s.t.} \tag{2.34}$$

$$\mathbf{u} = \hat{\boldsymbol{\mu}} + \mathbf{y}_1 + \mathbf{C}^T \mathbf{y}_2, \quad \|\mathbf{y}_1\|_2 \leq \Gamma_1(\delta/2, N), \quad \|\mathbf{y}_2\|_2 \leq \sqrt{\frac{1-\epsilon}{\epsilon}} \Bigg\},$$

where $\mathbf{C}^T\mathbf{C} = \hat{\boldsymbol{\Sigma}} + \Gamma_2(\delta/2, N)$ *is a Cholesky decomposition. The support function of* $\mathcal{U}^{CS}$ *is given explicitly by the right-hand side of Eq.* (2.33)*. Moreover, if the conditions of Thm. 2.9 hold,* $\mathcal{U}^{CS}$ *implies a probabilistic guarantee for* $\mathbb{P}^*$*.*

See A.1.4 for a proof.

2.6.2 Uncertainty Set Motivated by Delage and Ye, 2010

[52] propose a data-driven approach for solving distributionally robust optimization problems. Their method relies on a slightly more general version of the following:

Theorem 2.12 (Delage and Ye, 2010)**.** *Let* R *be such that* $\mathbb{P}^*((\tilde{\mathbf{u}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\tilde{\mathbf{u}} - \boldsymbol{\mu}) \leq R^2) = 1$ *where* $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ *are the true mean and covariance of* $\tilde{\mathbf{u}}$ *under* $\mathbb{P}^*$*. Let,* $\gamma_1 \equiv \frac{\beta}{1-\alpha-\beta}$*,* $\gamma_2 \equiv \frac{1+\beta}{1-\alpha-\beta}$*,* $\beta \equiv \frac{R^2}{N}\left(2 + \sqrt{2\log(2/\delta)}\right)^2$*,* $\alpha \equiv \frac{R^2}{\sqrt{N}}\left(\sqrt{1 - \frac{N}{R^4}} + \sqrt{\log(4/\delta)}\right)$*, and suppose also that* N *is large enough so that* $1 - \alpha - \beta > 0$*. Finally suppose* $supp(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$*. Then with probability at least* $1 - \delta$ *with respect to the sampling,* $\mathbb{P}^* \in \mathcal{P}^{DY}$ *where*

$$\mathcal{P}^{DY} \equiv \Bigg\{ \mathbb{P} \in \Theta[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] : (\mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}] - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}] - \hat{\boldsymbol{\mu}}) \leq \gamma_1,$$

$$\mathbb{E}^{\mathbb{P}}[(\tilde{\mathbf{u}} - \hat{\boldsymbol{\mu}})(\tilde{\mathbf{u}} - \hat{\boldsymbol{\mu}})^T] \preceq \gamma_2 \hat{\boldsymbol{\Sigma}} \Bigg\}.$$

Since R is typically unknown, the authors describe an estimation procedure for R and prove a modified version of the theorem using this estimate and different constants. We treat the simpler case where R is known here. Extensions to the other case are straightforward. In contrast to Thm. 2.9, the condition on N is required for

the confidence region to be well-defined. In our experiments, we have noticed that N must frequently be in the thousands.

The authors do not connect this confidence region to designing uncertainty sets. Using our schema, though, it is straightforward to do so.

Theorem 2.13.

i) We have $\sup_{\mathbb{P}\in\mathcal{P}^{DY}} VaR_\epsilon^{\mathbb{P}}(\mathbf{x}) = \max_{\mathbf{u}\in\mathcal{U}^{DY}} \mathbf{u}^T\mathbf{x}$ *where*

$$\begin{aligned}\mathcal{U}^{DY} = \Big\{\mathbf{u}\in[\hat{\mathbf{u}}^{(0)},\hat{\mathbf{u}}^{(N+1)}] : &\exists\lambda\in\mathbb{R},\mathbf{w},\mathbf{v},\in\mathbb{R}^d,\mathbf{A},\hat{\mathbf{A}}\succeq\mathbf{0}\ s.t.\\ &\|\mathbf{w}\|\le\lambda\le\frac{1}{\epsilon},\quad (\lambda-1)\hat{\mathbf{u}}^{(0)}\le\mathbf{v}\le(\lambda-1)\hat{\mathbf{u}}^{(N+1)},\\ &\begin{pmatrix}\lambda-1 & \mathbf{v}^T\\ \mathbf{v} & \mathbf{A}\end{pmatrix}\succeq\mathbf{0},\quad \begin{pmatrix}1 & \mathbf{u}^T\\ \mathbf{u} & \hat{\mathbf{A}}\end{pmatrix}\succeq\mathbf{0},\\ &\lambda\hat{\boldsymbol{\mu}}+\sqrt{\gamma_1}\mathbf{C}\mathbf{w}=\mathbf{u}+\mathbf{v},\\ &\mathbf{A}+\hat{\mathbf{A}}\preceq\lambda(\gamma_2\hat{\boldsymbol{\Sigma}}+\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)+2\sqrt{\gamma_1}\sum_{i=1}^d w_i\mathbf{C}\mathbf{e}_i\hat{\boldsymbol{\mu}}^T\Big\},\end{aligned}\tag{2.35}$$

and $CC^T=\hat{\boldsymbol{\Sigma}}$ *is a Cholesky-decomposition.*

ii) Assume the conditions in Theorem 2.12 are met. With probability at least $1-\delta$ *with respect to the sampling,* $\mathcal{U}_{DY}$ *implies a probabilistic guarantee for* $\mathbb{P}^*$ *at level* ϵ*.*

The proof utilizes techniques from [97] to compute the worst-case violation probability $\sup_{\mathbb{P}\in\mathcal{P}^{DY}}\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x}>t)$ and then a technique from [55] to convert this probability into a bound on the Value at Risk. See Appendix A.1.5.

2.6.3 Solving Robust Optimization Problems over $\mathcal{U}^{CS}$ and $\mathcal{U}^{DY}$

Using the formula for the support function of $\mathcal{U}^{CS}$ (cf. Eq. (2.33)) we can reformulate robust linear constraints as second order cone constraints. For $\mathcal{U}^{DY}$, we can use strong conic duality to reformulate robust linear constraints over $\mathcal{U}^{DY}$ as a linear matrix

inequality. Generally speaking, second order cone constraints are computationally simpler than LMIs and may be preferred.

We note that $\mathcal{U}^{CS}$ can in general be strengthened by intersecting it with $\text{supp}(\mathbb{P}^*)$. The complexity of this constraint depends on the form of $\text{supp}(\mathbb{P}^*)$. When $\text{supp}(\mathbb{P}^*)$ is second order cone representable, we can still reformulate robust constraints over $\mathcal{U}^{CS} \cap \text{supp}(\mathbb{P}^*)$ as second order constraints using standard techniques.

2.6.4 Connections to Hypothesis Testing

In this section, we provide a new perspective on the above methods in light of hypothesis testing. This alternative perspective provides a common ground on which to compare the methods and will motivate our use of the bootstrap and Gaussian approximation.

To this end, consider the hypotheses:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0,\ \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \ \text{ vs.} H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,\ \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0, \tag{2.36}$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are given constants. From Thm. 2.9, the test which rejects the null hypothesis at level δ if $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|_2 > \Gamma_1(\delta/2, N)$ or $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0\|_F > \Gamma_2(\delta/2, N)$ is a valid test. The confidence region of this test is

$$\begin{aligned}\mathcal{P}^{CS} = \Big\{ \mathbb{P} \in \Theta(R) : &\mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}, \quad \mathbb{E}^{\mathbb{P}}[\tilde{u}\tilde{u}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \boldsymbol{\Sigma}, \\ &\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \Gamma_1(\delta/2, N), \quad \|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_F \leq \Gamma_2(\delta/2, N) \Big\},\end{aligned}$$

where $\Theta(R)$ is the set of Borel probability measures on the ball of radius R. By construction, $\sup_{\mathbb{P} \in \mathcal{P}^{CS}} \text{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x})$ is given by the right-hand side of Eq. (2.33).

A similar interpretation applies to $\mathcal{P}^{DY}$. Indeed, consider the test which rejects H_0 above if $(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}) > \gamma_1$ or $\boldsymbol{\Sigma}_0 + (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T \not\preceq \gamma_2 \hat{\boldsymbol{\Sigma}}$. By Thm. 2.12, this is a valid test at level δ since if H_0 is true, $\boldsymbol{\Sigma}_0 + (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T \not\preceq \gamma_2 \hat{\boldsymbol{\Sigma}} \iff \mathbb{E}^{\mathbb{P}^*}[(\tilde{\mathbf{u}} - \hat{\boldsymbol{\mu}})(\tilde{\mathbf{u}} - \hat{\boldsymbol{\mu}})^T] \not\preceq \gamma_2 \hat{\boldsymbol{\Sigma}}$.

This correspondence is not unique to these two methods. There is a one-to-one

correspondence between families of measures which contain $\mathbb{P}^*$ with probability at least $1-\delta$ with respect to the sampling and the confidence regions of hypothesis tests. This correspondence is sometimes called the "duality between confidence regions and hypothesis testing" in the statistical literature [99]. It implies that any data-driven method predicated on a family of measures that contain $\mathbb{P}^*$ with probability $1-\delta$ can be interpreted in the light of hypothesis testing.

We feel this observation is interesting for two reasons. First, it unifies several distinct methods in the literature, and, in particular, ties them to the well-established theory of hypothesis testing in statistics. Secondly, and in our opinion, most importantly, there is a wealth of practical experience using hypothesis tests. We know empirically which tests are best suited to various applications and which tests perform well even when the underlying assumptions on $\mathbb{P}^*$ that motivated the test may be violated. In the next section, we leverage some of this practical experience with hypothesis testing to refine some of our earlier constructions.

2.7 Refining Sets via the Bootstrap and Gaussian Approximations

The hypothesis tests for the mean and covariance introduced in the previous section are not typically used in applied statistics. These tests have low *power*, i.e., they often require a great deal of data to correctly reject the null-hypothesis when it is false. Two common approaches in applied statistics to addressing this issue are bootstrapping and Gaussian approximation. We next show how these two approaches can be used to refine the sets $\mathcal{U}^{CS}$, $\mathcal{U}^{DY}$ and $\mathcal{U}^{FB}$.

2.7.1 Bootstrapped versions of $\mathcal{U}^{CS}$ and $\mathcal{U}^{DY}$

Our review of bootstrapping is necessarily brief. See [54] or [83, Chapt. 15] for a more thorough treatment. Loosely speaking, in bootstrapping, the distribution of a statistic under $\mathbb{P}^*_{\mathcal{S}}$ is approximated by an empirical distribution formed by repeatedly

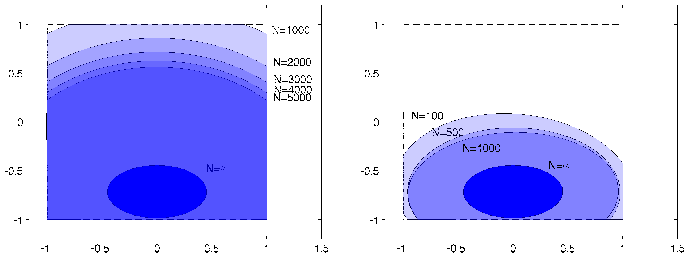
sampling from the data (with replacement) and calculating the statistic on each of these samples. This empirical distribution can then be used to form a confidence region (resp. hypothesis test) for the statistic. This confidence region (resp. hypothesis test) often performs comparably to the best possible confidence region (resp. hypothesis test) had we known $\mathbb{P}^*_{\mathcal{S}}$ exactly. (See [83, p. 15.4] for more precise statements.)

As an example, consider the thresholds $\Gamma_1(\delta, N), \Gamma_2(\delta, N)$ in the definition of $\mathcal{U}^{CS}$. Instead of computing them as in Thm. 2.9, we consider computing them via the bootstrap as follows. We first compute $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ from $\mathcal{S}$. We then construct N_B distinct bootstrap samples. Each bootstrap sample consists of N points drawn with replacement from the original data set. Using the j^{th} sample, we compute its mean $\hat{\boldsymbol{\mu}}_j$, covariance $\hat{\boldsymbol{\Sigma}}_j$ and the two statistics $\Gamma_{1j} \equiv \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_j\|$ and $\Gamma_{2j} \equiv \|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_j\|_F$. Finally, we define $\Gamma_1^B(\delta, N)$ as the $\lceil N_B * \delta \rceil$ largest value among Γ_{1j}, $j = 1, \ldots, N_B$ and similarly for $\Gamma_2^B(\delta, N)$.

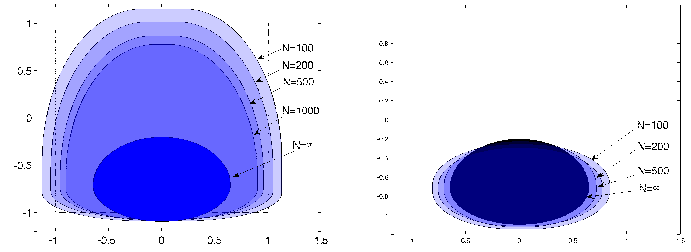
Standard results from the theory of the bootstrap ensure that under mild conditions, the hypothesis test that rejects the null hypothesis in Eq. (2.36) if $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|_2 > \Gamma_1^B(\delta/2, N)$ or $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0\|_F > \Gamma_2^B(\delta/2, N)$ is a valid test at level δ as $N \to \infty$. For finite N, the significance level is approximately δ. Computational experience suggests the approximation is *extremely* good. Moreover, the thresholds $\Gamma_1^B(\delta, N), \Gamma_2^B(\delta, N)$ are often *much* smaller than the corresponding thresholds $\Gamma_1(\delta, N), \Gamma_2(\delta, N)$. Consequently, the set $\mathcal{U}^{CS}$ with bootstrapped thresholds will be much smaller than the original thresholds, but will satisfy the same probabilistic guarantee.

We illustrate this with a numerical example in Fig. 2-4. The true data are generated by the same distribution $\mathbb{P}^*$ as in Fig. 2-3. On the top left, we show the set $\mathcal{U}^{CS}$ with the thresholds $\Gamma_1(\delta, N), \Gamma_2(\delta, N)$. Notice that for $N = 1000$, the set is almost as big as the full support and shrinks slowly to its infinite limit. On the top right, we show the set same set using the bootstrapped thresholds with $N_B = 10,000$. The bootstrapped set with $N = 100$ points is smaller than the non-bootstrapped version with 50 times as many points.

The above argument can be adapted to $\mathcal{U}^{DY}$ to compute bootstrapped thresholds in an entirely analogous manner. A benefit of computing bootstrapped thresholds



(a) U^{CS} Original (b) U^{CS} Bootstrapped



(c) $\mathcal{U}^{FB}$ Original (d) $\mathcal{U}^{FB}$ Bootstrapped

Figure 2-4: The sets $\mathcal{U}^{CS}$ and $\mathcal{U}^{FB}$ for different amounts of data N with and without bootstrapped thresholds.

in this case is that the resulting sets are well-defined for all values of N, unlike the original thresholds γ_1, γ_2.

We stress that although, strictly speaking, hypothesis tests constructed using the bootstrap are only approximately valid, they are routinely used throughout the applied literature with great success, even with N as small as 100. Consequently, we believe practitioners can safely use bootstrapped thresholds in the above sets.

2.7.2 Refining $\mathcal{U}^{FB}$

Another common approach to hypothesis testing in applied statistics is to use tests designed for Gaussian data that are "robust to departures from normality." The best known example of this approach is the t-test from Sec. 2.2.1, for which there is a great deal of experimental evidence to suggest that the test is still approximately valid when the underlying data is non-Gaussian [83, Chapt. 11.3]. Moreover, certain nonparametric tests of the mean for non-Gaussian data are asymptotically equivalent to the t-test, so that the t-test, itself, is asymptotically valid for non-Gaussian data [83, p. 180]. Consequently, the t-test is routinely used in practice, even when the

Gaussian assumption may be invalid.

We next use the t-test in combination with bootstrapping to refine $\mathcal{U}^{FB}$. We replace m_{fi}, m_{bi} in Eq. (2.27), with the upper and lower thresholds of a t-test at level $\delta'/2$. We expect these new thresholds to correctly bound the true mean μ_i with probability approximately $1-\delta'/2$ with respect to the data. We then use the bootstrap to calculate bounds on the forward and backward deviations σ_{fi}, σ_{bi}. Specifically, using the j^{th} bootstrap sample, we compute the statistics $\sigma_{fi}^j, \sigma_{bi}^j$ by evaluating Eq. (2.22) replacing the expectation over $\mathbb{P}_i^*$ with the sample average over the bootstrap sample. Define σ_{fi}^B as the $\lceil N_B(1-\delta'/2)\rceil$ largest value among σ_{fi}^j, $j = 1, \ldots, N_B$ and define σ_{bi}^B similarly. We expect that $\sigma_{fi}^B, \sigma_{bi}^B$ will upper bound the true forward and backward deviations with probability approximately $1 - \delta'/2$.

Refining the thresholds for $\mathcal{U}^{FB}$ in this way makes a substantial difference. For a numerical example see the bottom panels of Fig. 2-4.

We stress not all tests designed for Gaussian data are robust to departures from normality. Applying Gaussian tests that lack this robustness will likely yield poor performance. Consequently, some care must be taken when choosing an appropriate test.

2.8 Guidelines for Practitioners

From a practitioner's point of view, the most critical question is: *"Which set should I use to model my problem?"* Choosing the right set requires striking a balance between faithfully modeling the underlying randomness and the tractability of the resulting model. Based on our preliminary computational experience, we offer some guidelines as to which sets to use in various modeling scenarios for various computational budgets. As experience with these sets in real applications grows and computing power increases, we expect to revisit these guidelines.

When working with discrete distributions, $\mathcal{U}^{\chi^2}$ should be preferred to $\mathcal{U}^G$. As discussed in the text, these sets are nearly identical for large N, but it is substantially easier to solve robust optimization problems over $\mathcal{U}^{\chi^2}$. When working with distri-

butions with independent marginals with a moderate computational budget, the set $\mathcal{U}^{FB}$ should be preferred to $\mathcal{U}^{I}$. Although $\mathcal{U}^{FB}$ is a superset of $\mathcal{U}^{I}$, it often not too much larger and the computational savings are substantial. Finally, when working with distributions with potentially correlated marginal distributions we recommend using $\mathcal{U}^{CS}$ for its simplicity over $\mathcal{U}^{DY}$. Finally, wherever possible, we advocate using bootstrapped thresholds. In our computational experiments, the gains from bootstrapping are typically significant.

2.9 Applications

We demonstrate how our new sets may be used in two applications: portfolio management and queueing theory. Our goals are to, first, emphasize the practical application of these sets and, second, to compare them to one another. We summarize our major insights.

- As expected, our data-driven sets are smaller than their traditional counterparts, and they continue to shrink as more data becomes available.
- Our sets are able to learn features of $\mathbb{P}^*$ like correlation structure and skewness directly from the data.
- Finally, solutions of robust models built from our sets are stable with respect to perturbations in the data resulting from random sampling.

2.9.1 Portfolio Management

Portfolio management has been well-studied in the robust optimization literature [e.g., 65, 89]. For simplicity, we will consider the one period allocation problem:

$$\max_{\mathbf{x}} \left\{ \min_{\mathbf{r}\in\mathcal{U}} \quad \mathbf{r}^T\mathbf{x} : \quad \mathbf{e}^T\mathbf{x} = 1, \quad \mathbf{x} \geq \mathbf{0} \right\}, \tag{2.37}$$

which seeks the portfolio $\mathbf{x}$ with maximal worst-case return over the set $\mathcal{U}$. If $\mathcal{U}$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ, then the optimal value z^* of this

optimization is a conservative bound on the ϵ-worst case return for the optimal solution $\mathbf{x}^*$.

We consider a synthetic market with $d = 10$ assets. Returns are generated according to the following factor model:

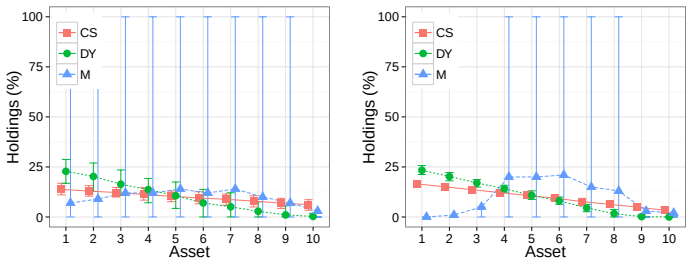
$$\tilde{r}_i = \beta_i \tilde{z} + \tilde{\zeta}_i, \quad \beta_i \equiv \frac{i-1}{9}, \quad i = 1, \ldots, 10, \tag{2.38}$$

where $\tilde{z}$ is a common market factor that follows a normal distribution with mean 3% and standard deviation 5% truncated at ± 3 standard deviations, and $\tilde{\zeta}_i$ is an idiosyncratic contribution following a normal distribution with mean 0% and standard deviation 5% truncated at ± 3 standard deviations. The random variables $\tilde{z}, \tilde{\zeta}_i$ are mutually independent.

This setup mirrors the classical Capital Asset Pricing Model. The common market factor correlates the assets. Lower indexed assets have lower returns and are less correlated to the market. Moreover, lower indexed asset have smaller support. Specifically, the support of Asset 1 is approximately [-15%, 15%], while the support of Asset 10 is approximately $[-26\%, 31\%]$.

In the absence of data, the only uncertainty set which guarantees a probabilistic guarantee is the support of $\mathbb{P}^*$. Using this set in Eq. (2.37) yields a portfolio which invests all its wealth in the first asset since this asset has the largest lower bound on its support.

Using Eq. (2.38), we simulate $N = 120$ historical monthly returns and use these data to construct the uncertainty sets $\mathcal{U}^M$, $\mathcal{U}^{CS} \cap \text{supp}(\mathbb{P}^*)$ and $\mathcal{U}^{DY}$ with $\epsilon = 10\%$ and $\delta = 20\%$. We use bootstrapped thresholds for the latter two. We then solve Eq. (2.37) for each of these sets and record the optimal allocation $\mathbf{x}^*$ and the objective value z^*. We also compute the true 10% worst-case return for each of these allocations and then repeat this procedure 100 times. Figure 2-5a shows the average holdings for each method with a 90% confidence interval, and Table 2.2 shows summary statistics for z^* and the true, out-of-sample 10% worst-case return for each method. We also include the average out-of-sample return and the statistics for the "No Data" set in



(a) $N = 120$ (b) $N = 1,200$

Figure 2-5: Each panel shows the holdings by asset for each method for a single-factor market (cf. Eq. (2.38)). The error bars indicate a 90% confidence interval over the 100 simulations. For comparison, the "No data" set (not shown) holds 100% of the first asset in both cases.

Table 2.2: Results for single-factor market (cf. Eq. (2.38)). Values in %.

	$N = 120$			$N = 1,200$		
Set	Avg. z^*	$\text{VaR}_{10\%}$	Expected Return	Avg. z^*	$\text{VaR}_{10\%}$	Expected Return
No Data	-15.0	-6.4	0.0	-15.0	-6.4	0.0
$\mathcal{U}^M$	-6.9	-5.8	1.3	-6.1	-5.9	1.4
$\mathcal{U}^{CS}$	-8.9	-2.1	1.1	-7.2	-2.1	1.0
$\mathcal{U}^{DY}$	-8.4	-2.3	0.7	-7.9	-2.4	0.7

Table 2.2 for comparison.

Since $\mathcal{U}^M$ does not use the joint distribution, it sees no benefit to diversification. It consequently invests all of its wealth in whichever asset appears to have the best worst-case quantile given the data. Since it is difficult to estimate this quantile with only 120 data points, the asset it chooses varies greatly depending on the particular run. $\mathcal{U}^{DY}$ and $\mathcal{U}^{CS}$ (with bootstrapping) are both able to learn the covariance structure and consequently diversify across the assets. $\mathcal{U}^{CS}$ holds a more evenly diversified portfolio than $\mathcal{U}^{DY}$. Unfortunately, as can be seen, since $N = 120$ is a relatively small number of data points, the value of z^* is a very conservative estimate of the 10% worst-case return for each of these sets.

To better understand the effect of using more data, we repeat the above experiment with $N = 1,200$ data points. Admittedly, this amount of data is rarely available in typical applications of portfolio allocation. Nonetheless, the results are shown in

Figs. 2-5b and Table 2.2. With the added data $\mathcal{U}^M$ is able to more accurately learn the quantiles of the marginal distributions, and identify that the best, single asset is one of Assets 5, 6, 7, or 8 depending on the run. $\mathcal{U}^{CS}$ and $\mathcal{U}^{DY}$ hold very similar portfolios as they did with $N = 120$, although the error bars are considerably smaller. Finally, the quality of the bound z^* on the worst-case return in Table 2.2 is improved for all of our sets.

2.9.2 Queueing Analysis

Recently, [10] proposed a robust optimization approach to analyzing queueing networks. Their method yields *approximations* to a desired performance metric, such as the waiting time. In this section, we combine our data-driven uncertainty sets with their methodology to generate upper bounds on these performance metrics that satisfy *probabilistic guarantees*. For concreteness, we will focus on the waiting time in a G/G/1 queue and use our set $\mathcal{U}^{FB}$. Similar results can be derived for our other sets. Extending this analysis to more complex queueing networks is an open question, but likely can be accomplished along the lines in [10].

Let $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ denote the service times of the first n customers in a queue, and let $\tilde{\mathbf{T}} = (\tilde{T}_1, \ldots, \tilde{T}_n)$ denote the interarrival times. We assume that $\tilde{X}_i$ (resp. $\tilde{T}_i$) is i.i.d. for all i and that the service times and interarrival times are independent. Moreover, we have bounds $\overline{X}, \overline{T}$ such that $0 \leq \tilde{X}_i \leq \overline{X}$ and $0 \leq \tilde{T}_i \leq \overline{T}$ almost surely. Let $\hat{x}_1, \ldots, \hat{x}_N, \hat{t}_1, \ldots, \hat{t}_N$ be drawn from these service and interarrival distributions respectively. We compute $m_{fX}, \sigma_{fX}, m_{bT}, \sigma_{bT}$ from Eq. (2.24). Recall, these quantities depend on the significance level δ. We use them to form the set $\mathcal{U}^{FB}$ at level ϵ.

From Lindley's recursion [84], the waiting time of the n^{th} customer is

$$\tilde{W}_n = \max_{1 \leq j \leq n} \left(\max \left(\sum_{l=j}^{n-1} X_l - \sum_{l=j+1}^{n} T_l, 0 \right) \right) = \max \left(0, \max_{1 \leq j \leq n} \left(\sum_{l=j}^{n-1} X_l - \sum_{l=j+1}^{n} T_l \right) \right). \tag{2.39}$$

The optimizing index j represents the last customer to arrive when the queue is empty. Eq. (2.39) holds path by path; using the data, we can derive a similar recursion that

holds with high probability. Let $\tilde{N}$ denote the number of customers served in a typical busy period. From the sequences $\hat{x}_1, \ldots, \hat{x}_N$ and $\hat{t}_1, \ldots, \hat{t}_N$, we compute the number of customers served in each busy period of the queue, denoted $\hat{n}_1, \ldots, \hat{n}_K$, which are i.i.d. realizations of $\tilde{N}$. Using the KS test at level δ', we observe that with probability $1 - \delta'$ with respect to the sampling,

$$\mathbb{P}(\tilde{N} > \hat{n}^{(k)}) \leq 1 + \Gamma^{KS}(\delta') - \frac{k}{K} \equiv \epsilon'. \tag{2.40}$$

In other words, the queue empties every $\hat{n}^{(k)}$ customers with probability at least $1-\epsilon'$.

Motivated by [10], we consider a worst-case realization of a Lindley recursion truncated at $\hat{n}^{(k)}$, namely

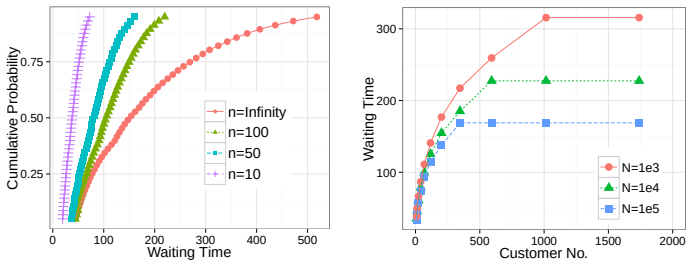
$$W_n^{Rob} \equiv \max\left(0, \max_{1\leq j\leq n_{min}} \max_{(\mathbf{x},\mathbf{t})\in\mathcal{U}^{FB}} \left(\sum_{l=j}^{n_{min}-1} x_l - \sum_{l=j+1}^{n_{min}} t_l\right)\right), \tag{2.41}$$

where $n_{min} \equiv \min(n^{(k)}, n)$. The inner optimization can be solved in closed-form. Using Thm. 2.6 and Eq. (2.26),

$$\begin{aligned}
&\max_{1\leq j\leq n_{min}} \max_{(\mathbf{x},\mathbf{t})\in\mathcal{U}} \left(\sum_{l=j}^{n_{min}-1} x_l - \sum_{l=j+1}^{n_{min}} t_l\right) \\
&\qquad = \max_{1\leq j\leq n_{min}} (m_{fX} - m_{bT})(n_{min} - j) + \sqrt{2\log(1/\epsilon)(\sigma_{fX}^2 + \sigma_{bT}^2)}\sqrt{n_{min} - j} \\
&\qquad \leq \max_{0\leq z\leq n_{min}} (m_{fX} - m_{bT})z + \sqrt{2\log(1/\epsilon)(\sigma_{fX}^2 + \sigma_{bT}^2)}\sqrt{z},
\end{aligned}$$

where the last optimization follows from the transformation $z = n_{min} - j$ and relaxing the integrality on j. Examining the first order conditions for this last optimization yields

$$W_n^{Rob} \leq \begin{cases} (m_{fX} - m_{bT})n_{min} + \sqrt{2\log(\frac{1}{\epsilon})(\sigma_{fX}^2 + \sigma_{bT}^2)}\sqrt{n_{min}} & \text{if } n_{min} < \frac{\log(\frac{1}{\epsilon})(\sigma_{fX}^2+\sigma_{bT}^2)}{2(m_{bT}-m_{fX})^2} \\ & \text{or } m_{fx} > m_{bT}, \\ \frac{\log(\frac{1}{\epsilon})(\sigma_{fX}^2+\sigma_{bT}^2)}{2(m_{bT}-m_{fX})} & \text{otherwise.} \end{cases} \tag{2.42}$$



(a) CDF of $\tilde{W}_n$ (b) Median Waiting Time

Figure 2-6: The left panel presents bounds on the cumulative distribution function of $\tilde{W}_n$ for $n = 100, 1000, \infty$. The right panel presents our upper bound on the median of the steady-state waiting time given N data points.

The qualitative form of the solution matches our intuition for the stochastic system. Indeed, if the mean service time exceeds the mean interarrival time ($m_{fX} > m_{bT}$), then the system is unstable. Otherwise, there exists a relaxation time $\frac{\log(1/\epsilon)(\sigma_{fX}^2+\sigma_{bT}^2)}{2(m_{bT}-m_{fX})^2}$ before which the waiting time grows linearly, and after which the waiting time converges to a constant which has the same functional form as the well-known Kingman bound on the mean waiting time [80].

Most importantly, since $\mathcal{U}^{FB}$ was chosen to imply a probabilistic guarantee, we can interpret Eq. (2.42) as a guarantee on a quantile of the waiting time distribution of the n^{th} customer. Specifically, note that, conditioned on the sampling, Eq. (2.41) holds with probability at least $1-\epsilon'$, and, by reversing the two maximzations, there are at most $n^{(k)}$ linear functions of the uncertainty on the righthand side. Since $\mathcal{U}^{FB}$ implies a probabilistic guarantee at level ϵ, the probability that any one of these linear functions exceeds W_n^{Rob} is at most ϵ. From a union bound, it follows that Eq. (2.42) holds with probability at least $1-\epsilon'-n^{(k)}\epsilon$. Thus, Eq. (2.42) represents a bound on the $1-\epsilon'-n^{(k)}\epsilon$ quantile of the waiting time distribution of the n^{th} customer. Moreover, this analysis holds for any choice of ϵ, ϵ', so that by varying these parameters we can obtain bounds on the entire waiting time distribution for the n^{th} customer, both in transient and steady-states.

We illustrate these ideas numerically. Service times are distributed as a Pareto distribution with parameter 1.1, and the interarrival times are distributed as an exponential distribution with rate 3.92. Both distributions are then truncated at their

95^{th} percentiles, i.e., approximately 15.23 and 11.74, respectively. The resulting truncated distributions have means of approximately 3.39 and 3.72, respectively, yielding an approximate 90% utilization.

As a first experiment, we simulate $N = 10^5$ services and arrivals from these distributions and then use Eq. (2.42) to compute W_n^{Rob} for various choices of ϵ_1 and ϵ_2 to bound the quantiles of $\tilde{W}_n$ for $n = 10, 50, 100, \infty$. The resulting bounds are shown in the left panel of Fig. 2-6. We have used $\delta = \delta' = 10\%$.

As a second experiment, we look at the dependence of our bound on the amount of data. For varying values of N, we use Eq. (2.42) to compute an upper bound on the median of the waiting for the n^{th} customer. The resulting bounds are shown in the right panel of Fig. 2-6. For reference, the true median as $n \to \infty$ (obtained by simulation) is approximately 24. As can be seen, as the amount of data available increases, the upper bound on the median waiting time improves.

2.10 Conclusion

In this chapter, we took a first step towards adapting traditional robust optimization techniques to this new paradigm. Specifically, we proposed a novel schema for designing uncertainty sets for robust linear optimization from data using hypothesis tests. Sets designed using our schema imply a probabilistic guarantee and are typically much smaller than corresponding data poor variants. Models built from these sets are thus less conservative than conventional robust approaches, yet retain the same robustness guarantees.

Chapter 3

Data-Driven Approaches to Unit Commitment

3.1 Introduction

This chapter presents a modification of our data-driven uncertainty set technique to adaptive optimization and an in-depth application to the unit commitment (UC) problem in electricity markets. Throughout the chapter we present our results in the context of a real data set drawn from the New England electricity market to emphasize their practicality. (See Sec. 3.4 for a discussion of the data used.)

As discussed in the introduction, UC is a significant, high-impact problem in power systems operations. In words, the objective of UC is to choose a dispatch schedule – i.e., which generators to activate and how to operate them to produce electricity – that minimizes operational costs while maintaining a near 100% service level for consumers. (We give a precise formulation of the problem in Section 3.2.) Instances of UC are routinely solved in power-system design and operations. The New England Independent System Operator (NE ISO), for example, uses a UC formulation as the basis of its real-time market-clearing mechanism to determine the price of electricity. A slightly different variant is used to direct its capacity expansion and transmission investment programs. Finally, UC is also used routinely in reliability analysis of the power grid to assess its robustness to potential disruptions and mechanical failures.

These applications have substantial scope. There are over 6.5 million households and businesses receiving electricity from NE ISO's power grid. In 2012, the aforementioned real-time energy market cleared $5 billion dollars in revenue. Moreover, there is an additional $5 billion slated for investment upgrading transmission lines between 2013 and 2018. [41]. Consequently, even small improvements to existing UC algorithms can yield large impacts.

Unsurprisingly, the problem has garnered much research attention (see [93] for a survey). Following [36, 116], we model UC as a two-stage adaptive optimization problem which can be formulated as a robust, mixed-binary linear optimization problem (See Sec. 3.2). This type of formulation has been well-motivated in the literature, including the need for incorporating uncertainty and the rationale for using adaptive optimization (see, e.g., [113] for a discussion of mixed-integer formulations, and [36] for motivation of the adaptive formulation). Consequently, we take this particular formulation as *given* in what follows. Instead, we focus our efforts, first, on adapting our uncertainty set constructions from Chapt. 2 to this adaptive optimization context and, second, on providing an efficient method for computing a near-optimal affinely adaptive policy.

With respect to the first goal, we observe that electricity demand, or *load*, is not drawn i.i.d. from a distribution. Rather, it is a time-series, exhibiting seasonality, non-stationarity and significant day-on-day autocorrelations (see Sec. 3.4.2). Consequently, our previous constructions cannot be applied "out of the box." Rather, we propose a suitable, simple modification to those techniques to apply them to time-series data. Our modification is versatile and can be tailored as appropriate to a variety of modeling situations.

With respect to our second goal, it is well-known that computing the optimal adaptive policy to an adaptive optimization problem can be NP-hard [15]. Consequently, many authors focus on fixing simple parametric functional forms for later stage decisions to ensure tractability [15, 64]. An important class of parametric forms are affinely adaptive policies, i.e., policies in which later stage decisions are affine functions of the uncertainties. Affinely adaptive solutions have been proven to be

optimal or near optimal for some classes of problems [26, 30, 77], and practically effective for many others [16]. In what follows, we, too, concentrate on computing affinely adaptive solutions.

Unfortunately, even when restricting to the class of affine policies, solving our UC formulation can be computationally prohibitive for the large-scale instances frequently encountered in practice. Motivated by principal components analysis in statistics, we propose a subset of these policies, which we call *projected affine policies*. Projected affine policies are affine in a low-dimensional projection of the uncertainty. When the underlying data also has a low-dimensional structure (as load does), an optimal projected affine policy can be computed very efficiently with only a small loss in optimality to the full affine policy. We illustrate this idea numerically with our load data.

As mentioned, we are not the first to propose using an adaptive optimization approach to model UC. [36, 116] both proposed adaptive formulations for UC. Both works focus on computing the optimal, fully-adaptive policy. [36] uses a combination of a general purpose Bender's decomposition and a vertex enumeration scheme to compute a policy, while [116] uses a more tailored decomposition. Moreover, both works employ a relatively simple, uncertainty set that does not heavily leverage data. Arguably, this simplicity is critical to computing the optimal adaptive policy efficiently.

By contrast we consider computing a suboptimal policy (i.e., the affine policy), but for a more sophisticated, data-driven uncertainty set. Extensive computational backtesting in the New England market confirms that our approach yields high quality solutions with relatively little computational time. In some sense, then, a message of this chapter is that combining suboptimal policies with data-driven uncertainty sets is a viable alternative to computing optimal policies for computationally simpler sets.

We summarize our contributions as follows:

- We adapt the constructions of Chapt. 2 to deal with vector, time-series data which are not drawn i.i.d. Our proposal is simple, versatile and computationally tractable.

- We use principal components analysis as a dimensionality reduction to motivate the class of projected affine policies and show that these policies perform almost as well as fully-affine policies for load data with data-driven uncertainty sets.

- We apply these ideas to an extensive numerical case study based on real data for generators and load taken from NE ISO. Overall, we show through numerical experiments that solving the projected affinely adaptive optimization problem over our data-driven sets can be done tractably, and, in out-of-sample testing with historical data, produces high-quality solutions.

The structure of the remainder of this chapter is as follows. In Sec. 3.2 we provide mathematical formulations of the nominal, robust and affinely adaptive UC problem. In Sec. 3.3 we discuss adapting our uncertainty set constructions to time series data. Sec. 3.4 summarizes the data from NE ISO that will be used in our experiments, and Sec. 3.5 applies the previous technique to this data to construct several sets. Sec. 3.6 discusses some algorithmic details and, importantly, introduces projected affinely adaptive policies. Finally, Sec. 3.7 presents the numerical results of our case-study.

Before proceeding, we remark that while it is common in the operations management literature to use the terms "supply" and "demand," the more common terms in the power systems literature are "generation" and "load," respectively, which we adopt.

3.2 Formulation

We now give a precise formulation of the UC problem we treat. We are given a set $\mathcal{G}$ of generators (indexed by $g \in \mathcal{G}$), and a planning horizon $1, \ldots, H$, indexed by h. (In applications, the planning horizon is typically 24 hours.) For each generator g, we are given parameters describing its operating characteristics, namely:

Ecomin $\underline{p}_{gh}$ The minimum amount of electricity in MWh that g can produce (if on) in hour h.

Ecomax $\overline{p}_{gh}$ The maximum amount of electricity in MWh that g can produce (if on) in hour h.

Minimum down hours $\underline{d}_g$ The minimum number of hours g must remain off once it is stopped.

Minimum up hours $\overline{d}_g$ The minimum number of hours g must remain on and producing electricity once it is started.

Maximum Starts $\overline{z}_g$ The maximum number of times g can be started over the horizon.

Start-up cost K_g The dollar cost incurred for each start of g.

Offer curve $c_{gh}(\cdot)$ A function $c_{gh} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $c_{gh}(t)$ represents the dollar cost of g producing t MWh of electricity in hour h. This function is typically piecewise-linear and convex.

With the exception of the offer cost curve, these parameters depend on technical specifications of each generator, and, hence, they are typically known to the system operator and not subject to uncertainty. In deregulated markets like NE ISO, the offer curves c_{gh} are updated each day by each generator and represent their offers into the electricity auction.

Rather, the primary source of uncertainty we will consider is load. Namely, let $\tilde{u}_h \in \mathbb{R}_+$ represent uncertain load at hour h, and $\tilde{\mathbf{u}} \in \mathbb{R}^H$ denote the corresponding vector for the horizon. At the beginning of the horizon, $\tilde{\mathbf{u}}$ is unknown, although we may have some forecast for it denoted by $\mathbf{f}(\tilde{\mathbf{u}})$ ("f" is for *forecast*). (We discuss forecasts more fully in Section 3.5.1.) There are other sources of uncertainty in UC, e.g., generator failure and variable production from wind farms. The effects of these uncertainties are small in NE ISO and are ignored in what follows.

3.2.1 Nominal Formulation

We are now in a position to formulate our nominal optimization problem. Let $x_{gh} \in \{0, 1\}$ denote a binary decision variable indicating whether or not unit g is on at hour

h. Similarly, let p_{gh} represent the amount of electricity produced by generator g in hour h. Finally, let h^- denote hour $h-1$. Our nominal formulation of UC is given by:

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{p}} \quad \sum_{g\in\mathcal{G}}\sum_{h=1}^{H}(K_g z_{gh} + t_{gh}) + \kappa s$$

$$\text{s.t.} \quad x_{gh} - x_{gh^-} \le z_{gh}, \quad x_{gh^-} - x_{gh} \le y_{gh}, \quad \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1a}$$

$$z_{gh} \le 1 - x_{gh^-}, \quad y_{gh} \le x_{gh^-}, \quad \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1b}$$

$$z_{gh} \le x_{gh'}, \quad h \le h' < h + \overline{d}_g,\ \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1c}$$

$$y_{gh} \le 1 - x_{gh'}, \quad h \le h' < h + \underline{d}_g,\ \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1d}$$

$$\sum_{h=1}^{H} z_{gh} \le \overline{z}_g,\ \forall g \in \mathcal{G} \tag{3.1e}$$

$$c_{gh}(p_{gh}) \le t_{gh},\ \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1f}$$

$$\sum_{h=1}^{H}\left| f_h(\tilde{u}) - \sum_{g\in\mathcal{G}} p_{gh}\right| \le s \tag{3.1g}$$

$$\underline{p}_{gh} x_{gh} \le p_{gh} \le \overline{p}_{gh} x_{gh} \quad \forall g \in \mathcal{G},\ h = 1,\dots,H, \tag{3.1h}$$

To simplify the notation, we have adopted the convention that a constraint whose indices extend beyond the range $h = 1,\dots,H$ is considered vacuous. For example, constraint (3.1c) is meant to be omitted when $h > H - \overline{d}_g$. The exogenous parameter κ controls the tradeoff between the immediate, first-stage operating costs, and the delayed second-stage mismatch costs. We discuss this constant more fully in Sec. 3.2.4.

The binary variables $y_{gh}, z_{gh} \in \{0,1\}$ represent if generator g stops (resp. starts) in hour h. Constraints (3.1a), (3.1b) define the relationship between x_{gh}, y_{gh} and z_{gh}. Constraint (3.1c) (resp. (3.1d)) captures the minimum up hours (resp. minimum down hours) requirement on the unit. Constraint (3.1e) constrains the total number of starts over the horizon. Constraint (3.1f) is an epigraph constraint for $c_{gh}(t)$. Since c_{gh} is piecewise-linear, convex, this constraint can easily be rewritten as a series of linear constraints using standard linear optimization techniques. The same holds for constraint (3.1g). (See, e.g., [34].) Finally, Constraint (3.1h) constrains the unit

to operate within the ecomin and ecomax parameters (defined above) if on. With these constraints, notice that nominal UC can be solved as a mixed binary linear optimization problem.

NE ISO solves a UC problem nearly identical to Problem (3.1) as part of its market-clearing procedures and reliability analysis with three noteworthy differences. First, they incorporate some additional operating constraints on generators, e.g., "ramping constraints" which control how quickly a generator can change its production from hour to hour. Second, they use the *reserve methodology* which requires that a certain amount of additional generation be available on top of what is needed to serve the load at each hour. The precise amount of additional generation is determined by a proprietary heuristic. Finally, and most significantly, Problem (3.1) neglects the underlying network structure. In reality, generators and consumers are not co-located, but rather dispersed geographically and connected via the transmission network. When dispatching generators, the system operator must ensure that there exists a feasible flow on this network to ship electricity. Feasibility of a flow is determined by the transmission line capacity constraints and Kirchoff's laws.

Recall our goal in this chapter is not to study the precise mechanism used by NE ISO, but rather to illustrate how our data-driven techniques can be applied to an important, real-world problem. Consequently, we will continue to make these three simplifications going forwards. Incorporating them needlessly complicates the exposition, but can be done in a conceptually simple manner by linearizing appropriately. (See, e.g., [27, 36, 113].) Nonetheless, we still consider Problem (3.1) to be a reasonable approximation to the current methodology used by NE ISO for solving UC problems and use it as a benchmark in our computational studies.

3.2.2 Robust Formulation

Notice that Problem (3.1) is formulated with the forecasted loads $\mathbf{f}(\tilde{\mathbf{u}})$. A natural robust formulation then is to replace these forecasts by uncertain loads $\mathbf{u} \in \mathcal{U}$ for some choice of uncertainty set $\mathcal{U} \subseteq \mathbb{R}^H_+$. Notice this change only affects constraint (3.1g),

yielding

$$\sum_{h=1}^{H} \left| u_h - \sum_{g \in \mathcal{G}} p_{gh} \right| \leq s \quad \forall \mathbf{u} \in \mathcal{U} \tag{3.2}$$

This robust constraint is a sum of maxima of linear functions. Recall that our analysis in Chapter 2 dealt exclusively with robust linear constraints.

[66] discuss robust constraints which are sums of maxima of linear functions in detail and offer several reformulations of various computational complexity and conservativeness for treating them. The simplest approach parallels the nominal formulation. Namely, we replace eq. (3.2) by the constraints

$$\begin{aligned} u_h - \sum_{g \in \mathcal{G}} p_{gh} &\leq s_h \quad \forall \mathbf{u} \in \mathcal{U},\ h = 1, \ldots, H, \\ u_h - \sum_{g \in \mathcal{G}} p_{gh} &\geq -s_h \quad \forall \mathbf{u} \in \mathcal{U},\ h = 1, \ldots, H, \\ \sum_{h=1}^{H} s_h &\leq H \end{aligned} \tag{3.3}$$

[66] proves that in general this type of "LP"-style reformulation is conservative approximation of the original robust constraint, and, in some instances, can be very conservative. On the other hand, this type of reformulation is by far the most common one in the applied robust optimization literature. We will adopt this simple "LP"-style reformulation and will call the problem obtained by replacing eq. (3.1g) by eqs. (3.3) the *Robust UC* problem, or RUC.

Given the simple structure of eqs. (3.3), we can solve the robust counterpart explicitly:

Proposition 3.1. *Consider the robust constraints* (3.2). *Let* $B(\mathcal{U}) = \{\mathbf{u} \in \mathbb{R}^H_+ : \underline{\mathbf{u}} \leq \mathbf{u} \leq \overline{\mathbf{u}}\}$ *be the smallest bounding box containing* $\mathcal{U}$. *Then, constraints* (3.2) *are equivalent to the same constraints with* $B(\mathcal{U})$ *replacing* $\mathcal{U}$, *and, furthermore, these*

constraints are equivalent to the (non-robust) constraints

$$\begin{aligned}
\overline{u}_h - \sum_{g\in\mathcal{G}} p_{gh} &\leq s_h \quad h = 1,\dots,H,\\
\underline{u}_h - \sum_{g\in\mathcal{G}} p_{gh} &\geq -s_h \quad h = 1,\dots,H,\\
\sum_{h=1}^{H} s_h &\leq H.
\end{aligned}$$

Proof. The key observation is that the uncertainties decouple by coordinate over the constraints. Specifically, consider the first set of constraints for hour h. The worst case is attained at $\arg\max_{\mathbf{u}\in\mathcal{U}} \overline{u}_h$ which equals $\overline{u}_h$ by definition. The second set of constraints is similar. □

A consequence of Proposition 3.1 is that solving RUC is no harder than solving nominal UC. Indeed, existing nominal solvers can be adjusted very slightly to solve the corresponding robust problem.Proposition 3.1 also invites a potential criticism of RUC. Namely, regardless of the initial choice of uncertainty set $\mathcal{U}$, this formulation is equivalent to solving the same problem replacing $\mathcal{U}$ by $B(\mathcal{U})$. In other words, any dependence information that $\mathcal{U}$ might contain about the coordinates of $\tilde{\mathbf{u}}$ is ignored.

3.2.3 Affinely Adaptive Formulation

Notice that both the nominal and robust formulations do not distinguish between first and second stage decision variables. By contrast, adaptive formulations for UC *do* distinguish between these variables, allowing the second stage variables p_{gh} to depend on the realized load $\mathbf{u}$; i.e., $p_{gh}(\mathbf{u})$. In what follows, we focus on affine policies $p_{gh}(\mathbf{u}) = \boldsymbol{\ell}_{gh}^T\mathbf{u} + \ell_{gh0}$ for some decision variables $\boldsymbol{\ell}_{gh} \in \mathbb{R}^H, \ell_{gh0} \in \mathbb{R}$ for all $g \in \mathcal{G}$, $h = 1,\dots,H$.

Affine adaptability only affects constraints (3.1f), (3.1g), and (3.1h). In the case

of (3.1g), e.g., we obtain:

$$d_h - \sum_{g \in \mathcal{G}} p_{gh}(\mathbf{d}) \leq s_h \quad \forall \mathbf{d} \in \mathcal{U},\ h = 1, \ldots, H,$$
$$d_h - \sum_{g \in \mathcal{G}} p_{gh}(\mathbf{d}) \geq -s_h \quad \forall \mathbf{d} \in \mathcal{U},\ h = 1, \ldots, H,$$
$$\sum_{h=1}^{H} s_h \leq H$$

Notice that these constraints no longer separate coordinate-wise by uncertainty. Hence, a result like Proposition 3.1 no longer applies. In other words, this formulation will take advantage of any dependence structure enforced by the set $\mathcal{U}$ between the coordinates. On the other hand, the above constraints (after expanding the functions $p_{gh}(\mathbf{d})$) are robust linear constraints in the decision variables. We can reformulate these constraints using standard techniques.

The other sets of constraints (3.1f) and (3.1h) can be treated similarly; first rewrite them as a system of linear equations using standard linear programming techniques, then, expand the adaptive variables (retaining the linear structure), and, finally, robustify each constraint separately. We omit the details. We call such a formulation our *Affine* UC, or AUC.

3.2.4 Specifying κ

In each of the above formulations, there are two types of costs in the objective: operating costs and mismatch costs. Operating costs are measured in dollars and consist of the (fixed) startup costs and the (variable) cost to produce electricity for each unit. These costs are easily quantifiable and are actually paid by consumers to the suppliers in the electricity market.

By contrast, the mismatch costs are not as easily quantifiable. Mismatches between generation and load in the second stage are naturally measured in GWh. The constant κ (in units \$/GWh) converts these mismatches into mismatch costs (in units \$). The constant κ represents costs due to potential load-shedding (i.e., unmet de-

mand), turning on rapid-start, fuel-inefficient generators to try to cover shortfalls in production, increased C02 emissions from these fuel-inefficient generators, and capital depreciation from operating units outside of their technical specifications to better match demand. Many of these costs are not explicitly paid in normal market operations. Rather they are implicitly borne as negative externalities in the system. It is this feature that makes them difficult to quantify exactly.

The magnitude of κ controls the tradeoff between these two types of costs, and it is debatable what an appropriate value should be. In our numerical experiments, we adopt the value \$5M/ GWh as used by NE ISO in their daily operations.

3.3 Data-Driven Uncertainty Sets for Time Series

Both RUC and AUC require specifying an uncertainty set $\mathcal{U}$ for load. Given a data set $\mathcal{S} = \{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_N\}$ of historical loads, one approach to forming such a data-driven uncertainty sets would be to directly apply the formulas of the previous chapter. In the specific case of load, we expect such an approach to perform poorly.

The issue is that the data $\mathcal{S}$ does not truly represent an i.i.d. sample from some unknown distribution $\mathbb{P}^*$, but rather a time-series which we emphasize by re-indexing it as $\mathcal{S} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_T)$. As a time-series, the mean and variability of $\tilde{u}_t$ may change over time, and it may exhibit significant autocorrelation for different t. (The interested reader may want to skip ahead to Sec. 3.4.2 to see these effects as they pertain to load data from NE ISO.) There are a multitude of classical techniques for analyzing time series models, many of which involve making fairly stringent assumptions about the data generating process (see., e.g., [45] for an overview.) We propose a different approach that requires fewer assumptions by leveraging a forecasting model.

By a *forecasting model*, we mean any deterministic mechanism which uses only information available at time t to form a prediction of $\tilde{\mathbf{u}}_{t+1}$. We make no assumptions on the particular structure of the forecasting model, only that its inputs are known at time t. Numerous techniques exist in time-series analysis, statistics and machine learning for constructing such models. For example, one might use nonparametric

regression to fit a function that predicts $\tilde{\mathbf{u}}_{t+1}$ from features known at time t. (Indeed, we follow this approach in Sec. 3.5.1.) Alternatively, one might fit an ARMA process to the data $\mathcal{S}$, and then use extrapolation to predict future elements [46].

We extend our previous notation and let $\mathbf{f}_t(\tilde{\mathbf{u}}_{t+1}) \in \mathbb{R}_+^H$ denote the forecast of $\tilde{\mathbf{u}}_{t+1}$ computed at time t. We now describe our proposal for constructing uncertainty sets for time series data based on $\mathcal{S}$ and a forecasting model $\mathbf{f}(\cdot)$. To simplify the notation, we will concentrate on the case where we seek to construct an uncertainty set for $\tilde{\mathbf{u}}_{T+1}$ for an optimization to be used at time T.

1. Using $\mathcal{S}$, form the residual forecast errors $\mathcal{S}' = \{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_T - 1\}$ defined by

$$\hat{\mathbf{r}}_{t+1} \equiv \hat{\mathbf{u}}_{t+1} - \mathbf{f}_t(\tilde{\mathbf{u}}_{t+1}), \quad t = 1, \ldots, T-1.$$

2. Apply any of the data-driven uncertainty set construction from Chapt. 2 to $\mathcal{S}'$ to construct an uncertainty set $\mathcal{U}_r$ for these residuals.

3. Let $\mathcal{U} = \{\mathbf{f}_T(\tilde{\mathbf{u}}_{T+1}) + \mathbf{r} : \mathbf{r} \in \mathcal{U}_r\}$.

We stress that in contrast to other proposals for constructing uncertainty sets from statistical models such as [65, 117], our proposal is agnostic to the particular form of the forecasting model. This feature makes the approach versatile. It can be applied in conjunction with any of the aforementioned statistical techniques. Moreover, the forecasting model only serves to translate the set $\mathcal{U}$, and, consequently, does not affect the computational complexity of separating robust constraints over $\mathcal{U}$. This complexity is entirely determined by the complexity of separating over $\mathcal{U}_r$ which can be controlled by choosing an appropriate construction in Step 2.

We also observe that if a forecasting model is not given a priori, but T is large, one can simply split the initial data $\mathcal{S}$ in half, use the first half to train a forecasting model using any desired statistical algorithm, and then proceed with the above procedure using the second half of the data.

We next outline the intuition behind above scheme. We focus on the special case that, under true data generating distribution, each component of $\tilde{\mathbf{u}}_t$ is given by the

time-series

$$\tilde{u}_{ht} \equiv s_h(t) + \sum_{i=1}^{p} \theta_{hi} \tilde{u}_{h,t-i} + \tilde{\xi}_{ht}, \tag{3.4}$$

where $s(t)$ is a deterministic periodic function with period $L < \infty$, and $\tilde{\xi}_{ht}$ are an i.i.d. innovation process. In other words, each component of $\tilde{\mathbf{u}}_t$ is the sum of a seasonality term and autoregressive process of order p. We further assume for simplicity that the forecasting mechanism is constructed by training separate nonparametric regressions for each h of $\hat{u}_{ht}$ on the predictors $(t, \hat{u}_{h,t-1}, \ldots, \hat{u}_{h,t-\tau})$, for $t = \tau + 1, \ldots, T$ and $\max(L, p) < \tau$. Finally, we assume the uncertainty set in Step 2 is constructed to imply a probabilistic guarantee at level ϵ.

Then, given eq. (3.4), $\mathbb{E}_t[\tilde{u}_{h,t+1}] = s_h(t+1) + \sum_{i=1}^{p} \theta_{hi} \tilde{u}_{h,t-i}$, where $\mathbb{E}_t$ refers to expectation conditional on time t. In other words, the conditional expectation is a fixed measurable function of the input data to the random forest. Standard results on random forests (see, e.g., [37]) show then that for each h,

$$\mathbb{E}\left[(f_{ht}(\tilde{\mathbf{u}}_{t+1}) \to E_t[\tilde{u}_{h,t+1}])^2\right] \to 0, \quad \forall t \quad \text{as } T \to \infty.$$

Thus, for large enough T, $\mathcal{S}'$ approximates an i.i.d sample of the innovation process $\tilde{\boldsymbol{\xi}}_t$. If $\mathcal{S}'$ were truly i.i.d., the results of Chapt. 2 would imply that $\mathcal{U}_r$ implies a probabilistic guarantee for $\tilde{\boldsymbol{\xi}}$ at level ϵ. In reality, because $\mathcal{S}'$ is only approximately i.i.d., $\mathcal{U}_r$ will imply a probabilistic guarantee at level slightly greater than ϵ. Finally, note that for large enough T

$$\begin{aligned} \tilde{\mathbf{u}}_{T+1} &= \mathbb{E}_T[\tilde{\mathbf{u}}_{T+1}] + (\tilde{\mathbf{u}}_{T+1} - \mathbb{E}_T[\tilde{\mathbf{u}}_{T+1}]) \\ &\approx \mathbf{f}_t(\tilde{\mathbf{u}}_{T+1}) + \xi_{T+1}. \end{aligned}$$

Therefore, $\mathcal{U}$ also implies a probabilistic guarantee for $\tilde{\mathbf{u}}_T$ at a level slightly greater than ϵ.

The specific form in eq. (3.4) was not crucial to the argument. Rather, the key requirement was that the innovations $\tilde{\mathbf{u}}_{t+1} - \mathbb{E}_t[\tilde{\mathbf{u}}_{t+1}]$ were (approximately) representable by a fixed, measurable function of the data in the regression. Many time-series models

have this property, including ARIMA and ARMAX models and their nonlinear counterparts. At least intuitively, this suggests the above approach should be practically effective in many circumstances.

3.3.1 Tuning Uncertainty Sets in Adaptive Optimization

There remains the question of to what level ϵ (and/or δ) should be set in the above construction in Step 2. From the formulations of RUC and AUC, there is no obvious choice for either parameter. These problems are two-stage optimization problems with full-recourse, not chance-constrained problems or single-stage optimization problems with uncertain data. In principle, infeasibility is impossible as we have full recourse.

We propose an alternative approach based on cross-validation for selecting these parameters in two-stage optimization problems with full-recourse. Specifically, for any specified value of ϵ, we can apply the above construction, and solve our optimization for some day in our test set. After fixing the first stage decisions, we can observe historically how the uncertainty for that day realized, and re-solve the second stage problem with the new realized uncertainty, fixing those first-stage decisions. This process mirrors real-time operations had we used our model to schedule first-stage dispatch, and yields some total cost. Embedding this procedure within a cross-validation scheme, we can estimate the expected total cost from our model (over the test set) for that choice of ϵ. (Some type of cross-validation scheme is required to account for in-sample bias.) Choosing the optimal ϵ can be done via an exhaustive grid search.

We remark that cross-validation can often be simplified by reparametrizing the uncertainty set construction. For example, instead of specifying the parameters ϵ, δ in the set $\mathcal{U}^M$ in eq. (2.32), one can alternatively specify the constant s, or the ratio s/N directly. Similarly, instead of specifying ϵ, δ in the set $\mathcal{U}^{CS}$ in eq. (2.34), one can alternatively specify Γ_1 and Γ_2 and ϵ.

Table 3.1: Generator Composition. Capacity refers to capacity at 12:00 pm (Hour ending 13). CT stands for "Combustion Turbine" and CC stands for "Combined Cycle."

	Original Data		Filtered Data		Test Instance	
Type	Number	Capacity (GWh)	Number	Capacity (GWh)	Number	Capacity (GWh)
Steam	115	20.62	108	20.32	10	1.78
CT	90	5.13	81	4.99	13	0.79
CC	3	0.12	3	0.12	3	0.12
Diesel	29	0.15	23	0.13	2	0.02
Hydro	65	0.83	-	-	-	-
Nuclear	5	4.67	5	4.67	5	4.67
Wind	1	-	-	-	-	-

3.4 NE ISO Data Overview

Our numerical case-study is based on data taken from NE ISO. In this section, we describe this data. Portions of the dataset are publicly available while others are confidential and were provided to us by the ISO.[1]

3.4.1 Generator Data

NE ISO contains over 300 different generators. The precise number differs slightly day to day as generators are taken out of service for maintenance. We were provided operating characteristics for 308 generators in the region. In addition, we are given the cost curves for each of these generators for some particular (anonymized) day. This data is not publicly available.

Of the 300 generators, we removed 88 generators from the dataset due to questionable data entries. The remaining 220 generators constitute approximate 96% of the peak generation capacity. See Table 3.1 for details.

In some of our particularly computationally intensive experiments, we restrict ourselves to an even smaller test-instance of generators, also shown in Table 3.1. These 33 generators constitute 24% of the initial peak generation capacity.

[1]In the spirit of reproducible research, we are working with the ISO to provide appropriately anonymized versions of these confidential datasets for publication.

In our experiments to follow, we scale the load data by 96% (resp. 24%) when using the filtered set (resp. test instance) of generators.

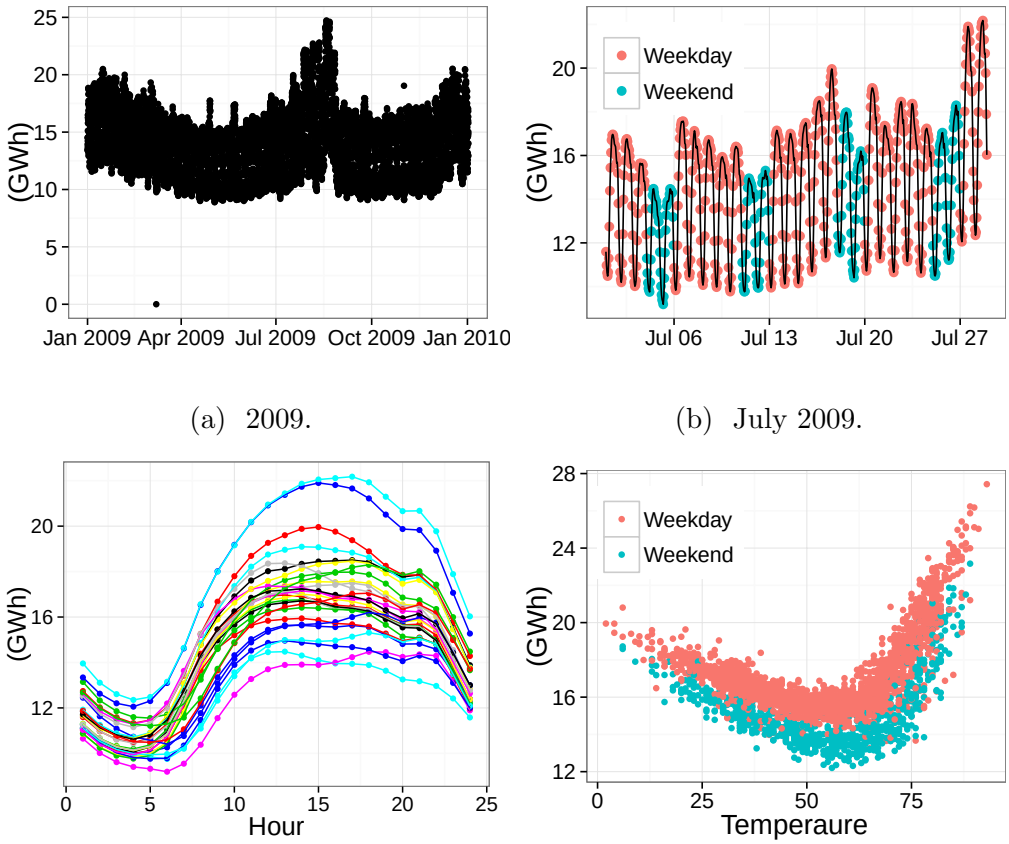
3.4.2 Load Data

System and nodal level load data is publicly available from the NE ISO [75]. We restrict ourselves to load data from 1 March 2002-28 Feb. 2014. Hourly loads for 2009, hourly loads for July 2009 and the daily profile of hourly loads for weekdays in July 2009 are shown in the first 3 panels Figure 3-1. Some features are immediately evident:

- Yearly Periodicity: Load is periodic, peaking in both the winter and summer.
- Weekday/Weekend Effects: Load during weekdays is typically larger than load on weekends.
- Daily Profiles: Although daily profiles follow a similar shape from day-to-day, there is still considerable variability, especially in the peak. Hour ending 15 (3:00 pm), e.g., can spike as much as 4.5 GWh from its median value.
- Smoothness: Hourly profiles are generally smooth.

Beyond its periodic features, load also demonstrates a well-documented dependence on weather, most prominently temperature and humidity. Intuitively, when it is very hot outside, consumers turn on their air-conditioning. When it is cold, they use electric heaters. These behavioral responses drive a large portion of electrical usage in NE ISO.

This phenomenon suggests that local, hourly temperature would be a good predictor of local energy usage. Obtaining local temperature information throughout New England, however, can be daunting. To simplify models and analysis, however, NE ISO typically uses weighted combination of hourly dry-bulb (resp. wet-bulb) temperatures from 8 large cities (Boston, MA; Bridgeport, CT; Burlington, VT; Concord, NH; Portland, ME; Providence, RI; Windsor Lakes, CT; and Worcester, MA) to rep-



(a) 2009.

(b) July 2009.

(c) Daily profiles.

(d) Load vs. Temperature.

Figure 3-1: Total (system) load in NE-ISO. Panel 3-1c plots only weekdays from July 2009. Panel 3-1d plots the dependence for hour ending 13 (12:00 pm) only.

resent the "average" temperature across New England.[2] Fig. 3-1d shows system load versus this average dry-bulb temperature in hour ending 13 (12:00 pm). A strong functional dependence is visible. In what follows, we will also use the hourly values of this weighted combination of dry-bulb (resp. wet-bulb) temperatures at these cities in our predictions.

We stress that these load data do not appear i.i.d. day-on-day, but rather demonstrate many of the aforementioned time-series characteristics discussed in Sec. 3.3.

3.5 Constructing Uncertainty Sets for NE ISO

We next follow the strategy outlined in Sec. 3.3 to fit an uncertainty set to this data. For the purposes of out-of-sample validation, we divide the data into a training set (1 March 2002 to 1 Aug. 2006), a test set (2 Aug. 2006 to 1 July 2010) and a validation set (2 July 2010 to 28 Feb. 2014). The training set will be used to construct our forecasting model, the test set to form our uncertainty set and tune parameters, and the validation set to generate out-of-sample results.

3.5.1 Forecasting Model

There are an immense number of models in the power systems literature of varying sophistication for forecasting load. Some authors (e.g., [76] propose models based on the periodic and autoregressive features in load data, while others [49, 108] propose models based on dependence on weather data. Most modern forecasting models blend both sources of information (e.g. [57, 74, 95]). Moreover, each year since 2012 the Kaggle has hosted the Global Energy Forecasting Competition in conjunction with the IEEE Power and Energy Systems. The contest has still more sophisticated models with higher forecast accuracy [63].

In what follows, we use a very simple forecasting model based on temperature dependence in order to streamline the exposition. Admittedly, more sophisticated

[2]Dry-bulb temperatures refer to temperatures taken with an ordinary thermometer. Wet-bulb temperatures refer to temperatures taken with a thermometer wrapped in a wet-cloth. The difference in these temperatures is a measure of humidity.

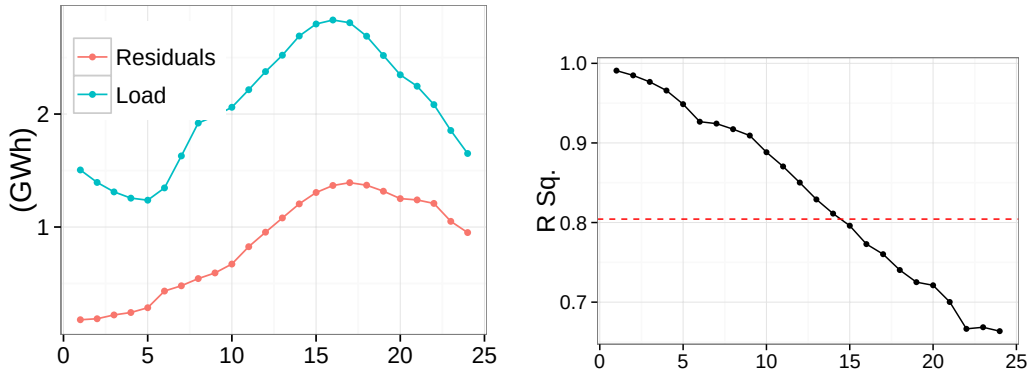
models might produce better forecasts. Specifically, we first split the training data by season (Winter: December to February, Spring: March to May, Sumer: June to August and Fall: September to November). Isolating the seasons helps control for the yearly seasonality effects observed in Fig. 3-1a. We then separate weekdays from weekends, again to control for the weekday effect in Fig. 3-1b. Finally, for each season and each type of day, we fit 24 separate random forest regressions to predict the load in each hour based on the 24 hourly dry temperature readings and 24 hourly wet-bulb temperature readings from the *previous* day. We emphasize that we use the previous days temperatures, because, as discussed in Sec. 3.3 we must build a forecasting model that only uses information up to the beginning of today.

The goodness-of-fit metrics below focus on the random forests corresponding to weekdays in the summer because these days will be most pertinent to our case-study in Sec. 3.7. There are 241 observations in each of the training, testing and validation sets when limited to weekdays in the summer. Fig. 3-2a presents the residual standard deviation and the standard deviation of the full load, by hour, computed over the test set. (Recall that the training set was used to construct the random forest, so that this figure represents out-of-sample results.) Fig. 3-2b presents the residual R^2 by hour. Clearly, hours later in the day are predicted less accurately, because we limit ourselves to information available at time by the beginning of the day. Nonetheless, even for these hours our forecasting mechanism reduces the variability by about 65%. In the earlier hours, the variability is reduced by about 98%

3.5.2 Uncertainty Set

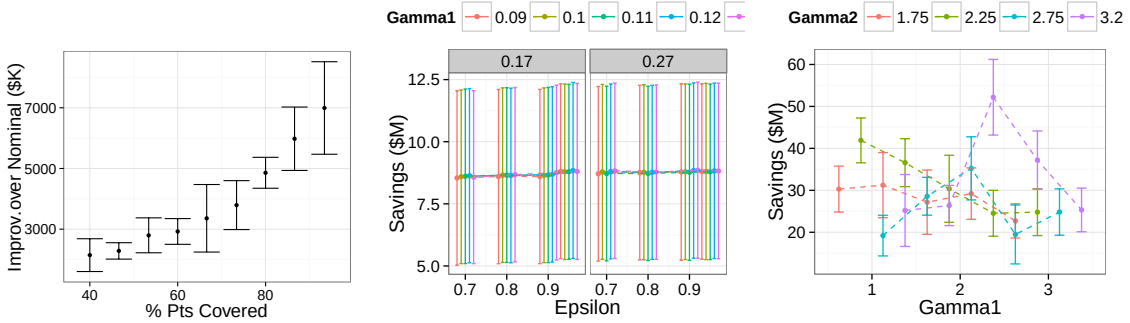
In the case of Robust UC, Proposition 3.1 proves that it suffices to consider box uncertainty, like U^M (cf. eq. (2.32)). Consequently, we use this set in Step 2 of our construction, and tune the parameter s directly by 5-fold cross-validation on the test set. The results are presented in Fig. 3-3 in terms of the rescaled parameter $1-2s/N$, i.e., the percentage of data points covered by the box in each coordinate direction. Based on these results, we fix a value of $1 - 2s/N = 93\%$ for our experiments.

In the case of Affine UC, we have more freedom in choosing $\mathcal{U}$. We will use $\mathcal{U}^{CS}$



(a) Hourly standard deviations. (b) R^2 by hour for our forecasting model.

Figure 3-2: Forecasting model goodness of fit. In the right hand panel, the dotted red line represents the average R^2 across all weekday, summer observations in the test set.



(a) Robust (b) U^{CS} (c) U^{B}

Figure 3-3: Cross-validation results for each method. For the robust method, we use 5-fold cross-validation on the full test set. For the affine methods, we use a jackknife procedure on 10 representative days described below.

(cf. eq. (2.34)) in our experiments for two reasons. First, empirically, the residuals on our test set do not appear to be independent hour by hour. Secondly, compared to $\mathcal{U}^{DY}$ (cf. eq. (2.35)), $\mathcal{U}^{CS}$ is substantially more tractable and admits a closed-form separation oracle.

We then need to tune the parameters $\epsilon, \Gamma_1, \Gamma_2$ in the construction. 5-fold cross-validation is particularly computationally expensive for this example. Instead, we use a jackknife procedure on a limited test set of 10 well-chosen days to tune these parameters. Specifically, we first perform k-means clustering on our forecasts over the test set with $k = 10$. We treat the historical data point closest to the center of each cluster as a prototypical day for that cluster. Finally, for a fixed choice of $\epsilon, \Gamma_1, \Gamma_2$ and each of these 10 prototypical days, we fit the set $\mathcal{U}^{CS}$ excluding that day and then compute the total cost of dispatch on that day. We seek the combination of $\epsilon, \Gamma_1, \Gamma_2$ that minimizes the mean performance over these 10 samples. Partial results are summarized in Fig. 3-3b. The difference between various choices are small, so we have only included results for $\Gamma_2 = .17, .27$. In our experiments, we use the parameters $\epsilon = .1, \Gamma_1 = .124$, and $\Gamma_2 = .241$ which seem to slightly outperform other choices.

Finally, to provide a more comprehensive comparison to existing robust optimization formulations of UC, we also consider using a budget uncertainty set for the residuals in Step 2 of our procedure. Recall, the budget uncertainty set is defined by

$$\begin{aligned}\mathcal{U}^B = \Bigg\{ \mathbf{u} \in \mathbb{R}^H : &-\Gamma_2 \leq \frac{u_h - \mu_h}{\sigma_h} \leq \Gamma_2, \quad h = 1, \ldots, H, \\ &-\Gamma_1\sqrt{H} \leq \sum_{h=1}^{H} \left| \frac{u_h - \mu_h}{\sigma_h} \right| \leq \Gamma_1\sqrt{H} \Bigg\}.\end{aligned} \tag{3.5}$$

Budget uncertainty is used by [36] in the context of UC, albeit for a fully-adaptive problem with simulated data and without our forecasting or tuning techniques. In our context, we take $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ to be the sample mean and sample standard deviations of the residuals over the test set, and tune the parameters Γ_1, Γ_2. Again, for computational tractability, we use a jackknife procedure on the same 10 points used for $\mathcal{U}^{CS}$. Partial results are shown in Fig. 3-3c. Differences between values are very volatile. We use

the parameters $\Gamma_1 = 6.28$, $\Gamma_2 = 3.0$ for our experiments.

3.6 Solving Affinely Adaptive Problems over $\mathcal{U}^{CS}$ and $\mathcal{U}^{B}$

Notice that $\mathcal{U}^{CS}$ is second-order cone representable. Reformulating the robust constraints in our affine formulation, then, yields a mixed binary second order cone problem. Mixed binary second order cone problems are generally considered more difficult (sometimes significantly more difficult) than mixed binary linear optimization problems.

One approach to addressing this difficulty is to use a polyhedral outer approximation to $\mathcal{U}^{CS}$ In [13], the authors propose a general purpose polyhedral approximation technique for second order cone constraints. The technique uses the projection of an extended formulation of a polyhedron to represent an exponential number of facets with a polynomial number of variables and constraints. Applying this technique to approximate the set $\mathcal{U}^{CS}$ within an approximation factor γ yields a polyhedral representation using approximately $2(H-2)(2v+3)$ variables and $2(H-2)(3v+6)$ constraints where $v = \log_2\left(\frac{\pi}{2\arccos(1/\gamma)}\right)$. (Recall, H is the length of the horizon, typically 24 hours.)

Unfortunately, such an approach is generally computationally infeasible for problems like AUC because they contain an extremely large number of robust constraints. This explosion in the number of robust constraints is typical of affinely adaptive optimization problems. In our case-study, for example, there are

$$\begin{aligned} 24 \text{ hrs } &\times 220 \text{ generators } \times 2 \text{ constraints/ generator / hr} \\ &+ 48 \text{ load balancing constraints} \\ &= 10,608 \text{ robust constraints.} \end{aligned}$$

For even a mild value like $\gamma = \sqrt{2}-1 \approx 41\%$, outer approximating $\mathcal{U}^{CS}$ by this method would require 230 additional variables and 414 additional constraints. Reformulating

each of the above robust constraints, then, would require and addition 2.4 million variables and 4.4 million constraints. This is clearly impractical. Although there exist other, coarser polyhedral approximations to second order cone representable sets, (e.g., the d-norm approximation of [31]), they, too, are generally impractical for AUC.

Consequently, we follow a cut-generation approach using "lazy cuts" as outlined in Chapt. 2. Recently, cut generation approaches have been shown to be competitive if not more effective for certain types of robust MILO problems with ellipsoidal sets [25]. Although $\mathcal{U}^{CS}$ is not an ellipsoid, it is second order cone representable and its support function can be evaluated in closed form making cut generation extremely efficient.

In the case of $\mathcal{U}^{B}$, even though the set is polyhedral, reformulating a robust constraint requires $4H + 2$ additional variables and $2H + 2$ additional constraints, again making reformulation impractical for AUC. Consequently, we also apply cut-generation. Notice that the support function for this set can again be computed closed form (see [25]), making cut generation extremely efficient.

In our experiments, we found that a typical instance of AUC generated about a thousand cuts, making this approach dramatically more efficient than reformulation.

3.6.1 Projected Affine Policies

When considering affine policies, the number of variables in AUC can be prohibitively large. We next introduce the class of projected affine policies, aimed at reducing the number of variables in the formulation.

Many real-world phenomena (including load) exhibit low dimensional structure. In other words, there exists a k dimensional subspace $V \subset \mathbb{R}^H$ with $k < H$ such that the projection error of projecting $\tilde{\mathbf{u}}$ onto V is small. (Recall that $\tilde{\mathbf{u}} \in \mathbb{R}^H$ and typically $H = 24$.) Numerous techniques exist in statistics and compressed sensing for identifying such low-dimensional subspaces, most notably principal components analysis [71]. Intuitively, a good uncertainty set $\mathcal{U}$ for $\tilde{\mathbf{u}}$ should capture this low-dimensional structure, i.e., the projection error of projecting $\mathcal{U}$ onto V should also

be small. We next show how to leverage low dimensional structure in $\mathcal{U}$ to reduce the variable dimension of an affinely adaptive optimization problem. We present the approach in context of AUC, but it can just as easily be applied to other affinely adaptive optimization problems.

Specifically, suppose for a moment that $\mathcal{U} \subset V$. Let $V^\perp \subset R^H$ denote the orthogonal complement of V. Finally let

$$p_{gh}(\mathbf{u}) = \boldsymbol{\ell}_{gh}^T \mathbf{u} + \ell_{gh0}, \quad g \in \mathcal{G}, \quad h = 1, \ldots, H. \tag{3.6}$$

be an arbitrary affine policy. Standard linear algebra yields a decomposition $\boldsymbol{\ell}_{gh} = \boldsymbol{\ell}_{gh}^V + \boldsymbol{\ell}_{gh}^\perp$ with $\boldsymbol{\ell}_{gh}^V \in V$ and $\boldsymbol{\ell}_{gh}^\perp \in V^\perp$. It follows that

$$\begin{aligned} p_{gh}(u) &= (\boldsymbol{\ell}_{gh}^V + \boldsymbol{\ell}_{gh}^\perp)^T \mathbf{u} + \ell_{gh0} \\ &= (\boldsymbol{\ell}_{gh}^V)^T \mathbf{u} + \ell_{gh0} \qquad\qquad \text{since } \mathbf{u} \in V \text{ and } \boldsymbol{\ell}_{gh}^\perp \in V^\perp \ . \end{aligned}$$

Thus, it suffices to consider affine policies such that $\boldsymbol{\ell}_{gh} \in V$. Said another way, letting $\Pi_V \in R^{k \times H}$ denote the orthogonal matrix representing projection onto V, it suffices to consider policies of the form

$$p_g h(\mathbf{u}) = \overline{\boldsymbol{\ell}_{gh}}^T \Pi_V^T \mathbf{u} + \overline{\ell}_{gh0}. \tag{3.7}$$

Here $(\overline{\boldsymbol{\ell}_{gh}}, \overline{\ell}_{gh0}) \in \mathbb{R}^{k+1}$. If k is much smaller than H, this can yield a significant reduction in the variable dimension of our optimization.

In most real applications, however, we will not have $\mathcal{U} \subset V$ for a low dimensional subspace V. A more realistic assumption is that there exists a low-dimensional V such that the projection error $\max_{\mathbf{u} \in \mathcal{U}} \|\mathbf{u} - \Pi_V \mathbf{u}\|^2$ is small. In this case, restricting attention to policies of the form eq. (3.7) incurs some loss of optimality related to the size of the projection error. Depending on the application, the benefit from the reducing the size of the optimization may outweigh this loss in optimality.

We now use this observation to construct a class of projected affine policies from the data. We first consider $\mathcal{U}^{CS}$. Let $\hat{\boldsymbol{\Sigma}}$ denote the sample covariance of our matrix,

and let $\mathbf{v}_1, \ldots, \mathbf{v}_H$ and $\lambda_1, \ldots, \lambda_H$ denote its eigenvectors and eigenvalues, respectively, sorted so that $\lambda_1 \geq \ldots \geq \lambda_H > 0$. For any k, we consider the k dimensional affine policy given by eq. (3.7) with $V = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. It is well known that of all k dimensional subspaces V, $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ minimizes the projection error over the data. Since $\mathcal{U}^{CS}$ was constructed to represent the data, we expect that

$$\max_{\mathbf{u} \in \mathcal{U}^{CS}} \|\mathbf{u} - \Pi_{\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)} \mathbf{u}\|^2$$

should be small as well.

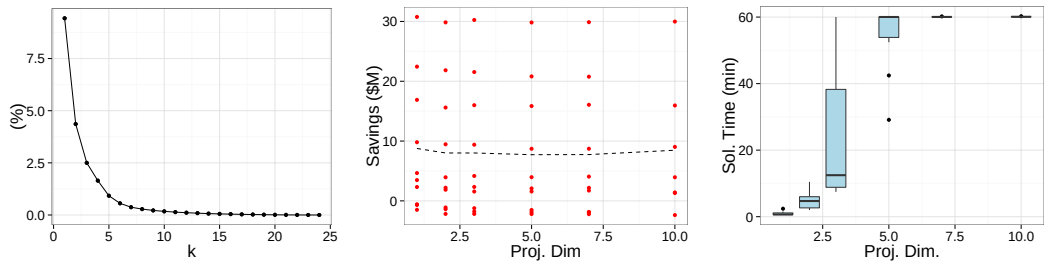
Fig. 3-4 illustrates this idea with our load data. In the first panel, we show the residual (unexplained) variance over the test set when considering the space $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Notice with $k = 1$, we already explain 93% of the variability. Furthermore, for each additional eigenvalue, we must add an addition $5,280$ optimization variables to our formulation for the full generator instance. The second and third panels show the improvement over nominal and solution time for various choices of k over our 10 prototypical days in the test set (cf. Sec. 3.5.2).[3]

The optimal value for $k = 1$ is nearly identical to that for $k = 10$, but solves in a fraction of the computational time. This accords with our observation that a single eigenvalue explains most of the sample covariance matrix. In our experiments in Sec. 3.7 we will use $k = 1$.

Since $\mathcal{U}^B$ is not built from the full covariance matrix of the data, we define projected affine policies slightly differently. Let $H_k \subseteq \{1, \ldots, H\}$ be the set of indices which maximize $\sum_{i \in \mathcal{H}_k} \sigma_i^2$ in eq. (3.5). For any k, we consider the k dimensional affine policy given by eq. (3.7) with $V = \text{span}(\mathbf{e}_i)_{i \in H_k}$. It is not hard to show that of all k dimensional subspaces V, $\text{span}(\mathbf{e}_i)_{i \in H_k}$ minimizes $\max_{\mathbf{u} \in \mathcal{U}^B} \|\mathbf{u} - \Pi_V \mathbf{u}\|^2$.

Fig. 3-5 parallels Fig. 3-4. The first panel of Fig. 3-4 shows $\sqrt{\frac{\sum_{i \notin \mathcal{H}_k} \sigma_i^2}{\sum_{i=1}^H \sigma_i^2}}$ for various k. Notice that we require a significantly larger value of k, such as $k = 15, 20$ to cover a reasonable amount of the variability. The second and third panels show improvement over the nominal method and solution time for our 10 scenarios. Unlike with U^{CS},

[3] All experiments in this section were run using the JuMP modeling language ([86]) in conjunction with Gurobi version 5.6.0 [69] on machine with a a quad-core 3.6 GHz processor and 48 GB RAM.

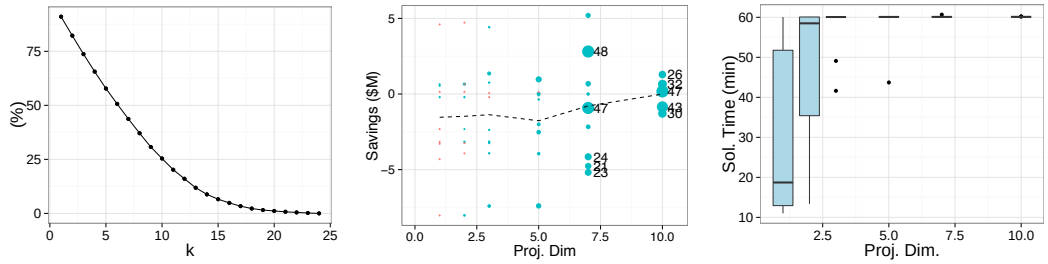


(a) Unexplained variance for $\mathcal{U}^{CS}$.

(b) Savings over nominal for our 10 scenarios.

(c) Solution time (s) for our 10 scenarios.

Figure 3-4: Projected Affine Policies for $\mathcal{U}^{CS}$. The dotted line in the middle panel corresponds to mean. Optimizations were allocated 1 hour of solution time to mimic realtime operations.



(a) Unexplained variance for $\mathcal{U}^{B}$.

(b) Savings over nominal for our 10 scenarios.

(c) Solution time (s) for our 10 scenarios.

Figure 3-5: Projected Affine Policies for $\mathcal{U}^{B}$. The dotted line in the middle panel corresponds to mean. Optimizations were allocated 1 hour of solution time to mimic realtime operations.

not all of the instances were solved to optimality within one hour. Instances that were interrupted after one hour are indicated by blue-green dots (instead of red dots), and the size of dot is proportional to the optimality gap at that time. Gaps that exceed 20% are labeled in the plot. Observe that for large values of k, we are essentially unable to find provably high-quality solutions within the timeframe. This poses are difficult tradeoff. Since $\mathcal{U}^{B}$ does not represent the data as closely as $\mathcal{U}^{CS}$, projected affine policies with small values of k may be very suboptimal, however, we cannot reliably compute good solutions with larger values of k. To keep the size of the optimization problem similar to the $\mathcal{U}^{CS}$ case, we will take $k = 1$ in our study in Sec. 3.7.

3.7 Case-Study: UC in the NE ISO

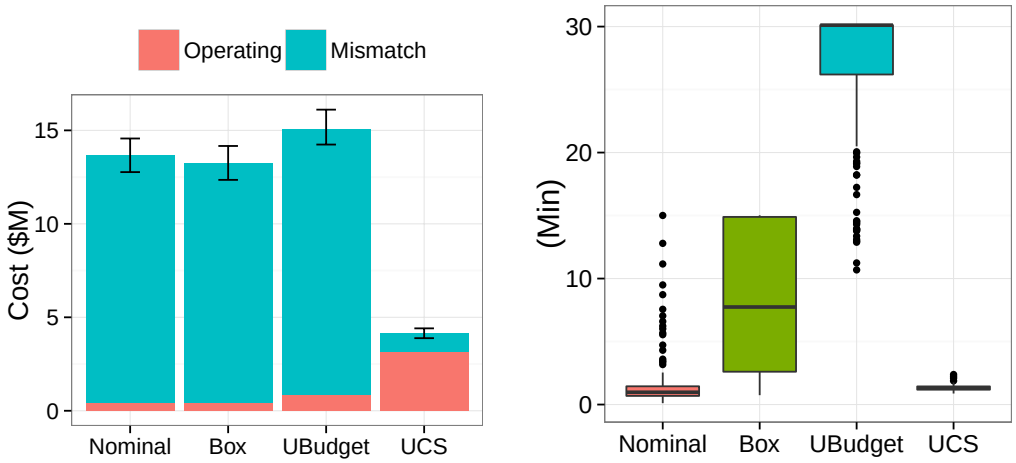
We now proceed to our main numerical study. For each day in the validation set, we solve nominal, robust and affine UC problems using our forecasts. We use the set $\mathcal{U}^M$ (box) for the robust model, and both $\mathcal{U}^{CS}$ and $\mathcal{U}^B$ for our affine models. After fixing the first stage decisions according to these models, we resolve the second stage problem using the actual, realized load on that day historically with these fixed decisions. We repeat this procedure for all 241 days in our validation set. As discussed in Sec. 3.2.4, we fix the value of $\kappa = \$5000/MWh$ (as is used at NE ISO).

We first present the solution times for each method.[4] To mimic realtime operations, we allow 1 hour of computation for each method, and at the conclusion of the 1 hour, use the best feasible solution found so far. Fig. 3-6b presents the results. Notice that the $\mathcal{U}^{CS}$ method requires significantly less time than the other methods. Although it is difficult to ascribe a precise reason for this difference, intuitively, our projected affine policies effectively reduce the variable dimension and isolate which variables are most critical to the optimization (i.e., which ones correspond to large directions of uncertainty). These features likely significantly speed-up the branch and bound process.

Next, we present the total costs of each method – broken down by startup costs, production costs and load shedding – in Table 3.2. The histogram for each type of costs and load mismatches are also presented in Fig. 3-7. A first observation is that while the other methods incur a similar startup cost initially dispatching units, the $\mathcal{U}^{CS}$ method incurs significantly larger costs (cf. Fig. 3-7a). Perhaps surprisingly, this difference in startup cost is not only caused by the $\mathcal{U}^{CS}$ method dispatching more units, but rather, by its choosing to dispatch more *flexible*, expensive units.

To clarify, see Fig. 3-8. Here we have broken down the initial dispatch decision by fuel-type for hour ending 1 (12:00 AM) and hour ending 13 (12:00 PM) for 2 July 2010 for both the $\mathcal{U}^{CS}$ method and the robust method over $\mathcal{U}^M$ (the box). Hour ending 1 (12:00 AM) is a fairly non-volatile hour for which we can predict load with high

[4] In contrast to the previous section, all experiments in this section were run using on a machine with a a quad-core 3.0 GHz processor and 8 GB RAM.



(a) (b)

Figure 3-6: The left panel shows the total cost broken down by operating costs and mismatch costs. Notice that in all cases, mismatch costs dominate. The right panel shows the solution times by method.

Table 3.2: The average costs ($K), broken down by source, and average load mismatch (GWh) for each method over the validation set. Standard errors are presented in parenthesis below each cell.

Method	Startup Cost ($K)	Production Cost ($K)	Mismatch (GWh)	Total Cost ($K)
Nominal	18.8	409.1	2.65	13,666
	(0.4)	(10.0)	(0.18)	(899)
Box	17.7	419.3	2.56	13,259
	(0.4)	(9.8)	(0.18)	(907)
UBudget	30.6	840.0	2.86	15,175
	(0.6)	(10.2)	(0.19)	(934)
UCS	238.7	2,928.9	0.20	4,146
	(0.6)	(43.2)	(0.05)	(263)

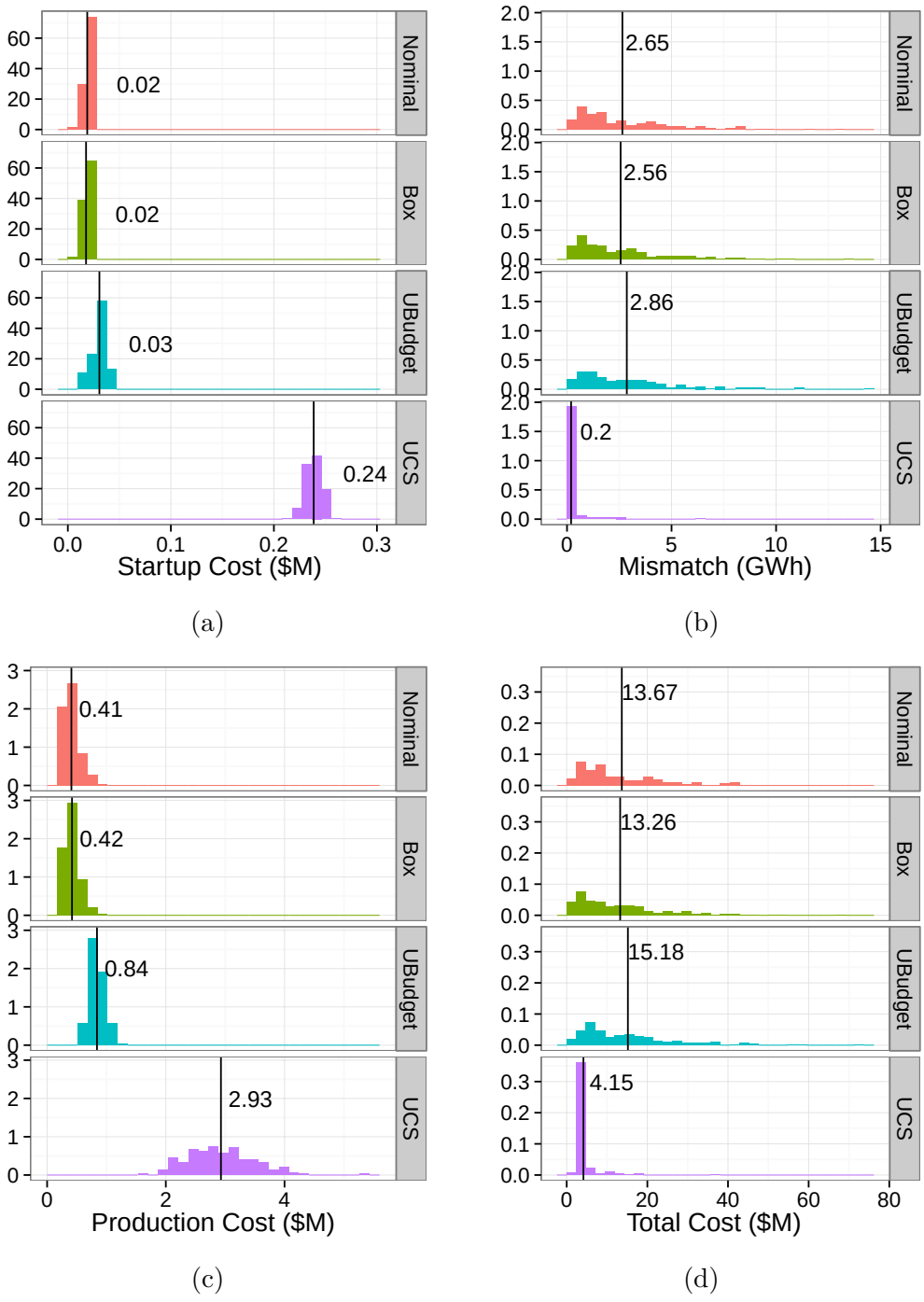
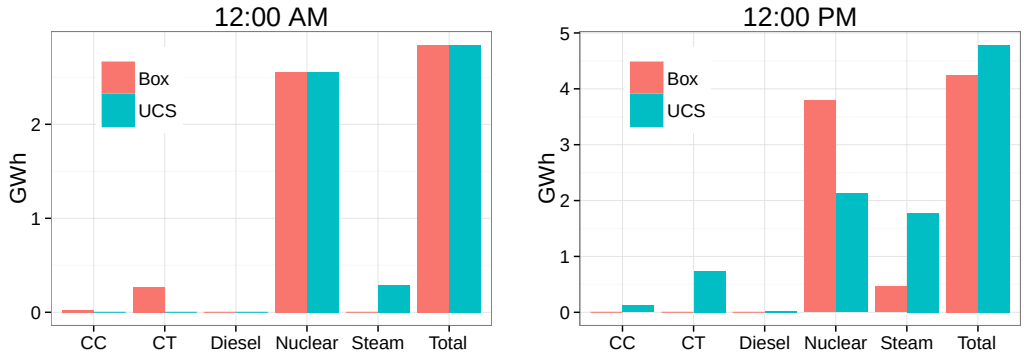


Figure 3-7: Out-of-sample results.



(a) (b)

Figure 3-8: The left panel shows the dispatch for hour ending 1 (12:00 AM) for the projected affine policy over UCS and the robust policy over the box. The right panel is similar but for hour ending 13 (12:00 PM).

accuracy. In this hour, the $\mathcal{U}^{CS}$ and $\mathcal{U}^{M}$ methods dispatch nearly the same amount of total capacity in GWh. By contrast, hour ending 13 (12:00 PM) is a highly volatile hour for which it is much more difficult to predict load. In this hour, $\mathcal{U}^{CS}$ schedules more total capacity. This scheduling of additional capacity is not entirely surprising since $\mathcal{U}^{CS}$ more accurately captures the volatility of the load profiles. What is perhaps more interesting is that, in both hours, $\mathcal{U}^{CS}$ prefers to schedule more expensive and more flexible steam generators over cheaper and less flexible nuclear generators. This choice has two important consequences: First, $\mathcal{U}^{CS}$ in general incurs larger production costs in the second stage because the generators it has available are more expensive. Second, it incurs much smaller load mismatches (in GWh) because it has access to much more flexible units.

Using the conversion $\kappa = \$5000/MWh$, yields the total costs seen in Table 3.2. Figure 3-6a presents an alternative view of these costs, breaking them down between operating costs and mismatch costs. It is easily seen that mismatch costs dominate in all cases. As discussed in Sec. 3.2.4, valuing these mismatch costs is subtle, and the choice of κ is debatable. A different choice of κ would change this tradeoff and, consequently, change the dispatch decisions and total costs of the $\mathcal{U}^{CS}$ method. One might ask what the "break-even" value of κ is, i.e., what is the smallest value of κ such that the $\mathcal{U}^{CS}$ method still outperforms the other methods? Since changing κ changes

the dispatch decisions, computing such a value requires an exhaustive grid search over κ and would be computationally expensive. Without performing an exhaustive search, however, we can still upper bound the break-even value for κ using the current solutions. A rough calculation based on this data then suggests that the $\mathcal{U}^{CS}$ method will still outperform the nominal method as long as κ exceeds \$1,150 / MWh and outperform the $\mathcal{U}^{B}$ method as long as κ exceeds \$861 / MWh.

Regardless of what value of κ is chosen, is is clear that our data-driven approach enjoys a number of important strengths. First, it appears to be more computationally appealing relative to other methods. Second, it is able to identify highly volatile, uncertain hours and appropriately allocate more capacity and more flexible units to those hours to mitigate mismatches. We consider both these features to be noteworthy strengths of the approach.

3.8 Conclusion

In this chapter we presented an application of our data-driven uncertainty sets to a real-world problem using real data. In contrast to existing works that use fully-adaptive optimal with non data-driven, simple uncertainty sets, we show that simpler, suboptimal, affinely adaptive policies in conjunction with more sophisticated, data-driven uncertainty sets can yield high quality solutions. In particular, we discussed how to tailor the our data-driven uncertainty set constructions to non-i.i.d. time-series data, and illustrated numerically that large-scale UC problems can still be solved efficiently for these sets.

Chapter 4

Inverse Variational Inequalities and Modeling Behavior

4.1 Introduction

In this chapter, we consider data-driven models for behavior. As discussed in the introduction, the most common models for behavior in operations research (utility maximization, Nash equilibrium, mean-field equilibrium, etc.) can be recast as variational inequalities (VI). VIs are a natural tool to describe equilibria with applications beyond behavioral modeling in fields including economics, physics, differential equations and optimization. Consequently, although our motivation and examples in what follows will center on behavioral modeling, we present our techniques in the case of a general VI, intending that they may be more generally applicable.

In what follows, we focus on posing and solving the *inverse variational inequality problem*. Namely, given data that we believe are equilibria, i.e., solutions to some VI, estimate the function which describes this VI, i.e. the model primitives. As an example from game theory, we might use a player's past actions in previous games to estimate her utility function assuming that her actions constituted (approximate) Nash equilibria.

Our formulation and analysis is motivated in many ways by the inverse optimization literature. In inverse optimization, one is given a candidate solution to an

optimization problem and seeks to characterize the cost function or other problem data that would make that solution (approximately) optimal. See [72] for a survey of inverse combinatorial optimization problems, [5] for the case of linear optimization and [78] for the case of conic optimization. The critical difference, however, is that we seek a cost function that would make the observed data equilibria, not optimal solutions to an optimization problem.[1] In general, optimization problems can be reformulated as variational inequalities (see Sec. 4.2.1), so that our inverse VI problem *generalizes* inverse optimization, but this generalization allows us to address a variety of new applications.

To the best of our knowledge, we are the first to consider inverse variational inequality problems. Previous work, however, has examined the problem of estimating parameters for systems assumed to be in equilibrium, most notably the structural estimation literature in econometrics and operations management ([6, 8, 92, 102]). Although there are a myriad of techniques collectively referred to as structural estimation, roughly speaking, they entail (1) assuming a parametric model for the system including probabilistic assumptions on random quantities, (2) deducing a set of necessary (structural) equations for unknown parameters, and, finally, (3) solving a constrained optimization problem corresponding to a generalized method of moments (GMM) estimate for the parameters. The constraints of this optimization problem include the structural equations and possibly other application-specific constraints, e.g., orthogonality conditions of instrumental variables. Moreover, this optimization problem is typically difficult to solve numerically, as it is can be non-convex with large flat regions and multiple local optima (see [6] for some discussion).

Our approach differs from structural estimation and other specialized approaches in a number of respects. From a philosophical point of view, the most critical difference is in the objective of the methodology. Specifically, in the structural estimation paradigm, one posits a "ground-truth" model of a system with a known parametric form. The objective of the method is to learn the parameters in order to provide

[1] The related work [29] adopts a hybrid viewpoint wherein the data are derived as a consequence of equilibrium assumptions, but used in an inverse optimization framework.

insight into the system. By contrast, in our paradigm, we make no assumptions (parametric or nonparametric) about the true mechanics of the system; we treat is as a "black-box." Our objective is to fit a model – in fact, a VI – that can be used to predict the behavior of the system. We make no claim that this fitted model accurately reflects "reality," merely that it has good predictive power.

This distinction is subtle, mirroring the distinction between "data-modelling" in classical statistics and "algorithmic modeling" in machine learning. (A famous, albeit partisaned, account of this distinction is [40].) Our approach is kindred to the machine learning point of view. For a more detailed discussion, please see Appendix B.2.

This philosophical difference has a number of *practical* consequences:

1. **Minimal Probabilistic Assumptions:** Our method has provably good performance in a very general setting with minimal assumptions on the underlying mechanism generating the data. (See Theorems 4.6-4.8 for precise statements.) By contrast, other statistical methods, including structural estimation, require a full-specification of the data generating mechanism and can yield spurious results if this specification is inaccurate.

2. **Tractability:** Since our fitted model need not correspond exactly to the underlying system dynamics, we have considerably more flexibility in choosing its functional form. For several interesting choices, including nonparametric specifications (see next point), the resulting inverse VI problem can be reformulated as a conic optimization problem. Conic optimization problems are both theoretically and numerically tractable, even for large scale instances ([38]), in sharp contrast to the non-convex problems that frequently arise in other methods.

3. **Nonparametric Estimation:** Like existing methods in inverse optimization and structural estimation, our approach can be applied in a parametric setting. Unlike these approaches, our approach also extends naturally to a nonparametric description of the function $\mathbf{f}$ defining the VI. To the best of our knowledge, existing methods do not treat this possibility. Partial exceptions are [21] and [68] which use nonparametric estimators for probability densities, but para-

metric descriptions of the mechanism governing the system. The key to our nonparametric approach is to leverage kernel methods from statistical learning to reformulate the infinite dimensional inverse variational inequality problem as a finite dimensional, convex quadratic optimization problem. In applications where we may not know, or be willing to specify a particular form for $\mathbf{f}$ we consider this non-parametric approach particularly attractive.

Although there are other technical differences between these approaches – for example, some structural estimation techniques can handle discrete features while our method applies only to continuous problems – we feel that the most important difference is the aforementioned intended purpose of the methodology. We see our approach as complementary to existing structural estimation techniques and believe in some applications practitioners may prefer it for its computational tractability and relatively fewer modeling assumptions. Of course, in applications where the underlying assumptions of structural estimations or other statistical techniques are valid, those techniques may yield potentially stronger claims about the underlying system.

We summarize our contributions below:

1. We propose the inverse variational inequality problem to model inverse equilibrium. We illustrate the approach by estimating market demand functions under Bertrand-Nash equilibrium and by estimating the congestion function in a traffic equilibrium.

2. We formulate an optimization problem to solve a parametric version of the inverse variational inequality problem. The complexity of this optimization depends on the particular parametric form of the function to be estimated. We show that for several interesting choices of parametric form, the parametric version of the inverse variational inequality problem can be reformulated as a simple conic optimization problem.

3. We formulate and solve a nonparametric version of the inverse variational inequality problem using kernel methods. We show that this problem can be

efficiently solved as a convex quadratic optimization problem whose size scales linearly with the number of observations.

4. Under very mild assumptions on the mechanism generating the data, we show that both our parametric and non-parametric formulations enjoy a strong generalization guarantee similar to the guarantee enjoyed by other methods in machine learning. Namely, if the fitted VI explains the existing data well, it will continue to explain new data well. Moreover, under some additional assumptions on the optimization problem, equilibria from the VI serve as good predictions for new data points.

5. We provide computational evidence in the previous two examples – demand estimation under Nash equilibrium and congestion function estimation under traffic equilibrium – that our proposed approach recovers reasonable functions with good generalization properties and predictive power. We believe these results may merit independent interest in the specialized literature for these two applications.

The remainder of this paper is organized as follows. Section 4.2 reviews background material on equilibrium modeling through VIs. Section 4.3 formally defines the inverse variational inequality problem and solves it in the case that the function to be estimated has a known parametric form. In preparation for the nonparametric case, Section 4.4 reviews some necessary background material on kernels. Section 4.5 formulates and solves the nonparametric inverse variational inequality problem using kernels, and Section 4.6 illustrates how to incorporate priors, semi-parametric modeling and ambiguity sets into this framework. Section 4.7 states our results on the generalization guarantees and predictive power of our approach. Finally, Section 4.8 presents some computational results, and Section 4.9 concludes. In the interest of space, almost all proofs are placed in the Appendix.

4.2 Variational Inequalities

4.2.1 Modelling Behavior

In this section, we briefly review some results on variational inequalities that we use in the remainder of the chapter. For a more complete survey, see [70]. Most importantly, we demonstrate how popular paradigms for modeling behavior can be recast as variational inequalities.

Given a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and a non-empty set $\mathcal{F} \subseteq \mathbb{R}^n$ the variational inequality problem, denoted VI$(\mathbf{f}, \mathcal{F})$, is to find an $\mathbf{x}^* \in \mathcal{F}$ such that

$$\mathbf{f}(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{4.1}$$

A solution $\mathbf{x}^*$ to VI$(\mathbf{f}, \mathcal{F})$ need not exist, and when it exists, it need not be unique. We can guarantee the existence and uniqueness of the solution by making appropriate assumptions on $\mathbf{f}(\cdot)$ and $\mathcal{F}$, e.g., $\mathbf{f}$ continuous and $\mathcal{F}$ convex and compact. See [70] for other less stringent conditions.

There are at least three classical applications of VI modeling that we will refer to throughout the paper: utility maximization (constrained optimization), Nash equilibrium, and Wardrop (or traffic) equilibrium.

Utility Maximization. The simplest example of a VI is in fact not an equilibrium, per se, but rather constrained optimization. Nonetheless, the specific example is very useful in building intuition about VIs. Using this formalism, one can derive many of the existing results in the inverse optimization literature as a special case of our results for inverse VIs in Section 4.3.2. Most importantly, as discussed in the introduction, utility maximization is the most popular paradigm for modeling single agent behavior. This example illustrates how utility maximization can be viewed in the VI framework.

Consider the problem

$$\min_{\mathbf{x} \in \mathcal{F}} F(\mathbf{x}). \tag{4.2}$$

The first order necessary conditions for an optimal solution of this problem are (see,

e.g., [22])

$$\nabla F(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{4.3}$$

These conditions are sufficient in the case that F is a convex function and $\mathcal{F}$ is a convex set. Observe, then, that solving (4.2) is equivalent to finding a point which satisfies Eq. (4.3), which is equivalent to solving VI($\nabla F, \mathcal{F}$).

Note that, in general, a VI with a function $\mathbf{f}$ whose Jacobian is symmetric models an optimization problem (see [70]).

Nash Equilibrium. Our second example of a VI is non-cooperative Nash equilibrium. This equilibrium concept is an incredibly popular paradigm for modeling behavior among a small number of agents.

Consider a game with p players. Each player i chooses an action from a set of feasible actions, $\mathbf{a}_i \in \mathcal{A}_i \subseteq \mathbb{R}^{m_i}$, and receives a utility $U_i(\mathbf{a}_1, \ldots, \mathbf{a}_p)$. Notice in particular, that player i's payoff may depend upon the actions of other players. We will assume that U_i is concave and differentiable in $\mathbf{a}_i$ and that $\mathcal{A}_i$ is convex for all i.

A profile of actions for the players $(\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_p^*)$ is said to be a Nash Equilibrium if no single player can unilaterally change her action and increase her utility. See [60] for a more complete treatment. In other words, player i plays her best response given the actions of the other players. More formally,

$$\mathbf{a}_i^* \in \arg\max_{\mathbf{a} \in \mathcal{A}_i} U_i(\mathbf{a}_1^*, \ldots, \mathbf{a}_{i-1}^*, \mathbf{a}, \mathbf{a}_{i+1}^*, \ldots, \mathbf{a}_p^*), \quad i = 1, \ldots, p. \tag{4.4}$$

This condition can be expressed as a VI. Specifically, a profile $\mathbf{a}^* = (\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_p^*)$ is a Nash Equilibrium, if and only if it solves VI($\mathbf{f}, \mathcal{F}$) where $\mathcal{F} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_p$,

$$\mathbf{f}(\mathbf{a}) = \begin{pmatrix} -\nabla_1 U_1(\mathbf{a}) \\ \vdots \\ -\nabla_p U_p(\mathbf{a}) \end{pmatrix} \tag{4.5}$$

and ∇_i denotes the gradient with respect to the variables $\mathbf{a}_i$ (see [70] for a proof.)

It is worth pointing out that many authors use Eq. (4.4) to conclude

$$\nabla_i U_i(\mathbf{a}_1^*, \ldots, \mathbf{a}_p^*) = \mathbf{0}, \quad i = 1, \ldots, p, \tag{4.6}$$

where ∇_i refers to a gradient with respect to the coordinates of $\mathbf{a}_i$. This characterization assumes that each player's best response lies on the strict interior of her strategy set $\mathcal{A}_i$. The assumption is often valid, usually because the strategy sets are unconstrained. Indeed, this condition can be derived as a special case of (4.5) in the case $\mathcal{A}_i = \mathbb{R}^{m_i}$. In some games, however, it is not clear that an equilibrium must occur in the interior, and we must use (4.5) instead. We will see an example in Sec. 4.3.3.

Wardrop Equilibrium. Our final example of a VI is Wardrop or user-equilibrium from transportation science. Wardrop equilibrium is extremely close in spirit to the Walrasian (market) equilibrium model in economics – see [50], [115] – and our comments below naturally extend to the Walrasian case. Wardrop equilibrium is an example of using a VI to model behavior when the number of agents is very large.

Specifically, we are given a directed network of nodes and arcs $(\mathcal{V}, \mathcal{A})$, representing the road network of some city. Let $\mathbf{N} \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{A}|}$ be the node-arc incidence matrix of this network. For certain pairs of nodes $\mathbf{w} = (w_s, w_t) \in \mathcal{W}$, we are also given an amount of flow $d^{\mathbf{w}}$ that must flow from w_s to w_t. The pair $\mathbf{w}$ is referred to as an origin-destination pair. Let $\mathbf{d}^{\mathbf{w}} \in \mathbb{R}^{|V|}$ be the vector which is all zeros, except for a $(-d^{\mathbf{w}})$ in the coordinate corresponding to node w_s and a $(d^{\mathbf{w}})$ in the coordinate corresponding to node w_t.

We will say that a vector of flows $\mathbf{x} \in \mathbb{R}_+^{|\mathcal{A}|}$ is feasible if $\mathbf{x} \in \mathcal{F}$ where

$$\mathcal{F} = \left\{ \mathbf{x} : \exists \mathbf{x}^{\mathbf{w}} \in \mathbb{R}_+^{|\mathcal{A}|} \text{ s.t.} \quad \mathbf{x} = \sum_{\mathbf{w} \in \mathcal{W}} \mathbf{x}^{\mathbf{w}}, \quad \mathbf{N}\mathbf{x}^{\mathbf{w}} = \mathbf{d}^{\mathbf{w}} \quad \forall \mathbf{w} \in \mathcal{W} \right\}.$$

Let $c_a : \mathbb{R}_+^{|\mathcal{A}|} \to \mathbb{R}_+$ be the "cost" function for arc $a \in \mathcal{A}$. The interpretation of cost, here, is deliberately vague. The cost function might represent the actual time it takes to travel an arc, tolls users incur along that arc, disutility from environmental

factors along that arc, or some combination of the above. Note that because of interdependencies in the network, the cost of traveling arc a may depend not only on $\mathbf{x}_a$, but on the flows on other arcs as well. Denote by $\mathbf{c}(\cdot)$ the vector-valued function whose a-th component is $c_a(\cdot)$.

A feasible flow $\mathbf{x}^*$ is a Wardrop equilibrium if for every origin-destination pair $\mathbf{w} \in W$, and any path connecting (w_s, w_t) with positive flow in $\mathbf{x}^*$, the cost of traveling along that path is less than or equal to the cost of traveling along any other path that connects (w_s, w_t). Here, the cost of traveling along a path is the sum of the costs of each of its constituent arcs. Intuitively, a Wardrop equilibrium captures the idea that if there exists a less congested route connecting w_s and w_t, users would find and use it instead of their current route.

It is well-known that a Wardrop equilibrium is a solution to VI($\mathbf{c}, \mathcal{F}$).

4.2.2 Approximate Equilibria

Let $\epsilon > 0$. We will say that $\hat{\mathbf{x}} \in \mathcal{F}$ is an ϵ-approximate solution to VI($\mathbf{f}, \mathcal{F}$) if

$$\mathbf{f}(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}) \geq -\epsilon, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{4.7}$$

This notion of an approximate solution is not new to the VI literature– it corresponds exactly to the condition that the primal gap function of the VI is bounded above by ϵ and is frequently used in the analysis of numerical procedures for solving the VI. We point out that ϵ-approximate solutions also frequently have a modeling interpretation. For example, consider the case of constrained convex optimization (cf. Eq. (4.2)). Let $\mathbf{x}^*$ be an optimal solution. Since F is convex, we have $F(\hat{\mathbf{x}}) - F(\mathbf{x}^*) \leq -\nabla F(\hat{\mathbf{x}})^T(\mathbf{x}^* - \hat{\mathbf{x}}) \leq \epsilon$. In other words, ϵ-approximate solutions to VIs generalize the idea of ϵ-optimal solutions to convex optimization problems. Similarly, in a Nash equilibrium, an ϵ-approximate solution to the VI (4.5) describes the situation where each player i does not necessarily play her best response given what the other players are doing, but plays a strategy which is no worse than ϵ from her best response.

The idea of ϵ-approximate solutions is not the only notion of an approximate

equilibrium. An alternative notion of approximation is that $\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq \delta$ where $\mathbf{x}^*$ is a solution to the VI$(\mathbf{f}, \mathcal{F})$. We say such a $\hat{\mathbf{x}} \in \mathcal{F}$ is δ-near a solution to the VI$(\mathbf{f}, \mathcal{F})$. As shown in Theorem 4.1, these two ideas are closely related. The theorem was proven in [94] to provide stopping criteria for certain types of iterative algorithms for solving VIs. We reinterpret it here in the context of approximate equilibria.

Before stating the theorem, we define strong monotonicity. We will say that $\mathbf{f}(\cdot)$ is *strongly monotone* if $\exists \gamma > 0$ such that

$$(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \gamma \|\mathbf{x} - \mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}.$$

When the VI corresponds to constrained optimization (cf. Eqs. (4.2), (4.3)), strong monotonicity of f corresponds to strong convexity of F. Intuitively, strong monotonicity ensures that f does not have large, flat regions.

Theorem 4.1 ([94])**.** *Suppose* $\mathbf{f}$ *is strongly monotone with parameter* γ*. Then an* ϵ*-approximate solution to VI*$(\mathbf{f}, \mathcal{F})$ *is* $\sqrt{\frac{\epsilon}{\gamma}}$*-near an exact solution.*

We require Theorem 4.1 in Section 4.7 to prove some of our generalization results.

4.2.3 Characterizing Approximate Solutions to VIs

In this section we provide an alternative characterization of an ϵ-approximate solution (cf. Eq. (4.7)) in the case when $\mathcal{F}$ is represented by the intersection of conic inequalities. Specifically, for the remainder of the paper, we will assume:

Assumption 1. $\mathcal{F}$ *can be represented as the intersection of a small number of conic inequalities in standard form,* $\mathcal{F} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in C\}$.

Assumption 2. $\mathcal{F}$ *satisfies a Slater-condition*

Moreover, for any proper cone C, i.e. C is pointed, closed, convex and has a strict interior, we will say $\mathbf{x} \leq_C \mathbf{y}$ whenever $\mathbf{y} - \mathbf{x} \in C$.

The assumption that $\mathcal{F}$ is given in standard form is not crucial. All of our results extend to the case that $\mathcal{F}$ is not given in standard form at the expense of some

notation. It is, however, crucial, that $\mathcal{F}$ is conic representable. Observe that when C is the nonnegative orthant, we recover the special case where $\mathcal{F}$ is a polyhedron. To stress the dependence on $\mathbf{A}, \mathbf{b}, C$, we will write VI($\mathbf{f}, \mathbf{A}, \mathbf{b}, C$).

The following result was proven in [4] to reformulate VI($\mathbf{f}, \mathbf{A}, \mathbf{b}, C$) as a single-level optimization problem. We reinterpret here as a characterization of approximate equilibria and sketch a short proof for completeness.

Theorem 4.2 ([4])**.** *Under assumptions **A1**, **A2**, the solution $\hat{\mathbf{x}}$ is an ϵ-approximate equilibrium to VI($\mathbf{f}, \mathbf{A}, \mathbf{b}, C$) if and only if $\exists \mathbf{y}$ s.t.*

$$\mathbf{A}^T\mathbf{y} \leq_C \mathbf{f}(\hat{\mathbf{x}}), \tag{4.8}$$

$$\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \mathbf{b}^T\mathbf{y} \leq \epsilon. \tag{4.9}$$

Proof. First suppose that $\hat{\mathbf{x}}$ is an ϵ-approximate equilibrium. Then, from Eq. (4.7), $\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \epsilon \leq \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x},\ \forall \mathbf{x} \in \mathcal{F}$, which is equivalent to $\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \epsilon \leq \min_{\mathbf{x}\in\mathcal{F}} \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x}$. The right hand side is a conic optimization problem in $\mathbf{x}$, and the above shows it is bounded below. Since $\mathcal{F}$ has non-empty interior, strong duality holds (see [38]), which implies that there exists a dual solution $\mathbf{y}$ that attains the optimum. In other words,

$$\min_{\mathbf{x}\in\mathcal{F}} \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x} = \max_{\mathbf{y}:\mathbf{A}^T\mathbf{y}\leq_C \mathbf{f}(\hat{\mathbf{x}})} \mathbf{b}^T\mathbf{y}.$$

Substituting this dual solution into the above inequality and rearranging terms yields the result. The reverse direction is proven analogously using weak conic duality. □

The above proof leverages the fact that the duality gap between an optimal primal and dual solution pair is zero. We can instead formulate a slightly different characterization by leveraging complementary slackness. In this case, Eq. (4.9) is replaced by the additional constraints

$$\sum_{i=1}^{n} x_i(f_i(\hat{\mathbf{x}}) - \mathbf{y}^T\mathbf{A}\mathbf{e}_i) \leq \epsilon. \tag{4.10}$$

Depending on the application, either the strong duality representation (cf. Eqs. (4.8),

(4.9)) or the complementary slackness representation (cf. Eqs. (4.8), (4.10) may be more natural. We will use the strong duality formulation in Section 4.8.3 and the the complementary slackness formulation in Section 4.8.1.

4.3 Inverse Variational Inequalities

4.3.1 Description

We are now in a position to pose the inverse variational inequality problem. We are given observations $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ for $j = 1, \ldots, N$. In this context, we modify Assumption **A2** to read

Assumption. *The set* $\mathcal{F}_j = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}_j\mathbf{x} = \mathbf{b}_j, \mathbf{x} \in C_j\}$ *is non-empty and satisfies a Slater condition for each* j*.*

We seek a function $\mathbf{f}$ such that $\mathbf{x}_j$ is an approximate solution to $\text{VI}(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j, C_j)$ for each j. Note, the function $\mathbf{f}$ is common to all observations. Specifically, we would like to solve:

$$\begin{aligned} \min_{\mathbf{f},\boldsymbol{\epsilon}} \quad & \|\boldsymbol{\epsilon}\| \\ \text{s.t.} \quad & \mathbf{x}_j \text{ is an } \epsilon_j\text{-approximate solution to } \text{VI}(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j, C_j), \quad j = 1, \ldots, N, \\ & \mathbf{f} \in \mathcal{S}. \end{aligned} \tag{4.11}$$

where $\|\cdot\|$ represents some choice of norm, and $\mathcal{S}$ represents a set of admissible functions. In the parametric case, treated in the following section, we will assume that $\mathcal{S}$ is indexed by a vector of parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^M$. In the nonparametric case, $\mathcal{S}$ will be a general set of functions that satisfy certain smoothness properties. We defer this extension until Section 4.5.

4.3.2 Parametric Formulation

In this section we assume that the function $\mathbf{f}$ is known to belong to a parametric family indexed by a vector $\boldsymbol{\theta} \in \Theta$. We write $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ to denote this dependence. We

will assume throughout that Θ is compact and $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. A direct application of Theorem 4.2 yields the following reformulation:

Theorem 4.3. *Under assumptions 1, 2 and the additional constraint that* $\mathbf{f} = \mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ *for some* $\boldsymbol{\theta} \in \Theta$*, problem Eq.* (4.11) *can be reformulated as*

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\Theta,\mathbf{y},\boldsymbol{\epsilon}} \quad & \|\boldsymbol{\epsilon}\| \\ s.t. \quad & \mathbf{A}_j^T\mathbf{y}_j \leq_C \mathbf{f}(\mathbf{x}_j;\boldsymbol{\theta}), \quad j = 1,\dots N, \\ & \mathbf{f}(\mathbf{x}_j;\boldsymbol{\theta})^T\mathbf{x}_j - \mathbf{b}_j^T\mathbf{y}_j \leq \epsilon_j, \quad j = 1,\dots,N, \end{aligned} \tag{4.12}$$

where $\mathbf{y} = (\mathbf{y}_1,\dots,\mathbf{y}_N)$.

Remark 4.1 (Multiple equilibria). We stress that since Theorem 4.2 is true for any ϵ-approximate solution to the VI, Theorem 4.3 is valid even when the function $\mathbf{f}$ might give rise to multiple distinct equilibria. This robustness to multiple equilibria is an important strength of our approach that distinguishes it from other specialized approaches that require uniqueness of the equilibrium.

Remark 4.2 (Equilibria on the boundary). In Theorem 4.2, we did not need to assume that the $\mathbf{x}_j$ or the solutions to $VI(\mathbf{f},\mathbf{A}_j,\mathbf{b}_j)$ belonged to the interior of $\mathcal{F}_j$. Consequently, Theorem 4.3 is valid even if the observations $\mathbf{x}_j$ or induced solutions to $VI(\mathbf{f},\mathbf{A}_j,\mathbf{b}_j)$ occur on the boundary. This is in contrast to many other techniques which require that the solutions occur on the relative interior of the feasible set.

Remark 4.3 (Computational complexity). Observe that $\mathbf{x}_j$ are data in Problem (4.12), not decision variables. Consequently, the complexity of this optimization depends on the cone C and the dependence of $\mathbf{f}$ on $\boldsymbol{\theta}$, but *not* on the dependence of $\mathbf{f}$ on $\mathbf{x}$. For a number of interesting parametric forms, we can show that Problem (4.12) is in fact tractable.

As an example, suppose we specify $\mathbf{f}(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{M}\theta_i\boldsymbol{\phi}_i(\mathbf{x})$ where $\boldsymbol{\phi}_1(\mathbf{x}),\dots,\boldsymbol{\phi}_M(\mathbf{x})$ is a set of (nonlinear) basis functions. Since $\mathbf{f}$ depends linearly on $\boldsymbol{\theta}$, Problem (4.12) is a conic optimization problem, even though the basis functions $\phi_i(\mathbf{x})$ may be arbitrary nonlinear functions. Indeed, if C is the nonnegative orthant, Problem (4.12) is a

linear optimization problem. Similarly, if C is the second-order cone, Problem (4.12) is a second-order cone problem.

Recall from the introduction that structural estimation is an alternative parametric technique for estimation in equilibrium from econometrics. Although structural estimation is not the focus of this chapter, in Appendix B.2 we briefly illustrate how to use Theorem 4.2 to formulate an alternate optimization problem that is similar to, but different from, Problem (4.12) and closer in spirit to structural estimation techniques. Moreover, we show that this formulation is equivalent to certain structural estimation techniques in the sense that they produce the same estimators. This section may prove useful to readers interested in comparing these methodologies.

4.3.3 Application: Demand Estimation in Bertrand-Nash Equilibrium

In this section we use Theorem 4.3 to estimate an unknown demand function for a product so that observed prices are approximately in Bertrand-Nash equilibrium. This is a somewhat stylized example inspired by various influential works in the econometrics literature, such as [19] and [20]. We include this styled example for two reasons: 1) To illustrate a simple problem where equilibria may occur on the boundary of the feasible region. 2) To further clarify how the choice of parameterization of $\mathbf{f}(\cdot;\boldsymbol{\theta})$ affects the computational complexity of the estimation problem.

For simplicity, consider two firms competing by setting prices p_1, p_2, respectively. Demand for firm i's product, denoted $D_i(p_1, p_2, \xi)$, is a function of both prices, and other economic indicators, such as GDP, denoted by ξ. Each firm sets prices to maximize its own revenues $U_i(p_1, p_2, \xi) = p_i D_i(p_1, p_2, \xi)$ subject to the constraint $0 \leq p_i \leq \overline{p}$. The upper bound $\overline{p}$ might be interpreted as a government regulation as is frequent in some markets for public goods, like electricity. We assume a priori that each demand function belongs to some given parametric family indexed by $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta$: $D_1(p_1, p_2, \xi; \boldsymbol{\theta}_1), D_2(p_1, p_2, \xi; \boldsymbol{\theta}_2)$. We seek to estimate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ so that the data (p_1^j, p_2^j, ξ) for $j = 1, \ldots, N$ correspond approximately to Nash equilibria.

Both [19] and [20] assume that equilibrium prices do not occur on the boundary, i.e., that $p_i < \overline{p}$ since they leverage Eq. (4.6) in their analysis. These methods are, thus, not directly applicable.

By contrast, Theorem 4.3 directly applies, yielding

$$\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\epsilon},(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta} \quad & \|\boldsymbol{\epsilon}\| \\
\text{s.t.} \quad & \mathbf{y}^j \geq \mathbf{0}, \quad j = 1,\ldots,N, \\
& y_i^j \geq p_i^j \frac{\partial}{\partial p_i} D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) + D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i), \quad i = 1,2,\ j = 1,\ldots,N, \\
& \sum_{i=1}^{2} \overline{p}^j y_i^j - (p_i^j)^2 \frac{\partial}{\partial p_i} D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) - p_i^j D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) \leq \epsilon_j, \quad j = 1,\ldots,N.
\end{aligned} \tag{4.13}$$

We stress that potentially more complex constraints on the feasible region can be incorporated just as easily.

Next, recall that the complexity of the optimization problem (4.13) depends on the parameterization of $D_i(p_1, p_2, \xi, \boldsymbol{\theta}_i)$. For example, when demand is linear,

$$D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = \theta_{i0} + \theta_{i1} p_1 + \theta_{i2} p_2 + \theta_{i3} \xi \tag{4.14}$$

problem (4.13) reduces to the linear optimization problem:

$$\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\epsilon},(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta,\mathbf{d}} \quad & \|\boldsymbol{\epsilon}\| \\
\text{s.t.} \quad & \mathbf{y}^j \geq \mathbf{0}, \quad j = 1,\ldots,N, \\
& y_i^j \geq d_i^j + \theta_{ii} p_i^j, \qquad i = 1,2, \quad j = 1,\ldots,N, \\
& \overline{p} \sum_{i=1}^{2} y_i^j - p_i^j d_i^j - (p_i^j)^2 \theta_{ii} \leq \epsilon_j, \quad j = 1,\ldots,N, \\
& d_i^j = \theta_{i0} + \theta_{i1} p_1^j + \theta_{i2} p_2^j + \theta_{i3} \xi^j, \qquad i = 1,2, \quad j = 1,\ldots,N.
\end{aligned} \tag{4.15}$$

Alternatively, if we assume demand is given by the multinomial logit model [61],

$D_i(p_1, p_2, \xi; \boldsymbol{\theta}) = \frac{e^{\theta_{i0}+\theta_{i1}p_i+\theta_{i3}\xi}}{e^{\theta_{10}+\theta_{11}p_1+\theta_{13}\xi}+e^{\theta_{20}+\theta_{21}p_2+\theta_{23}\xi}+\theta_{00}}$, the problem (4.13) becomes

$$\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\theta_1},\boldsymbol{\theta_2},\mathbf{d_1},\mathbf{d_2}} \quad & \|\boldsymbol{\epsilon}\| \\
\text{s.t.} \quad & \mathbf{y}^j \geq \mathbf{0}, \quad j = 1, \ldots, N, \\
& y_i^j \geq p_i^j \theta_{i1} d_1^j d_2^j + d_i^j, \quad i = 1, 2, \\
& \sum_{i=1}^{2} \overline{p}^j y_i^j + p_i^j d_i^j - (p_i^j)^2 \theta_{1i} d_i^j (1 - d_i^j) \leq \epsilon_j \\
& d_i^j = \frac{e^{\theta_{0i}+\theta_{i1}p_i^j+\theta_{i3}\xi^j}}{e^{\theta_{10}+\theta_{11}p_1^j+\theta_{13}\xi^j} + e^{\theta_{20}+\theta_{21}p_2^j+\theta_{23}\xi} + \theta_{00}}, \quad i = 1, 2, j = 1, \ldots N,
\end{aligned}$$

which is non-convex. Non-convex optimization problems can be challenging numerically and may scale poorly.

Finally, we point out that although it more common in the econometrics literature to specify the demand functions D_i directly as we have above, one could equivalently specify the marginal revenue functions $M_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = p_i \partial_i D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) + D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i)$ and then impute the demand function as necessary. We adopt this equivalent approach later in Section 4.8.1.

4.4 Kernel Methods: Background

Intuitively, our nonparametric approach in the next section seeks the "smoothest" function $\mathbf{f}$ which make the observed data approximate equilibria, where the precise notion of smoothness is determined by the choice of kernel. Kernel methods have been used extensively in machine learning, most recently for feature extraction in context of support-vector machines or principal component analysis. Our use of kernels, however, more closely resembles their application in spline interpolation and regularization networks ([62, 112]).

Our goal in this section is to develop a sufficiently rich set of scalar valued functions over which we can tractably optimize using kernel methods. Consequently, we first develop some background. Our review is not comprehensive. A more thorough treatment of kernel methods can be found in [58, 105, 110].

Let $k : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a symmetric function. We will say that k is a kernel if k is positive semidefinite over $\mathcal{F}$, i.e., if

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \text{ for any choice of } N \in \mathbb{N}, \quad \mathbf{c} \in \mathbb{R}^N, \quad \mathbf{x}_i \in \mathcal{F}.$$

Examples of kernels over $\mathbb{R}^n$ include:

Linear: $k(\mathbf{x}, \mathbf{y}) \equiv \mathbf{x}^T\mathbf{y}$,

Polynomial: $k(\mathbf{x}, \mathbf{y}) \equiv (c + \mathbf{x}^T\mathbf{y})^d$ for some choice of $c \geq 0$ and $d \in \mathbb{N}$,

Gaussian: $k(\mathbf{x}, \mathbf{y}) \equiv \exp(-c\|\mathbf{x} - \mathbf{y}\|^2)$ for some choice of $c > 0$.

Let $k_{\mathbf{x}}(\cdot) \equiv k(\mathbf{x}, \cdot)$ denote the function of one variable obtained by fixing the first argument of k to $\mathbf{x}$ for any $\mathbf{x} \in \mathcal{F}$. Define $\mathcal{H}_0$ to be the vector space of scalar valued functions which are representable as finite linear combinations of elements $k_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{F}$, i.e.,

$$\mathcal{H}_0 = \left\{ \sum_{j=1}^{N} \alpha_j k_{\mathbf{x}_j} : \mathbf{x}_j \in \mathcal{F},\ N \in \mathbb{N},\ \alpha_j \in \mathbb{R},\ j = 1, \ldots, N, N \in \mathbb{N} \right\}. \tag{4.16}$$

Observe that $k_{\mathbf{x}} \in \mathcal{H}_0$ for all $\mathbf{x} \in \mathcal{F}$, so that in a sense these elements form a basis of the space $\mathcal{H}_0$. On the other hand, for a given $f \in \mathcal{H}_0$, its representation in terms of these elements $k_{\mathbf{x}_j}$ for $\mathbf{x}_j \in \mathcal{F}$ need not be unique. In this sense, the elements $k_{\mathbf{x}}$ are not like a basis.

For any $f, g \in \mathcal{H}_0$ such that

$$f = \sum_{j=1}^{N} \alpha_j k_{\mathbf{x}_j}, \quad g = \sum_{i=1}^{N} \beta_i k_{\mathbf{x}_i}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^N \tag{4.17}$$

we define a scalar product

$$\langle f, g \rangle_{\mathcal{H}_0} \ = \ \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \beta_j \langle k_{\mathbf{x}_i}, k_{\mathbf{x}_j} \rangle_{\mathcal{H}_0} \ \equiv \ \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{4.18}$$

Since the representation in (4.17) is not unique, for this to be a valid definition one must prove that the right-hand side of the last equality is independent of the choice of representation. It is possible to do so. See [105] for the details. Finally, given this scalar-product, we define the norm $\|f\|_{\mathcal{H}_0} \equiv \sqrt{\langle f, f \rangle}_{\mathcal{H}_0}$.

In what follows, we will actually be interested in the closure of $\mathcal{H}_0$, i.e.,

$$\mathcal{H} = \overline{\mathcal{H}_0}. \tag{4.19}$$

We extend the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ and norm $\| \cdot \|_{\mathcal{H}_0}$ to $\mathcal{H}$ by continuity. (Again, see [105] for the details). Working with $\mathcal{H}$ instead of $\mathcal{H}_0$ simplifies many results.[2]

As an example, in the case of the linear and polynomial kernels, the space $\mathcal{H}$ is finite dimensional and corresponds to the space of linear functions and the space of polynomials of degree at most d, respectively. In the case of the Gaussian kernel, the space $\mathcal{H}$ is infinite dimensional and is a subspace of all continuous functions.

If $f \in \mathcal{H}_0$ admits a finite representation as in Eq.(4.17), note that from Eq. (4.18) we have for all $\mathbf{x} \in \mathcal{F}$

$$\langle k_{\mathbf{x}}, f \rangle_{\mathcal{H}} \;=\; \sum_{j=1}^{N} \alpha_j k(\mathbf{x}, \mathbf{x}_j) \;=\; f(\mathbf{x}). \tag{4.20}$$

In fact, it can be shown that this property applies to all $f \in \mathcal{H}$ ([87]). This is the most fundamental property of $\mathcal{H}$ as it allows us to relate the scalar product of the space to function evaluation. Eq. (4.20) is termed the *reproducing property* and as a consequence, $\mathcal{H}$ is called a *Reproducing Kernel Hilbert Space (RKHS).*

At this point, it may appear that RKHS are very restrictive spaces of functions. In fact, it can be shown that any Hilbert space of scalar-valued functions for which there exists a $c \in \mathbb{R}$ such that for each $f \in \mathcal{H}$, $|f(\mathbf{x})| \leq c\|f\|_{\mathcal{H}}$ for all $\mathbf{x} \in \mathcal{F}$ is an RKHS ([87]). Thus, RKHS are fairly general. Practically speaking, though, our three previous examples of kernels –linear, polynomial, and Gaussian –are by far the most common in the literature.

[2]For the avoidance of doubt, the closure in (4.19) is with respect to the norm $\| \cdot \|_{\mathcal{H}_0}$.

We conclude this section with a discussion about the norm $\|f\|_{\mathcal{H}}$. We claim that in each of our previous examples, the norm $\|f\|_{\mathcal{H}}$ makes precise a different notion of "smoothness" of the function f. For example, it is not hard to see that if $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, then under the linear kernel $\|f\|_{\mathcal{H}} = \|\mathbf{w}\|$. Thus, functions with small norm have small gradients and are "smooth" in the sense that they do not change value rapidly in a small neighborhood.

Similarly, it can be shown (see [62]) that under the Gaussian kernel,

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^n} \int |\tilde{f}(\omega)|^2 e^{\frac{\|\omega\|^2}{2c}} d\omega, \tag{4.21}$$

where $\tilde{f}$ is the Fourier transformation of f. Thus, functions with small norms do not have many high-frequency Fourier coefficients and are "smooth" in the sense that they do not oscillate very quickly.

The case of the polynomial kernel is somewhat more involved as there does not exist a *simple* explicit expression for the norm (see [62]). However, it is easily confirmed numerically using Eq. (4.18) that functions with small norms do not have large coefficients and do not have have high degree. Consequently, they are "smooth" in the sense that their derivatives do not change value rapidly in a small neighborhood.

Although the above reasoning is somewhat heuristic, it is possible to make the intuition that the norm on an RKHS describes a notion of smoothness completely formal. The theoretical details go beyond the scope of this paper (see [62]).

4.5 Nonparametric Formulation

4.5.1 Kernel Expansions

In this section we develop a nonparametric approach to the inverse variational inequality problem. The principal difficulty in formulating a nonparametric equivalent to (4.11) is that the problem is ill-posed. Specifically, if the set S is sufficiently rich, we expect there to be many, potentially infinitely many, different functions $\mathbf{f}$ which all reconcile the data, and make each observation an exact equilibrium. Intuitively, this

multiplicity of solutions is similar to the case of interpolation where, given a small set of points, many different functions will interpolate between them exactly. Which function, then, is the "right" one?

We propose to select the function $\mathbf{f}$ of minimal $\mathcal{H}$-norm among those that approximately reconcile the data. This choice has several advantages. First, as mentioned earlier, functions with small norm are "smooth", where the precise definition of smoothness will be determined by the choice of kernel. We feel that in many applications, assuming that the function defining a VI is smooth is very natural. Second, as we shall prove, identifying the function $\mathbf{f}$ with minimal norm is computationally tractable, even when the RKHS $\mathcal{H}$ is infinite dimensional. Finally, as we will show in Section 4.7, functions with bounded $\mathcal{H}$-norm will have good generalization properties.

Using Theorem 4.2, we reformulate Problem (4.11) as

$$
\begin{aligned}
\min_{\mathbf{f},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \|f_i\|_{\mathcal{H}}^2 \\
\text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq \mathbf{f}(\mathbf{x}_j), \qquad j = 1, \ldots, N, && \text{(4.22a)} \\
& \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j, \quad j = 1, \ldots, N, && \text{(4.22b)} \\
& \|\boldsymbol{\epsilon}\| \leq \kappa, \quad \boldsymbol{\epsilon} \geq \mathbf{0}, \quad f_i \in \mathcal{H}, \quad i = 1, \ldots, n, \\
& \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) = 1. && \text{(4.22c)}
\end{aligned}
$$

Here f_i is the i-th component of the vector function $\mathbf{f}$ and $\mathcal{H}$ is an RKHS. Since we may always scale the function $\mathbf{f}$ in $\text{VI}(\mathbf{f}, \mathcal{F})$ by a positive constant without affecting the solution, we require the last constraint as a normalization condition. Finally, the exogenous parameter κ allows us to balance the norm of $\mathbf{f}$ against how closely $\mathbf{f}$ reconciles the data; decreasing κ will make the observed data closer to equilibria at the price of $\mathbf{f}$ having greater norm.

Problem (4.22) is an optimization over functions, and it is not obvious how to solve it. We show in the next theorem, however, that this can be done in a tractable way. This theorem is an extension of a representation theorem from the kernel literature

(see [112]) to the constrained multivariate case. See the appendix for a proof.

Theorem 4.4. *Suppose Problem* (4.22) *is feasible. Then, there exists an optimal solution* $\mathbf{f}^* = (f_1^*, \ldots, f_n^*)$ *with the following form:*

$$f_i^* = \sum_{j=1}^{N} \alpha_{i,j} k_{\mathbf{x}_j}, \tag{4.23}$$

for some $\alpha_{i,j} \in \mathbb{R}$*, where* k *denotes the kernel of* $\mathcal{H}$*.*

By definition of $\mathcal{H}$, when Problem (4.22) is feasible, its solution is a potentially infinite expansion in terms of the kernel function evaluated at various points of $\mathcal{F}$. The importance of Theorem 4.4 is that it allows us to conclude, first, that this expansion is in fact finite, and second, that the relevant points of evaluation are exactly the data points $\mathbf{x}_j$. This fact further allows us to replace the optimization problem (4.22), which is over an infinite dimensional space, with an optimization problem over a finite dimensional space.

Theorem 4.5. *Problem* (4.22) *is feasible if and only if the following optimization problem is feasible:*

$$\begin{aligned}
\min_{\boldsymbol{\alpha},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha}^T \mathbf{e}_i \\
s.t. \quad & \mathbf{A}_j \mathbf{y}_j \leq \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j \quad j = 1, \ldots, N, \\
& \mathbf{x}_j^T \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j \quad j = 1, \ldots, N, \\
& \|\boldsymbol{\epsilon}\| \leq \kappa, \quad \boldsymbol{\epsilon} \geq \mathbf{0}, \\
& \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^T \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j = 1.
\end{aligned} \tag{4.24}$$

Here $\boldsymbol{\alpha} = (\alpha_{ij})_{i=1,j=1}^{i=n,j=N} \in \mathbb{R}^{n \times N}$*, and* $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{i,j=N}$*. Moreover, given an optimal solution* $\boldsymbol{\alpha}$ *to the above optimization problem, an optimal solution to Problem* (4.22) *is given by Eq.* (4.23)*.*

See the appendix for a proof. Given the optimal parameters $\boldsymbol{\alpha}, \mathbf{f}$ can be evaluated at new points $\mathbf{t}$ using (4.23). Note that $\mathbf{K}$ is positive semidefinite (as a matrix) since

k is positive definite (as a function). Thus, (4.24) is a convex, quadratic optimization problem. Such optimization problems are very tractable numerically and theoretically, even for large-scale instances. (See [38]). Moreover, this quadratic optimization problem exhibits block structure – only the variables $\boldsymbol{\alpha}_j$ couple the subproblems defined by the $\mathbf{y}_j$ — which can be further exploited in large-scale instances. Finally, the size of this optimization scales with N, the number of observations, not with the dimension of the original space $\mathcal{H}$, which may be infinite.

Observe that Problem (4.22) is bounded, but may be infeasible. We claim it will be feasible whenever κ is sufficiently large. Indeed, let $\hat{\mathbf{f}}_i \in \mathcal{H}$ be any functions from the RKHS. By scaling, we can always ensure (4.22c) is satisfied. The following convex optimization $\min_{\mathbf{x}:\mathbf{A}_j\mathbf{x}=\mathbf{b}_j,\mathbf{x}\geq\mathbf{0}} \hat{\mathbf{f}}(\mathbf{x}_j)^T\mathbf{x}$ is bounded and satisfies a Slater condition by Assumption **A2**. Let $\hat{\mathbf{y}}_j$ be the dual variables to this optimization, so that $\hat{\mathbf{y}}_j$ satisfy (4.22a) and define $\hat{\epsilon}_j$ according to (4.22b). Then as long as $\kappa \geq \|\hat{\boldsymbol{\epsilon}}\|$, Problem (4.22), and consequently Problem (4.24), will be feasible and obtain an optimal solution.

Computationally, treating the possible infeasibility of (4.24) can be cumbersome, so in what follows, we find it more convenient to dualize this constraint so that the objective becomes,

$$\min_{\boldsymbol{\alpha},\mathbf{y}} \boldsymbol{\alpha}^T K\boldsymbol{\alpha} + \lambda\|\boldsymbol{\epsilon}\|, \tag{4.25}$$

and then solve this problem for various choices of $\lambda > 0$. Note this version of the problem is always feasible, and, indeed, we will employ this formulation later in Sec. 4.8.

We conclude this section by contrasting our parametric and nonparametric formulations. Unlike the parametric approach, the nonparametric approach is always a convex optimization problem. This highlights a key tradeoff in the two approaches. The parametric approach offers us fine-grained control over the specific form of the function $\mathbf{f}$ at the potential expense of the tractability of the optimization. The nonparametric approach offers less control but is more tractable.

We next illustrate our nonparametric approach below with an example.

4.5.2 Application: Estimating Cost Functions in Wardrop Equilibrium.

Recall the example of Wardrop equilibrium from Section 4.2.1. In practice, while the network $(\mathcal{V}, \mathcal{A})$ is readily observable, the demands $\mathbf{d}^w$ and cost function $c_a(\cdot)$ must be estimated. Although several techniques already exist for estimating the demands $\mathbf{d}^w$ ([1], [114]), there are fewer approaches for estimating $c_a(\cdot)$. Those techniques that do exist often use stylized networks, e.g., one origin-destination pair, to build insights. See [88] for a maximum likelihood approach, and [96] for kinematic wave analyses.

By contrast, we focus on estimating $c_a(\cdot)$ from observed flows or traffic counts on real, large scale networks. Specifically, we assume we are given networks $(\mathcal{V}_j, \mathcal{A}_j)$, $j = 1, \ldots, N$, and have access to estimated demands on these networks $\mathbf{d}^{\mathbf{w}_j}$ for all $\mathbf{w}_j \in W_j$. In practice, this may be the same network observed at different times of day, or different times of year, causing each observation to have different demands.

In the transportation literature, one typically assumes that $c_a(\cdot)$ only depends on arc a, and in fact, can be written in the form $c_a(x_a) = c_{0a} g\left(\frac{x_a}{m_a}\right)$, for some nondecreasing function g. The constant c_{0a} is sometimes called the free-flow travel time of the arc, and m_a is the effective capacity of the arc. These constants are computed from particular characteristics of the arc, such as its length, the number of lanes or the posted speed limit. (Note the capacity m_a is not a hard constraint; it not unusual to see arcs where $x_a^* > m_a$ in equilibrium.) We will also assume this form for the cost function, and seek to estimate the function $g(\cdot)$.

Using (4.24) and (4.25) we obtain the quadratic optimization problem

$$\min_{\boldsymbol{\alpha}, \mathbf{y}, \boldsymbol{\epsilon}} \quad \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \lambda \|\boldsymbol{\epsilon}\| \tag{4.26}$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{e}_a^T \mathbf{N}_j^T \mathbf{y}^{\mathbf{w}} \leq c_{0a} \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a, \quad \forall \mathbf{w} \in W_j,\ a \in \mathcal{A}_j,\ j = 1, \ldots, N, \\ & \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a \leq \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_{a'}, \qquad \forall a, a' \in \mathcal{A}_0 \quad \text{s.t. } \frac{x_a}{m_a} \leq \frac{x_{a'}}{m_{a'}}, \\ & \sum_{a \in \mathcal{A}_j} c_{0a} x_a \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a - \sum_{\mathbf{w} \in W_j} (\mathbf{d}^{\mathbf{w}})^T \mathbf{y}^{\mathbf{w}} \leq \epsilon_j, \quad , \forall \mathbf{w} \in W_j,\ j = 1, \ldots, N, \\ & \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_{a_0} = 1. \end{aligned} \tag{4.27}$$

In the above formulation $\mathcal{A}_0$ is a subset of $\bigcup_{j=1}^{N} \mathcal{A}_j$ and $\mathbf{K} \in \mathbb{R}^{\sum_{j=1}^{N}|\mathcal{A}_j| \times \sum_{j=1}^{N}|\mathcal{A}_j|}$. Constraint (4.27) enforces that the function $g(\cdot)$ be non-decreasing on these arcs. Finally, a_0 is some (arbitrary) arc chosen to normalize the function.

Notice, the above optimization can be quite large. If the various networks are of similar size, the problem has $O(N(|\mathcal{A}_1|+|W_1||\mathcal{V}_1|)$ variables and $O(N|W_1||\mathcal{A}_1|+|\mathcal{A}_0|)$ constraints. As mentioned previously, however, this optimization exhibits significant structure. First, for many choices of kernel, the matrix K is typically (approximately) low-rank. Thus, it is usually possible to reformulate the optimization in a much lower dimensional space. At the same time, for a fixed value of α, the optimization decouples by $\mathbf{w} \in W_j$ and j. Each of these subproblems, in turn, is a shortest path problem which can be solved very efficiently, even for large-scale networks. Thus, combining an appropriate transformation of variables with block decomposition, we can solve fairly large instances of this problem. We take this approach in Section 4.8.3.

4.6 Extensions

Before proceeding, we note that Theorem 4.4 actually holds in a more general setting. Specifically, a minimization over an RKHS will admit a solution of the form (4.23) whenever

a) the optimization only depends on the norms of the components $\|f_i\|_{\mathcal{H}}$ and the function evaluated at a finite set of points $\mathbf{f}(\mathbf{x}_j)$, and

b) the objective is nondecreasing in the norms $\|f_i\|_{\mathcal{H}}$.

The proof is identical to the one presented above, and we omit it for conciseness. An important consequence is that we can leverage the finite representation of Theorem 4.4 in a number of other estimation problems and to facilitate inference. In this section, we describe some of these extensions.

4.6.1 Priors and Semi-Parametric Estimation

Suppose we believe a priori that the function $\mathbf{f}$ describing the VI should be close to a particular function $\mathbf{f}_0$ (a prior). In other words, $\mathbf{f} = \mathbf{f}_0 + \mathbf{g}$ for some function $\mathbf{g}$ which we believe is small. We might then solve

$$\begin{aligned}
\min_{\mathbf{g},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \|g_i\|_{\mathcal{H}}^2 \\
\text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq \mathbf{f}_0(\mathbf{x}_j) + \mathbf{g}(\mathbf{x}_j) \qquad j = 1, \ldots, N, \\
& \mathbf{x}_j^T (\mathbf{f}_0(\mathbf{x}_j) + \mathbf{g}(\mathbf{x}_j)) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j \quad j = 1, \ldots, N, \\
& \|\boldsymbol{\epsilon}\| \leq \kappa, \quad \boldsymbol{\epsilon} \geq \mathbf{0}, \quad g_i \in \mathcal{H}, \qquad i = 1, \ldots, n.
\end{aligned}$$

From our previous remarks, it follows that this optimization is equivalent to

$$\begin{aligned}
\min_{\boldsymbol{\alpha},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha}^T \mathbf{e}_i \\
\text{s.t.} \quad & \mathbf{A}_j \mathbf{y}_j \leq \mathbf{f}_0(\mathbf{x}_j) + \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j \quad j = 1, \ldots, N \\
& \mathbf{x}_j^T (\mathbf{f}_0(\mathbf{x}_j) + \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j \quad j = 1, \ldots, N, \\
& \|\boldsymbol{\epsilon}\| \leq \kappa, \quad \boldsymbol{\epsilon} \geq \mathbf{0},
\end{aligned}$$

which is still a convex quadratic optimization problem.

In a similar way we can handle semi-parametric variants where $\mathbf{f}$ decomposes into the sum of two functions, one of which is known to belong to a parametric family and the other of which is defined nonparametrically, i.e., $\mathbf{f}(\cdot) = \mathbf{f}_0(\cdot; \boldsymbol{\theta}) + \mathbf{g}$ for some $\boldsymbol{\theta}$ and $\mathbf{g} \in \mathcal{H}^n$.

Remark 4.4 (A Challenge with Partial Derivatives). There are natural modeling circumstances where Theorem 4.4 is not applicable. For example, recall in our demand estimation example from Section 4.3.3 that the inverse variational inequality problem depends not only on the demand functions $D_1(\cdot), D_2(\cdot)$ evaluated at a finite set of points (p_1^j, p_2^j), but also on their partial derivatives at those points. Intuitively, the partial derivative $\partial_i D_i(p_1^j, p_2^j)$ requires information about the function in a small

neighborhood of (p_1^j, p_2^j), not just at the point, itself. Consequently, Theorem 4.4 is not applicable. This is one of the reasons we choose to work directly with the marginal revenue functions in this example. Extending the above techniques to this case remains an open area of research.

4.6.2 Ambiguity Sets

In many applications, there may be multiple distinct models which all reconcile the data equally well. Breiman termed this phenomenon the "Rashomon" effect [40]. It can occur even with parametric models that are well-identified, since there may exist models outside the parametric family which will also reconcile the data. (See, e.g., Sec. 4.8.1.) Consequently, we would often like to identify the range of functions which may explain our data, and how much they differ.

We can determine this range by computing the upper and lower envelopes of the set of all functions within an RKHS that make the observed data approximate equilibria. We call this set the ambiguity set for the estimator. To construct these upper and lower bounds on the ambiguity set, consider fixing the value of κ in (4.22) and replacing the objective by $f_i(\hat{\mathbf{x}})$ for some $\hat{\mathbf{x}} \in \mathcal{F}$. This optimization problem satisfies the two conditions listed at the beginning of this section. Consequently, Theorem 4.4 applies, and we can use the finite representation to rewrite the optimization problem as a linear optimization problem in $\boldsymbol{\alpha}, \mathbf{y}$. Using software for linear optimization, it is possible to generate lower and upper bounds on the function $\mathbf{f}(\hat{\mathbf{x}})$ for various choices of $\hat{\mathbf{x}}$ quickly and efficiently.

To what value should we set the constant κ? One possibility is to let κ be the optimal objective value of (4.12) , or a small multiple of it. This choice of κ yields the set of functions which "best" reconcile the given data. We discuss an alternative approach in Section 4.7 that yields a set of functions which are statistically similar to the current estimator.

Regardless of how we choose, κ, though, ambiguity sets can be combined with our previous parametric formulations to assess the appropriateness of the particular choice of parametric family. Indeed, the ambiguity set formed from the nonparametric kernel

contains a set of alternatives to our parametric form which are, in some sense, equally plausible from the data. If these alternatives have significantly different behavior from our parametric choice, we should exercise caution when interpreting the fitted function.

Can we ever resolve the Rashomon effect? In some cases, we can use application-specific knowledge to identify a unique choice. In other cases, we need appeal to some extra, a priori criterion. A typical approach in machine learning is to focus on a choice with good generalizability properties. In the next section, we show that our proposed estimators enjoy such properties.

4.7 Generalization Guarantees

In this section, we seek to prove generalization guarantees on the estimators from Problem (4.12) and (4.22). Proving various types of generalization guarantees for algorithms is a central problem in machine learning. These guarantees ensure that the performance of our estimator on new, future data will be similar to its observed performance on existing data.

We impose a mild assumption on the generating process which is common throughout the machine learning literature:

Assumption 3. *The data* $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ *are i.i.d. realizations of random variables* $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ *drawn from some probability measure* $\mathbb{P}$*.*

Notice, we make no assumptions on potential dependence between $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$, nor do we need to know the precise form of $\mathbb{P}$. We adapt Assumption **A2** to this context as

Assumption 4. *The random set* $\tilde{\mathcal{F}} = \{\mathbf{x} : \tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}, \mathbf{x} \in \tilde{C}\}$ *satisfies a Slater Condition almost surely.*

and add

Assumption 5. $\tilde{\mathbf{x}} \in \tilde{\mathcal{F}}$ *almost surely.*

Assumptions 4 and 5 are not particularly stringent. If these condition may fail, we can consider pre-processing the data so that they succeed, and then consider a new measure $\mathbb{Q}$ induced by this processing of $\mathbb{P}$.

We now prove a bound for a special case of Problem (4.12). Let $z_N, \boldsymbol{\theta}_N$ denote the optimal value and optimal solution of (4.12). If for some N, there exist multiple optimal solutions, choose $\boldsymbol{\theta}_N$ by some tie-breaking rule, e.g., the optimal solution with minimal ℓ_2-norm. For any $0 < \alpha < 1$, define

$$\beta(\alpha) \equiv \sum_{i=0}^{\dim(\boldsymbol{\theta})} \binom{N}{i} \alpha^i (1-\alpha)^{N-i}.$$

Theorem 4.6. *Consider Problem* (4.12) *where the norm* $\|\cdot\| = \|\cdot\|_\infty$. *Suppose that this problem is convex in* $\boldsymbol{\theta}$ *and that Assumptions* ***A1****,* ***A3****-****A5*** *hold. Then, for any* $0 < \alpha < 1$*, with probability at least* $1 - \beta(\alpha)$ *with respect to the sampling,*

$$\mathbb{P}\left(\tilde{\mathbf{x}} \textit{ is a } z_N\textit{-approximate equilibrium for } \textit{VI}(\mathbf{f}(\cdot, \boldsymbol{\theta}_N), \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})\right) \geq 1 - \alpha.$$

The proof relies on relating Problem (4.12) to an uncertain convex program [44], and leveraging results on the randomized counterparts of such programs. See the appendix for the details.

Remark 4.5. There are two probability measures in the theorem. The first (explicit) is the probability measure of the new data point $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$. The second (implicit) is the probability measure of the random quantities z_N, $\boldsymbol{\theta}_N$. One way to interpret the theorem is as follows: One can ask, "For a fixed pair $z_N, \boldsymbol{\theta}_N$, is the probability that $\mathbf{x}_{N+1}$ is a z_N-approximate equilibrium for $\text{VI}(\mathbf{f}(\cdot, \boldsymbol{\theta}_N), \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$ with respect to the first measure at least $1 - \alpha$?" The theorem asserts the answer is, "Yes" with probability at least $1 - \beta(\alpha)$ with respect to the second measure. More loosely, the theorem asserts that for "typical" values of $z_N, \boldsymbol{\theta}_N$, the answer is "yes." This type of generalization result, i.e. result which is conditional on the data-sampling measure, is typical in machine learning.

Remark 4.6. Notice that $\beta(\alpha)$ corresponds to the tail probability of a binomial dis-

tribution, and, hence, converges exponentially fast in N.

Remark 4.7 (ℓ_1 Regularization). The value $\beta(\alpha)$ depends strongly on the dimension of θ. In [43], the authors show that including an ℓ_1 regularization of the form $\|\theta\|_1$ to reduce the effective dimension of θ can significantly improve the above bound in the context of uncertain convex programs.[3] Motivated by this idea, we propose modifying our original procedure by including a regularization $\lambda\|\boldsymbol{\theta}\|_1$ in the objective of Problem (4.12) where λ is chosen exogenously. Since the problem is convex this formulation is equivalent to including a constraint of the form $\|\boldsymbol{\theta}\|_1 \leq \kappa$ for some value of κ that implicitly depends on λ, and, consequently, Theorem 4.6 still applies but with z_N redefined to exclude the contribution of the regularization to the objective value.

Unfortunately, the proof of Theorem 4.6 does not generalize easily to other problems, such as other norms or Problem (4.22). A more general approach to proving generalization bounds is based upon Rademacher complexity. Rademacher complexity is a popular measure of the complexity of a class of functions, related to the perhaps better known VC-dimension. Loosely speaking, for function classes with small Rademacher complexity, empirical averages of functions in the class converge to their true expectation uniformly over the class, and there exist bounds on the rate of convergence which are tight up to constant factors. We refer the reader to [11] for a formal treatment.

We will use bounds based upon the Rademacher complexity of an appropriate class of functions to prove generalization bounds for both our parametric and nonparametric approaches. We limit our analysis to the case of ℓ_p-norm, $1 \leq p < \infty$in the objective of (4.12) and (4.22). To simplify the statement of our results we replace the objectives of both problems by $\frac{\|\boldsymbol{\epsilon}\|_p^p}{N}$. This monotonic transformation affects the optimal value, but not the optimal solution.

Moreover, in the case of our nonparametric approach, it will prove easier to analyze

[3]In fact, the authors show more: they give an algorithm leveraging ℓ_1 regularization to reduce the dimensionality of θ and then an improved bound based on the reduced dimension. The analysis of this improved bound can be adapted to our current context at the expense of more notation. We omit the details for space.

the following modified optimization problem instead of Problem (4.22):

$$\begin{aligned}
\min_{\mathbf{f},\mathbf{y},\boldsymbol{\epsilon}} \quad & \frac{\|\boldsymbol{\epsilon}\|_p^p}{N} \\
\text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq \mathbf{f}(\mathbf{x}_j), && j = 1, \ldots, N, \\
& \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j, && j = 1, \ldots, N, \\
& \|f_i\|_{\mathcal{H}}^2 \leq \kappa_i, \quad f_i \in \mathcal{H}, && i = 1, \ldots, n, \\
& \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) = 1.
\end{aligned} \tag{4.28}$$

Specifically, we have first used the dualized objective eq. (4.25), and secondly, using lagrangian duality again, moved the term $\lambda \sum_{i=1}^{n} \|f_i\|_{\mathcal{H}}$ from the objective to the constraints. Indeed, for any value of λ, there exists values κ_i so that these two problems are equivalent. Intuitively, if we can show that solutions to Problem (4.28) enjoy strong generalization guarantees, Problem (4.22) should satisfy similar guarantees.

Now, introduce

Assumption 6. *The set $\tilde{\mathcal{F}}$ is contained within a ball of radius R almost surely.*

Next, consider Problem (4.12) with our modified objective. Define

$$2 \sup_{\substack{\mathbf{x}:\|\mathbf{x}\|_2 \leq R \\ \boldsymbol{\theta} \in \Theta}} \|\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})\|_2 \equiv \overline{B}. \tag{4.29}$$

Observe that if Θ is compact and $\mathbf{f}$ is continuous, $\overline{B} < \infty$. Let $\mathbf{f}_N = \mathbf{f}(\cdot; \boldsymbol{\theta}_N)$ denote the function corresponding to the optimal solution of Problem (4.12). With a slight abuse of language, we call $\mathbf{f}_N$ a solution to Problem (4.12).

We define analogous quantities for Problem (4.28). Given a kernel $k(\cdot, \cdot)$, let $\overline{K}^2 \equiv \sup_{\mathbf{x}:\|\mathbf{x}\|_2 \leq R} k(\mathbf{x}, \mathbf{x})$. Notice, if k is continuous, $\overline{K}$ is finite by **A6**. For example,

$$\overline{K}^2 = \begin{cases} R^2 & \text{for the linear kernel} \\ (c + R^2)^d & \text{for the polynomial kernel} \\ 1 & \text{for the Gaussian kernel} \end{cases} \tag{4.30}$$

With a slight abuse of notation, let $z_N, \mathbf{f}_N$ denote the optimal value and an optimal solution to Problem (4.28), and let $\overline{B} \equiv 2R\overline{K}\sqrt{\sum_{i=1}^{n}\kappa_i^2}$. This mild abuse of notation allows us to express our results in a unified manner. It will be clear from context whether we are treating Problem (4.12) or Problem (4.28), and consequently be clear which definition of $\mathbf{f}_N, \overline{B}$ we mean.

Finally, define $\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ to be the smallest $\epsilon \geq 0$ such that $\tilde{\mathbf{x}}$ is an ϵ-approximate solution to $\text{VI}(\mathbf{f}_N, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$.

Theorem 4.7. *Let $z_N, \mathbf{f}_N$ be the optimal objective and an optimal solution to Problem* (4.12) *or* (4.28)*. Assume **A1**, **A3**-**A6**. For any $0 < \beta < 1$, with probability at least $1 - \beta$ with respect to the sampling,*

i)

$$\mathbb{E}[(\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}))^p] \leq z_N + \frac{1}{\sqrt{N}}\left(4p\overline{B}^p + 2\overline{B}^{p/2}\sqrt{2\log(2/\beta)}\right). \tag{4.31}$$

ii) For any $\alpha > 0$,

$$\begin{aligned}\mathbb{P}(&\tilde{\mathbf{x}} \textit{ is a } z_N{+}\alpha\textit{-approximate equilibrium for VI}(\mathbf{f}_N, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})) \\ &\geq 1 - \frac{1}{\alpha^p\sqrt{N}}\left(4p\overline{B}^p + 2\overline{B}^{p/2}\sqrt{2\log(2/\beta)}\right).\end{aligned}$$

Remark 4.8. To build some intuition, consider the case $p = 1$. The quantity z_N is the average error on the data set for $\mathbf{f}_N$. The theorem shows with high-probability, $\mathbf{f}_N$ will make a new data point an ϵ-approximate equilibrium, where ϵ is only $O(1/\sqrt{N})$ larger than z_N. In other words, the fitted function will perform not much worse than the average error on the old data. Note, this does not guarantee that z_N is small. Indeed, z_N will only be small if in fact a VI is a good model for the system.

Remark 4.9 (Specifying Ambiguity Sets). We can use Theorem 4.7 to motivate an alternate proposal for specifying κ in ambiguity sets as in Section 4.6. Specifically, let R_N denote the second term on the righthand side of (4.31). Given another feasible function $\mathbf{f}'$ in Problem (4.28) whose objective value is strictly greater than $z_N + R_N$, we

can claim that with probability at least $1-\beta$, $\mathbf{f}_N$ has a smaller expected approximation error than $\mathbf{f}'$. However, if the objective value of $\mathbf{f}'$ is smaller than $z_N + R_N$, we cannot reject it at level $1-\beta$; it is statistically as plausible as $\mathbf{f}_N$. Setting $\kappa = R_N$ in our ambiguity set recovers all such "statistically plausible" functions.

Theorems 4.6 and 4.7 provide a guarantee on the generalization error of our method. We may also be interested in its predictive power. To that end, given a new data point $(\mathbf{x}_{N+1}, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$, let $\hat{\mathbf{x}}$ be a solution to $\text{VI}(\mathbf{f}_N, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1}))$ The value $\hat{\mathbf{x}}$ is a prediction of the state of a system described by $(\mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$ using our fitted function, and $\mathbf{x}_{N+1}$ represents true state of that system. We have the following theorem:

Theorem 4.8. *Assume $\mathbf{f}_N$ is strongly monotone with parameter γ.*

i) Suppose the conditions of Theorem 4.6 hold. Then, for any $0 < \alpha < 1$, with probability at least $1 - \beta(\alpha)$ with respect to the sampling,

$$\|\mathbf{x}_{N+1} - \hat{\mathbf{x}}\| \leq \sqrt{\frac{z_N}{\gamma}}.$$

ii) Suppose the conditions of Theorem 4.7 hold. Then, for any $0 < \beta < 1$, with probability at least $1 - \beta$ with respect to the sampling, for any $\alpha > 0$

$$\mathbb{P}\left(\|\mathbf{x}_{N+1} - \hat{\mathbf{x}}\| > \sqrt{\frac{z_N + \alpha}{\gamma}}\right) \leq \frac{1}{\alpha^p \sqrt{N}} \left(4p\overline{B}^p + 2\overline{B}^{p/2}\sqrt{2\log(2/\beta)}\right).$$

In words, Theorem 4.8 asserts that solutions to our VI using our fitted function serve as good predictions to future data realizations. This is an important strength of our approach as it allows us to predict future behavior of the system. Again, this is contingent on the fact that z_N is small, i.e., that the VI well-explains the current data.

We conclude this section by noting that experimental evidence from machine learning suggests that bounds such as those above based on Rademacher complexity can be loose in small-samples. The recommended remedy is that, when computationally

feasible, to use a more numerically intensive method like cross-validation or bootstrapping to estimate approximation and prediction errors. This approach applies equally well to choosing parameters like the threshold in an ambiguity set κ as described in Remark 4.9. We employ both approaches in Section 4.8.

4.8 Computational Experiments

In this section, we provide some computational experiments illustrating our approach. For concreteness, we focus on our two previous examples: estimating the demand function in Bertrand-Nash equilibrium from Sec. 4.3.3 and estimating cost functions in traffic equilibrium from Sec. 4.5.2.

Before providing the details of the experiments, we summarize our major insights.

1. In settings where there are potentially many distinct functions that explain the data equally well, our nonparametric ambiguity sets are able to identify this set of functions. By contrast, parametric methods may misleadingly suggest there is only one possible function.

2. Even in the presence of endogenous, correlated noise, our parametric and nonparametric techniques are able to learn functions with good generalizability, even if the specified class does not contain the true function generating the data.

3. Sometimes, the functions obtained by our method are not strongly monotone. Nonetheless, they frequently still have reasonable predictive power.

4.8.1 Bertrand-Nash Equilibrium (Full-Information)

We first consider an idealized, full-information setting to illustrate the importance of our ambiguity set technique. Specifically, we assume the true, demand functions are given by the nonlinear model

$$D_i^*(p_1, p_2, \xi_i) = \log(p_i) + \theta_{i1}^* p_1 + \theta_{i2}^* p_2 + \theta_{i3}^* \xi_i + \theta_{i4}^*, \quad i = 1, 2$$

with $\boldsymbol{\theta}_1^* = [-1.2, .5, 1, -9]^T$ and $\boldsymbol{\theta}_2^* = [.3, -1, 1, -9]^T$. We assume (for now) that although we know the parametric form of these demand functions, we do not know the precise values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and seek to estimate them. The corresponding marginal revenue functions are

$$M_i^*(p_1, p_2, \xi_i; \boldsymbol{\theta}_i^*) = \log(p_i) + \theta_{i1}^* p_1 + \theta_{i2}^* p_2 + \theta_{i3}^* \xi_i + \theta_{i4}^* + 1 + \theta_{ii}^* p_i, \quad i = 1, 2. \quad (4.32)$$

Here ξ_1, ξ_2 are random variables representing firm-specific knowledge which change over time ("demand shocks") causing prices to shift.

Our idealized assumption is that $\xi_1 = \xi_2 \equiv \xi$, and ξ is common knowledge to both the firms and to the researcher (full-information). In our simulations, we take ξ to be i.i.d normals with mean 5 and standard deviation 1.5. Using these parameters with $\overline{p} = .45$, we simulate values of ξ and solve for the equilibrium prices p_1^j, p_2^j for $j = 1, \ldots, 250$. The values (ξ^j, p_1^j, p_2^j) constitute our data set.

To estimate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we substitute the functional form Eq. (4.32) into Problem (4.13), adding additional constraints that 1) the marginal revenue of firm i is positive for the minimal price p_i^j observed in the data, 2) the marginal revenue of firm i is decreasing in firm i's price, and 3) a normalization constraint. (See Appendix B.3.1 for an explicit formulation).

Unsurprisingly, solving this optimization recovers the true marginal revenue functions exactly. We say "unsurprisingly" because with full-information a correctly specified, known parametric form, we believe any reasonable estimation procedure should recover the true marginal revenue functions. We point out that the optimal solution to the optimization problem is *unique*, and the optimal value of the residuals is $\boldsymbol{\epsilon} = 0$

We plot the true marginal revenue functions for each firm (which is the same as our fitted function) in Figure 4-1 (dashed black line). To graph these functions we fixed ξ to be its median value over the dataset, and fixed the other firm's price to be the price observed for this median value. For convenience in what follows, we term this type of fixing of the other variables, *fixing to the median observation.*

Next consider the more realistic setting where we do not know the true para-

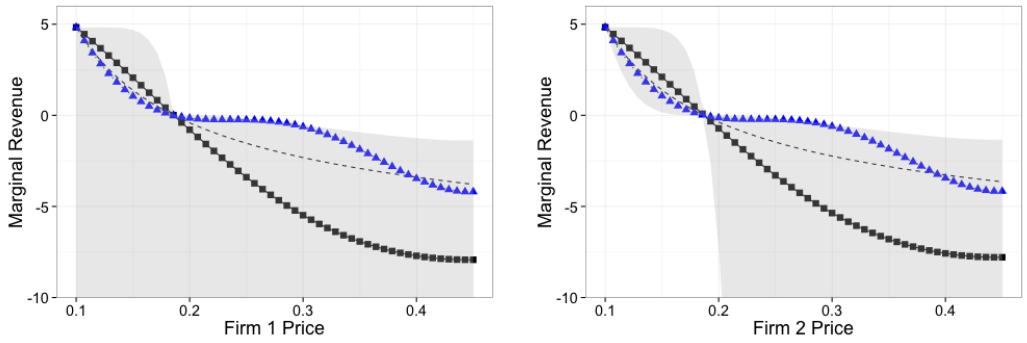


Figure 4-1: An idealized scenario. The true marginal revenue function (dashed black line), our nonparametric fit (black line, square markers), and the ambiguity set (grey region) for both firms. Every function in the ambiguity set *exactly* reconciles all the data. A sample member (blue line, triangle markers) shown for comparison. All variables other than the firm's own price have been fixed to the median observation.

metric form (4.32), and so use our nonparametric method (cf. Problem 4.22 with dualized objective (4.25)). We use a Gaussian kernel and tune the parameter c and regularization constant λ by 10-fold cross-validation. The resulting fitted function is shown in Figure 4-1 as a black line with square markers. Notice, in particular, this function does not coincide with the true function. However, this function also *exactly* reconciles the data, i.e. the optimal value of the residuals is $\boldsymbol{\epsilon} = 0$. This may seem surprising; the issue is that although there is only one function within the parametric family (4.32) which reconciles the data, there are many potential smooth, nonparametric functions which also exactly reconcile this data (Rashomon effect). Using our ambiguity set technique, we compute the upper and lower envelopes of this set of functions, and display the corresponding region as the grey ribbon in Figure 4-1. We also plot a sample function from this set (blue line with triangle markers).

This multiplicity phenomenon is not unusual; many inverse problems share it. Moreover, it often persists even for very large samples N. In this particular case, the crux of the issue is that, intuitively, the equilibrium conditions only give local information about the revenue function about its minimum. (Notice all three marginal revenue functions cross zero at the same price). The conditions themselves give no information about the global behavior of the function, even as $N \to \infty$.

We see our ambiguity set technique and nonparametric analysis as important tools

to protect against potentially faulty inference in these settings. Indeed, parametric estimation might have incorrectly led us to believe that the unique marginal revenue function which recovered the data was the dashed line in Figure 4-1 – its residual error is zero and it is well-identified within the class. We might then have been tempted to make claims about the slope of the marginal revenue function at the optima, or use it to impute a particular functional form for the demand function. In reality, however, ***any function from the ambiguity set might have just as easily generated this data***, e.g., the blue line with triangle markers. Those previous claims about the slope or demand function, then, need not hold. The data does not support them. Calculating nonparametric sets of plausible alternatives helps guard against these types of unwarranted claims.

Finally, in the absence of any other information, we argue that our proposed nonparametric fit (red line with circles) is a reasonable candidate function in this space of alternatives. By construction it will be smooth and well-behaved. More generally, of all those functions which reconcile the data, it has the smallest $\mathcal{H}$-norm, and thus, by our generalization results in Section 4.7, likely has the strongest generalization properties.

4.8.2 Bertrand-Nash Equilibrium (Unobserved Effects)

We now proceed to a more realistic example. Specifically, we no longer assume that $\xi_1 = \xi_2$, but rather these values represent (potentially different), firm-specific knowledge that is unobservable to us (the researchers). We assume instead that we only have access to the noisy proxy ξ. In our simulations, we take ξ_1, ξ_2, ξ' to be i.i.d normal with mean 5 and standard deviation 1.5, and let $\xi = (\xi_1 + \xi_2 + \xi')/3$. Moreover, we assume that we we have incorrectly specified that the marginal revenue functions are of the form

$$M_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = \sum_{k=1}^{9} \theta_{i1k} e^{-kp_1} + \sum_{k=1}^{9} \theta_{i2k} e^{-kp_2} + \theta_{i1} p_1 + \theta_{i2} p_2 + \theta_{i3} \xi_3 + \theta_{i4} \quad (4.33)$$

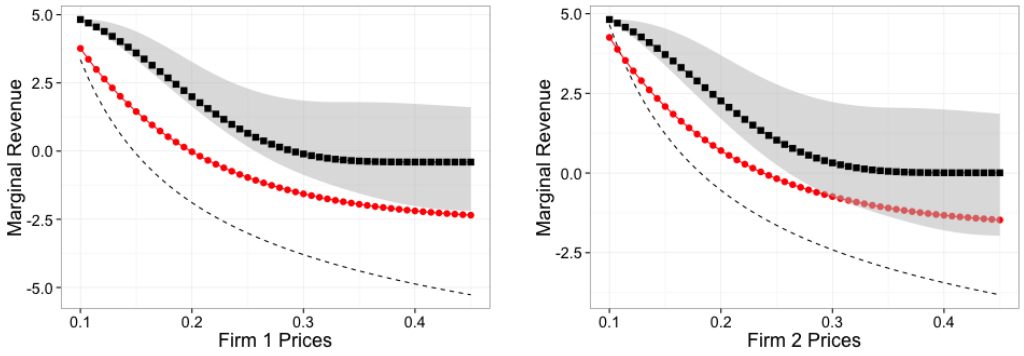


Figure 4-2: The true marginal revenue function (dashed line), fitted parametric marginal revenue function (solid red line, circular markers), fitted non-parametric marginal revenue function (solid black line, square markers)and ambiguity sets (grey region) for each firm. We fix all variables except the firm's own price to the median observation.

for some values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Notice that the true parametric is not contained in this class. This setup thus includes correlated noise, endogenous effects, and parametric mispecification. These features are known to cause statistical challenges in simple estimation procedures. We simulate $N = 40$ observations (ξ^j, p_1^j, p_2^j) from this model.

We again fit this model first by solving a modification of Problem (4.13) as before. (See Appendix B.3.1 for an explicit formulation). We only use half the data (20 observations) for reasons that will become clear momentarily. We use the ℓ_∞-norm for the residuals $\boldsymbol{\epsilon}$ and an ℓ_1-regularization of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ in the objective as discussed in Remark 4.7. We tune the value of λ in the regularization to minimize the mean squared error in price prediction obtaining the value $\lambda = .01$.

Unfortunately, because we used cross-validation to choose λ, Theorem 4.6 does not directly apply. Consequently, we now refit $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ with $\lambda = .1$ using the other half of our training set. The fitted marginal revenue functions for $\lambda = .01$ can be seen in Figure 4-2 (red line, circular markers). Notice that the fitted function does not exactly recover the original function, but does recover its approximate shape.

To assess the out of sample performance of this model, we generate a new set of $N_{out} = 200$ points. For each point we compute the approximation error (minimal ϵ to make this point an ϵ-approximate equilbria) and the prediction error had we

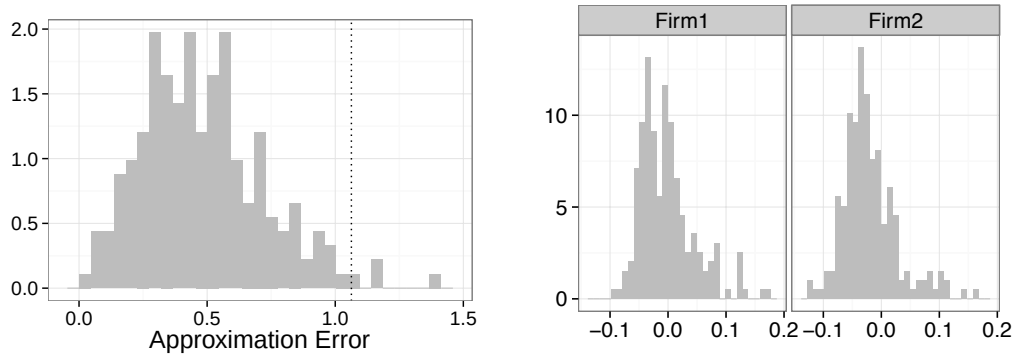


Figure 4-3: Bertrand-Nash example of Section 4.8.2. The left panel shows the out-of-sample approximation error. The right panel shows the out-of-sample prediction error.

attempted to predict this point by the solution to our VI with our fitted function. Histograms of both quantities are in Fig. 4-3.

The maximal residual on the second half of the training set was $z_N \approx 1.06$, indicated by the dotted line in the left panel. By Theorem 4.6, we would expect that with at least 90% probability with respect to the data sampling, a new point would not be an 1.06-equilibrium with probability at most .21. Our out-of-sample estimate of this probability is .025. In other words, our estimator has much stronger generalization than predicted by our theorem. At the same time, our estimator yields reasonably good predictions. The mean out-of-sample prediction error is $(-.002, 0.02)$ with standard deviation $(.048, .047)$.

Finally, we fit our a nonparametric estimator to this data, using a Gaussian kernel. We again tune the parameter c and regularization constant λ by cross-validation. The resulting fit is shown in Figure 4-2 (black line, square markers), along with the corresponding ambiguity set. We chose the value of κ to be twice the standard deviation of the ℓ_1-norm of the residuals, estimated by cross-validation as discussed in Remark 4.9 and the end of Sec. 4.7. The out-of-sample approximation error is similar to the parametric case. Unfortunately, the fitted function is not monotone, and, consequently, there exist multiple Nash equilibria. It is thus hard to compare prediction error on the out-of-sample set; which equilibria should we use to predict?

This non-monotonicity is a potential weakness of the nonparametric approach in this example.

4.8.3 Wardrop Equilibrium

Our experiments will use the Sioux Falls network [82], a standard benchmark throughout the transportation literature. It is modestly sized with 24 nodes and 76 arcs, and all pairs of nodes represent origin-destination pairs.

We assume that the true function $g(\cdot)$ is given by the U.S. Bureau of Public Roads (BPR) function, $g(t) = 1 + .15t^4$ which is by far the most commonly used for traffic modeling ([39], [98]). Baseline demand levels, arc capacities, and free-flow travel times were taken from the repository of traffic problems at [73]. We consider the network structure including arc capacities and free-flow travel times as fixed. We generate data on this network by first randomly perturbing the demand levels a relative amount drawn uniformly from $[0, 10\%]$. We then use the BPR function to solve for the equilibrium flows on each arc, x_a^*. Finally, we perturb these true flows by a relative amount, again drawn uniformly from $[0, 10\%]$. We repeat this process $N = 40$ times. Notice that because both errors are computed as relative perturbations, they both are correlated to the observed values. We use the perturbed demands and flows as our data set.

We then fit the function g nonparametrically using (4.26), again only using half of the data set. The use of low order polynomials in traffic modeling is preferred in the literature for a number of computational reasons. Consequently, we choose k to be a polynomial kernel with degree at most 6, and tune the choice of c by 5-fold cross-validation, minimizing the approximation error. The fitted functions for various choices of d are shown in left panel of Figure 4-4, alongside the true function. Notice that the fit is quite stable to choice of class of function, and matches the true function very closely. In what remains, we focus on our fit of polynomial of degree 3. We note, this class does not contain the true BPR function (which has degree 4). We refit the degree 3 polynomial with the second-half of our training set (not shown).

To assess the quality of our degree 3 fit, we create a new out-of-sample test-set of

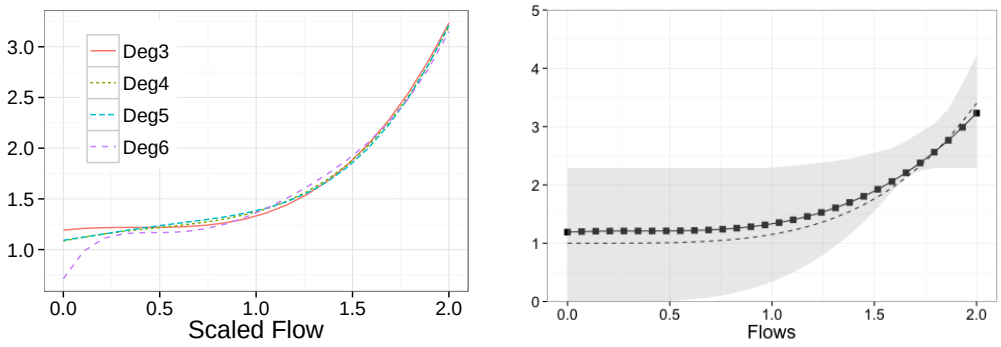


Figure 4-4: The left panel shows the true BPR function and fits based on polynomials of degrees 3, 4, 5, and 6. The right panel shows the true BPR function (dashed line), our degree 3 fit (solid black line with markers), and an ambiguity set around this function (grey region).

size $N_{out} = 500$. On each sample we compute the approximation error of our fit and the ℓ_2-norm of the prediction error when predicting new flows by solving the fitted VI. These numbers are large and somewhat hard to interpret. Consequently we also compute normalized quantities, normalizing the first by the minimal cost of travel on that network with respect to the fitted function and demands, and the second by the ℓ_2 norm of the observed flows. Histograms for the normalized quantities are shown in Figure 4-5. The mean (relative) approximation error is 6.5%, while the mean predictive (relative error) is about 5.5%.

The in-sample approximation error on the second-half of the training sample was $z_N \approx 8.14 \times 10^5$. By Theorem 4.7, we can compute that with probability at least 90% with respect to the data sampling, a new data point will be at most a 9.73×10^5 approximate equilibrium with respect to the fitted function with probability at least 90%. A cross-validation estimate of the same quantity is 6.24×10^5. Our out of sample estimate of this quantity from the above histogram is 7.78×10^5. In other words, the performance of our estimator is again better than predicted by the theorem. Cross-validation provides a slightly better, albeit biased, bound.

Finally, as in the previous section, we consider constructing an ambiguity set around the fitted function, selecting κ to be two standard deviations as computed by cross-validation. The resulting envelopes are also shown in the right panel of Figure 4-4. Notice that in contrast to the envelopes of the previous section, they are

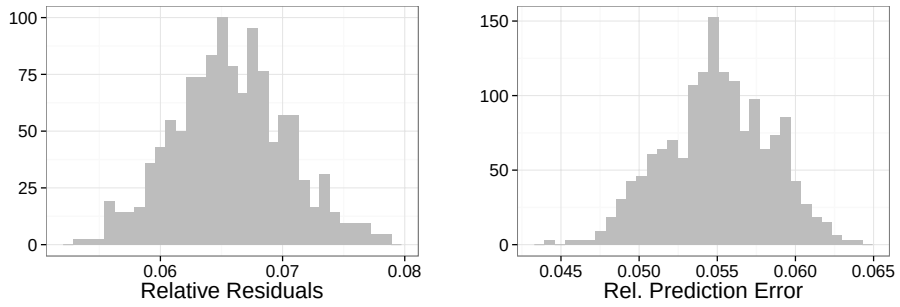


Figure 4-5: The left panel shows the histogram of of out-sample approximation errors induced by our nonparametric fit from Section 4.8.3. The right panel shows the norm of the difference of this flow from the observed flow, relative to the norm of the observed flow.

quite small, meaning we can have relatively high confidence in the shape of the fitted function.

4.9 Conclusion

In this chapter, we proposed a computationally tractable technique for estimation in equilibrium based on an inverse variational inequality formulation. Our approach can be used to fit many common paradigms for modeling behavior to data. We prove our method enjoys strong theoretical generalization guarantees and, also illustrate its usage in two applications – – demand estimation under Nash equilibrium and congestion function estimation under user equilibrium . Our results suggest this technique can successfully model systems presumed to be in equilibrium and make meaningful predictive claims about them.

Chapter 5

Concluding Remarks

The prevalence of high quality data is reshaping operations research, and a new data-centered paradigm is emerging. In this thesis, we proposed several new, data-driven models for uncertainty and human behavior. Throughout, the emphasis was on developing general purpose methodologies with provably good performance and practical relevance. We strived to ensure that our new data-driven models integrated naturally with existing, successful optimization paradigms.

Specifically:

- We proposed a new, data-driven schema for constructing uncertainty sets for robust optimization. We applied that schema to create a host of new uncertainty sets. As with more conventional uncertainty sets, solutions to robust optimization problems built from our sets enjoy a strong performance. Unlike conventional sets, however, our sets are able to learn the underlying features of a data distribution. We demonstrate through computational experiments that our new sets outperform conventional sets whenever data is available.

- We demonstrated the practical merits of our sets through an extensive case study of a real-world, large-scale application, namely unit commitment. In the process, we proposed a methodology for applying our set constructions to time-series data.

- Finally, we presented a new estimation procedure through an inverse varia-

tional inequality formulation for calibrating behavioral models. Importantly, we proved that our procedure satisfies a strong generalization guarantee, even when the underlying model may be misspecified. We again illustrated the approach through a variety of computational examples.

In each case, the key ideas involved drawing new connections between an existing optimization paradigm or technique and a statistical procedure or machine learning methodology. In particular:

- In our uncertainty set constructions, we leveraged a new connection between the well-established theory of hypothesis testing with robust optimization. This connection has a number of implications in a variety of optimization fields extending beyond uncertainty set design. Moreover, as we discussed, empirically vetted techniques from hypothesis testing might be leveraged to improve optimization approaches. As an example, we showed how the bootstrap algorithm might be used to tune the size of data-driven uncertainty sets, but there are many other exciting opportunities in this direction.

- In solving the UC problem, we combined a historical dataset of problem instances with machine learning techniques to identify specific techniques for improving the optimization algorithm. Developing similar techniques for other repeated optimization settings is another promising avenue of research.

- Finally, we proposed a new application of kernel methods in solving underdetermined inverse problems. Although we focused on the inverse variational inequality problem, this approach undoubtedly can be applied in other circumstances.

Overall, these connections have both practical implications (in terms of new algorithms and refined numerical procedures) and theoretical implications (in terms of new perspectives and avenues of future research.)

Appendix A

Supplement to Chapt. 2

A.1 Omitted Proofs

A.1.1 Proof of Proposition 2.2

Proof. Let $\Delta_j \equiv \frac{\hat{p}_j - p_j}{p_j}$. Then,

$$D(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{j=0}^{n-1} \hat{p}_j \log(\hat{p}_j/p_j) = \sum_{j=0}^{n-1} p_j(\Delta_j + 1)\log(\Delta_j + 1).$$

Using a Taylor expansion of $x \log x$ around $x = 1$ yields,

$$D(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{j=0}^{n-1} p_j \left(\Delta_j + \frac{\Delta_j^2}{2} + O(\Delta_j^3)\right) = \sum_{j=0}^{n-1} \frac{(\hat{p}_j - p_j)^2}{2p_j} + \sum_{j=0}^{n-1} O(\Delta_j^3),$$

where the last equality follows by expanding out terms and observing that $\sum_{j=0}^{n-1} \hat{p}_j = \sum_{j=0}^{n-1} p_j = 1$. Thus, the constraint defining $\mathcal{P}^{\chi^2}$ and the constraint defining $\mathcal{P}^G$ are identical up to a term of size $\sum_{j=0}^{n-1} O(\Delta_j^3)$.

Next, note $\mathbf{p} \in \mathcal{P}^G \implies \hat{p}_j/p_j \leq \exp(\frac{\chi^2_{n-1,1-\delta}}{2N\hat{p}_j})$. From the Strong Law of Large Numbers, for any $0 < \delta' < 1$, there exists M such that $\hat{p}_j \geq p_j^*/2$ with probability at least $1 - \delta'$ for all $j = 0, \ldots, n-1$, simultaneously. It follows that for N sufficiently large, with probability $1 - \delta'$, $\mathbf{p} \in \mathcal{P}^G \implies \hat{p}_j/p_j \leq \exp(\frac{\chi^2_{n-1,1-\delta}}{Np_j^*})$ which implies that $|\Delta_j| \leq \exp(\frac{\chi^2_{n-1,1-\delta}}{Np_j^*}) - 1 = O(N^{-1})$. This proves the claim. □

A.1.2 Proof of Theorem 2.6

Proof. We will show that $\phi_{\mathcal{U}^{FB}}(\mathbf{x})$ is given by Eq. (2.25). First observe

$$\max_{\mathbf{u}\in\mathcal{U}^{FB}} \mathbf{u}^T\mathbf{x} = \inf_{\lambda\geq 0} \lambda\log(1/\epsilon) + \max_{\substack{\mathbf{m}_b\leq\mathbf{y}_1\leq\mathbf{m}_b,\\ \mathbf{y}_2\geq,\mathbf{y}_3\geq\mathbf{0}}} \sum_{i=1}^{d} x_i(y_{1i}+y_{2i}-y_{3i}) - \lambda\sum_{i=1}^{d}\frac{y_{2i}^2}{2\sigma_{fi}^2}+\frac{y_{3i}^2}{2\sigma_{bi}^2}$$

by Lagrangian strong duality. The inner maximization decouples by i. The i^{th} subproblem further decouples into three sub-subproblems. The first is $\max_{m_{bi}\leq y_{i1}\leq m_{fi}} x_i y_{1i}$ with optimal value

$$y_{1i} = \begin{cases} m_{fi}x & \text{if } x_i \geq 0, \\ m_{bi}x_i & \text{if } x_i < 0. \end{cases}$$

The second sub-subproblem is $\max_{y_{2i}\geq 0} xy_{2i} - \lambda\frac{y_{2i}^2}{2\sigma_{fi}^2}$. This is maximizing a concave quadratic function of one variable. Neglecting the non-negativity constraint, the optimum occurs at $y_{2i}^* = \frac{x\sigma_{fi}^2}{\lambda}$. If this value is negative, the optimum occurs at $y_{2i}^* = 0$. Consequently,

$$\max_{y_{2i}\geq 0} \; x_i y_{2i} - \lambda\frac{y_{2i}^2}{2\sigma_{fi}^2} = \begin{cases} \frac{x_i\sigma_{fi}^2}{\lambda} & \text{if } x_i \geq 0, \\ 0 & \text{if } x_i < 0. \end{cases}$$

Similarly, we can show that the third subproblem has the following optimum value

$$\max_{y_{3i}\geq 0} \; -x_i y_{3i} - \lambda\frac{y_{3i}^2}{2\sigma_{bi}^2} = \begin{cases} \frac{x_i\sigma_{bi}^2}{\lambda} & \text{if } x_i \leq 0, \\ 0 & \text{if } x_i > 0. \end{cases}$$

Combining the three sub-subproblems, we see that the value of the i^{th} subproblem is exactly $\Psi(x/\lambda, m_{fi}, m_{bi}, \sigma_{fi}, \sigma_{bi})$. Combining the i subproblems we obtain Eq. (2.25).

□

A.1.3 Proof of Proposition 2.3

Proof. We need to show that under the null-hypothesis

$$\mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) \geq \overline{q}_i \text{ and } \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\mathbf{e}_i) \geq \underline{q}_i \text{ for all } i = 1, \ldots, d, \tag{A.1}$$

we have that $\mathbb{P}^*_{\mathcal{S}}(\hat{u}_i^{(s)} < \overline{q}_i \text{ or } -\hat{u}_i^{(N-s+1)} < \underline{q}_i \text{ for some } i) \leq \delta$.

We will prove the stronger result that, under assumption (A.1),

$$\mathbb{P}^*_{\mathcal{S}}(\hat{u}_i^{(s)} < \overline{q}_i) \leq \frac{\delta}{2d} \text{ and } \mathbb{P}^*_{\mathcal{S}}(-\hat{u}_i^{(N-s+1)} < \underline{q}_i) \leq \frac{\delta}{2d} \tag{A.2}$$

for any fixed i. The result then follows from the union bound.

Observe

$$\begin{aligned}
\mathbb{P}^*_{\mathcal{S}}\left(\overline{q}_i > \hat{u}_i^{(s)}\right) &\leq \mathbb{P}_{\mathcal{S}}\left(\mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) \geq \hat{u}_i^{(s)}\right) && \text{(by assumption (A.1))} \\
&= \sum_{j=s}^{N} \mathbb{P}^*_{\mathcal{S}}\left(\hat{u}_i^{(j)} \leq \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) < \hat{u}_i^{(j+1)}\right).
\end{aligned}$$

Each element of the sum is equal to the probability that a binomial random variable with N trials has $N - j$ successes, where the probability of a success is $\mathbb{P}^*(\tilde{u}_i \geq \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i))$. This last probability is at most ϵ/d. Thus,

$$\mathbb{P}_{\mathcal{S}}\left(\mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(x_i\mathbf{e}_i) > x_i\hat{u}_i^{(s)}\right) \leq \sum_{j=s}^{N} \binom{N}{N-j} (\epsilon/d)^{N-j}(1-\epsilon/d)^j \leq \frac{\delta}{2d},$$

where the last inequality follows from the definition of s. This proves Eq. (A.2) for $\hat{u}^{(s)}$. The proof for $\hat{u}^{(N-s+1)}$ is similar. □

A.1.4 Proof of Thm. 2.11

Proof. The proof follows directly from two applications of the Cauchy-Schwartz inequality. □

A.1.5 Proof of Thm 2.13

We will first require the following proposition:

Proposition A.1.

$$\sup_{\mathbb{P}\in\mathcal{P}_{DY}} \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > t) = \tag{A.3}$$

$$\begin{aligned}
\min_{r,s,\theta,\mathbf{y}_1,\mathbf{y}_2,\mathbf{Z}} \quad & r+s \\
s.t. \quad & \begin{pmatrix} r+\mathbf{y}_1^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^{-T}\hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}(\mathbf{q}-\mathbf{y}_1)^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_1) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& \begin{pmatrix} r+\mathbf{y}_2^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^{-T}\hat{\mathbf{u}}^{(N+1)} + \theta t - 1 & \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{x})^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{x}) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& s \geq (\gamma_2\hat{\mathbf{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)\circ\mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} + \sqrt{\gamma_1}\|\mathbf{q}+2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
& \mathbf{y}_1 = \mathbf{y}_1^+ + \mathbf{y}_1^-, \quad \mathbf{y}_2 = \mathbf{y}_2^+ + \mathbf{y}_2^-, \quad \mathbf{y}_1^+, \mathbf{y}_1^-, \mathbf{y}_2^+, \mathbf{y}_2^-\theta \geq \mathbf{0}.
\end{aligned}$$

Proof. In the spirit of linear programming duality, we claim that $\sup_{\mathbb{P}\in\mathcal{P}_{DY}} \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > t)$ has the following dual representation:

$$\begin{aligned}
\min_{r,s,\mathbf{q},\mathbf{Z},\mathbf{y}_1,\mathbf{y}_2,\theta} \quad & r+s \\
\text{s.t.} \quad & r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 0 \quad \forall \mathbf{u} \in [\hat{u}^{(0)}, \hat{u}^{(N+1)}], \\
& r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 1 \quad \forall \mathbf{u} \in [\hat{u}^{(0)}, \hat{u}^{(N+1)}] \cap \{\mathbf{u} : \mathbf{u}^T\mathbf{x} > t\}, \\
& s \geq (\gamma_2\hat{\mathbf{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)\circ\mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} +, \sqrt{\gamma_1}\|\mathbf{q}+2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
& \mathbf{Z} \succeq \mathbf{0}.
\end{aligned} \tag{A.4}$$

See [32] or the proof of Lemma 1 in [52] for details. The first two constraints are

robust constraints. Since $\mathbf{Z}$ is positive semidefinite, we use strong duality to write:

$$\begin{array}{lll}
\min\limits_{\mathbf{u}} & \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} & \max\limits_{\mathbf{y}_1,\mathbf{y}_1^+,\mathbf{y}_1^-} \quad -\frac{1}{4}(\mathbf{q}-\mathbf{y}_1)^T\mathbf{Z}^{-1}(\mathbf{q}-\mathbf{y}_1) + \mathbf{y}_1^+\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^-\hat{\mathbf{u}}^{(N+1)} \\
\text{s.t.} & \hat{\mathbf{u}}^{(0)} \le \mathbf{u} \le \hat{\mathbf{u}}^{(N+1)}, & \iff \quad \text{s.t.} \quad \mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-, \;\; \mathbf{y}_1^+, \mathbf{y}_1^- \ge \mathbf{0}. \\
\min\limits_{\mathbf{u}} & \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} & \max\limits_{\mathbf{y}_2,\mathbf{y}_2^+,\mathbf{y}_2^-} \quad -\frac{1}{4}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{x})^T\mathbf{Z}^{-1}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{x}) \\
 & & \qquad + \mathbf{y}_2^+\hat{u}^{(0)} - \mathbf{y}_2^-\hat{\mathbf{u}}^{(N+1)} + \theta t \\
\text{s.t.} & \hat{\mathbf{u}}^{(0)} \le \mathbf{u} \le \hat{\mathbf{u}}^{(N+1)}, & \iff \quad \text{s.t.} \quad \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-, \;\; \mathbf{y}_2^+, \mathbf{y}_2^- \ge \mathbf{0}, \; \theta \ge 0. \\
 & \mathbf{u}^T\mathbf{x} \ge t, &
\end{array}$$

Then, by using Schur-Complements, we can rewrite Problem (A.4) as in the proposition. □

We can now prove the theorem.

Proof of Thm. 2.13. Using Proposition A.1, we can characterize the worst-case quantile by

$$\sup_{\mathbb{P}\in\mathcal{P}_{DY}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{x}) = \inf\left\{t : r+s \le \epsilon, (r,s,t,\theta,\mathbf{y}_1,\mathbf{y}_2,\mathbf{Z}) \text{ are feasible in problem (A.3)}\right\}. \tag{A.5}$$

Notice the infimum on the right involves bilinear terms of the form θt and $\theta\mathbf{x}$. We rewrite this expression to eliminate the bilinear terms.

We first claim that $\theta > 0$ in any feasible solution to the infimum in Eq. (A.5). Suppose to the contrary that $\theta = 0$. Then this solution is also feasible when t is replaced by $t - \Delta t$. By taking $\Delta t \to \infty$, this shows that $\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{x} > -\infty) \le \epsilon$ for all $\mathbb{P} \in \mathcal{P}_{DY}$. Since $\mathrm{supp}(\mathbb{P})$ is bounded, this yields a contradiction.

Using $\theta > 0$, we can rescale all of the above optimization variables in problem (A.3)

by θ. Substituting this into Eq. (A.5) yields

$$
\begin{aligned}
\inf \quad & t \\
\text{s.t.} \quad & r + s \leq \theta\epsilon, \\
& \begin{pmatrix} r + \mathbf{y}_1^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^{-T}\hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}(\mathbf{q}-\mathbf{y}_1)^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_1) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& \begin{pmatrix} r + \mathbf{y}_2^{+T}\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^{-T}\hat{\mathbf{u}}^{(N+1)} + t - \theta & \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\mathbf{x})^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\mathbf{x}) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& s \geq (\gamma_2\hat{\mathbf{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} + \sqrt{\gamma_1}\|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
& \mathbf{y}_1 = \mathbf{y}_1^+ + \mathbf{y}_1^-, \quad \mathbf{y}_2 = \mathbf{y}_2^+ + \mathbf{y}_2^-, \quad \mathbf{y}_1^+, \mathbf{y}_1^-, \mathbf{y}_2^+, \mathbf{y}_2^-, \theta \geq \mathbf{0}.
\end{aligned}
$$

This optimization can be written as a semidefinite optimization problem which is positively homogenous and convex in $\mathbf{x}$. Its dual yields the explicit expression for $\mathcal{U}^{DY}$ in the theorem. This proves part (i) of the theorem. Part (ii) follows directly from our schema and Thm. 2.12. □

A.2 Generalizations of $\mathcal{U}^I$ and $\mathcal{U}^{FB}$

In this section we show how to extend our constructions for $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ to other EDF tests. We consider several of the most popular, univariate goodness-of-fit, empirical distribution function tests, namely:

Kuiper (K) Test The K test rejects the null hypothesis at level δ if

$$\max_{j=1,\ldots,N}\left(\frac{j}{N} - \mathbb{P}(\tilde{u} \leq \hat{u}^{(j)})\right) + \max_{j=1,\ldots,N}\left(\mathbb{P}(\tilde{u} < \hat{u}^{(j)}) - \frac{j-1}{N}\right) > V_{1-\delta}.$$

Cramer von-Mises (CvM) Test The CvM test rejects the null hypothesis at level δ if

$$\frac{1}{12N} + \sum_{j=1}^{N}\left(\frac{2j-1}{2N} - \mathbb{P}(\tilde{u} \leq \hat{u}^{(j)})\right)^2 > T_{1-\delta}.$$

Watson (W) Test The W test rejects the null hypothesis at level δ if

$$\frac{1}{12N} + \sum_{j=1}^{N} \left(\frac{2j-1}{2N} - \mathbb{P}(\tilde{u} \leq \hat{u}^{(j)}) \right)^2 - N \left(\frac{1}{N} \sum_{j=1}^{N} \mathbb{P}(\tilde{u} \leq \hat{u}^{(j)}) - \frac{1}{2} \right)^2 > U^2_{1-\delta}.$$

Anderson-Darling (AD) Test The AD test rejects the null hypothesis at level δ if

$$-N - \sum_{j=1}^{N} \frac{2j-1}{N} \left(\log \left(\mathbb{P}(\tilde{u} \leq \hat{u}^{(j)}) \right) + \log \left(1 - \mathbb{P}(\tilde{u} \leq \hat{u}^{(N+1-j)}) \right) \right) > A^2_{1-\delta}$$

Above, $K_{1-\delta}$, $V_{1-\delta}$, $T_{1-\delta}$, $U^2_{1-\delta}$ and $A^2_{1-\delta}$ are known constants, i.e., the $1-\delta$ quantiles of the appropriate null-distribution. Tables of such quantiles are readily available for various values of δ [e.g., 106, and references therein].

The confidence regions corresponding to these tests are, respectively,

$$\begin{aligned}
\overline{\mathcal{P}}^{K} =& \Big\{ \mathbb{P} \in \Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}] : \ \exists \Gamma_1, \Gamma_2 \text{ s.t. } \Gamma_1 + \Gamma_2 \leq \Gamma^K, \\
& \qquad \mathbb{P}(\tilde{u} \leq \hat{u}_j) \geq \frac{j}{N} - \Gamma_1, \ \ \mathbb{P}(\tilde{u} < \hat{u}_j) \leq \frac{j-1}{N} + \Gamma_2, \ \ j = 1, \ldots, N \Big\}, \\
\overline{\mathcal{P}}^{CvM} =& \left\{ \mathbb{P} \in \Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}] : \ \sum_{j=1}^{N} \left(\frac{2j-1}{N} - \mathbb{P}(\tilde{u} \leq \hat{u}_j) \right)^2 \leq \Gamma^{CvM} \right\}, \\
\overline{\mathcal{P}}^{W} =& \left\{ \mathbb{P} \in \Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}] : \ \sum_{j=1}^{N} \left(\frac{2j-1}{2N} - \mathbb{P}(\tilde{u} \leq \hat{u}_j) \right)^2 \right. \\
& \qquad \left. - N \left(\frac{1}{N} \sum_{j=1}^{N} \mathbb{P}(\tilde{u} \leq \hat{u}_j) - \frac{1}{2} \right)^2 \leq \Gamma^{W} \right\}, \\
\overline{\mathcal{P}}^{AD} =& \left\{ \mathbb{P} \in \Theta[\hat{u}^{(0)}, \hat{u}^{(N+1)}] : \ \sum_{j=1}^{N} \frac{2j-1}{N} \big(\log(\mathbb{P}(\tilde{u} \leq \hat{u}_j)) + \log(\mathbb{P}(\tilde{u} > \hat{u}_{N+1-j})) \big) \geq \Gamma^{AD} \right\},
\end{aligned}$$

where

$$\Gamma^K = V_{1-\delta}, \ \ \Gamma^{CvM} = T_{1-\delta} - \frac{1}{12N}, \ \ \Gamma^W = U^2_{1-\delta}, \ \ \Gamma^{AD} = -A^2_{1-\delta} - N,$$

The critical observation is that, as with $\overline{\mathcal{P}}^{KS}$, we compute worst-case expectations over each of these confidence regions by solving a finite dimensional, convex optimization problem. The precise optimization problems depends on the choice of

test. Specifically, for any $\theta \in [0,1]$, define

$$\mathcal{F}_\theta^{CvM} = \Bigg\{ (F_0, \ldots, F_{N+1})^T \in \mathbb{R}_+^{N+2} : \quad F_0 \le \cdots \le F_{N+1} = \theta, \quad \sum_{j=1}^{N} \left(\frac{2j-1}{2N} - F_j \right)^2 \le \theta \Gamma^{CvM} \Bigg\}, \tag{A.6}$$

$$\mathcal{F}_\theta^{W} = \Bigg\{ (F_0, \ldots, F_{N+1})^T \in \mathbb{R}_+^{N+2} : \quad F_0 \le \cdots \le F_{N+1} = \theta, \quad \sum_{j=1}^{N} \left(\frac{2j-1}{2N} - F_j \right)^2 - N \left(\frac{1}{N} \sum_{j=1}^{N} F_j - \frac{1}{2} \right)^2 \le \theta \Gamma^{W} \Bigg\}, \tag{A.7}$$

$$\mathcal{F}_\theta^{AD} = \Bigg\{ (F_0, \ldots, F_{N+1})^T \in \mathbb{R}_+^{N+2} : \quad F_0 \le \cdots \le F_{N+1} = \theta, \quad \sum_{j=1}^{N} \frac{2j-1}{N} \left[\log(F_j) + \log(1 - F_{N+1-j}) \right] \ge \theta \Gamma^{AD} \Bigg\}. \tag{A.8}$$

We then have the following theorem, paralleling Thm. 2.4.

Theorem A.1.

i) Suppose $g(u)$ is monotonic. Then,

$$\sup_{\mathbb{P} \in \overline{\mathcal{P}}^K} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})] = \max_{\theta \in [0,1]} \sum_{j=0}^{N+1} \left(\theta q_j^R(\Gamma^K) + (1-\theta) q_j^L(\Gamma^K) \right) g(\hat{u}^{(j)}). \tag{A.9}$$

ii) Suppose $g(u)$ is either non-decreasing or else non-increasing and right-continuous. Then,

$$\sup_{\mathbb{P} \in \overline{\mathcal{P}}^{CvM}} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})] = \max_{\mathbf{p} \in \mathcal{P}^{CvM}} \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}), \tag{A.10}$$

where

$$\mathcal{P}^{CvM} = \Big\{ \mathbf{p}^R + \mathbf{p}^L : \exists \theta \in [0,1], \ \mathbf{F}^R \in \mathcal{F}_\theta^{CvM}, \ \mathbf{F}^L \in \mathcal{F}_{(1-\theta)}^{CvM}, \quad F_j^R = \sum_{k=0}^{j} p_k, \ \ F_j^L = \sum_{k=0}^{j-1} p_k, \ \ j = 0, \ldots, N+1 \Big\}. \tag{A.11}$$

iii) Define $\mathcal{P}^W$ and $\mathcal{P}^{AD}$ as we have defined $\mathcal{P}^{CvM}$ but with $\mathcal{F}_\theta^{CvM}, \mathcal{F}_{(1-\theta)}^{CvM}$ replaced by $\mathcal{F}_\theta^{W}, \mathcal{F}_{(1-\theta)}^{W}$ and $\mathcal{F}_\theta^{AD}, \mathcal{F}_{(1-\theta)}^{AD}$ respectively. Then, Eq. (A.10) *remains true if*

$\overline{\mathcal{P}}^{CvM}, \mathcal{P}^{CvM}$ *is replaced by* $\overline{\mathcal{P}}^{W}, \mathcal{P}^{W}$ *or* $\overline{\mathcal{P}}^{AD}, \mathcal{P}^{AD}$.

iv) If $g(u)$ is non-decreasing (resp. non-increasing), then the optima of the right-hand side maximizations in parts i)-iii) of the Thm. are attained when $\theta = 1$ (resp. $\theta = 0$).

The spirit of the proof is very similar to that of Thm. 2.4. The only substantive difference is that the suprema above are achieved as the limit of a sequence of probability measures, and, consequently, we must make some limiting arguments. To this end, we first prove the following proposition. Let $\epsilon' = \min_{j=0,\dots,N} \hat{u}^{(j+1)} - \hat{u}^{(j)}$ be the minimum distance between two data points.

Proposition A.2. *Suppose $g(u)$ is non-increasing and right-continuous, and let $\overline{\mathcal{P}} \in \{\overline{\mathcal{P}}^{CvM}, \overline{\mathcal{P}}^{W}, \overline{\mathcal{P}}^{AD}\}$. Then $\sup_{\mathbb{P}\in\overline{\mathcal{P}}} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})]$ is achieved as the limit of a sequence of probability measures $\mathbb{Q}_n$, $(n = 1, 2, \dots,)$ such that*

$$supp(\mathbb{Q}_n) \subseteq \left((\hat{u}^{(0)}, \hat{u}^{(0)} + \frac{\epsilon'}{n}] \cup (\hat{u}^{(1)}, \hat{u}^{(1)} + \frac{\epsilon'}{n}] \cup \cdots \cup (\hat{u}^{(N)}, \hat{u}^{(N)} + \frac{\epsilon'}{n}] \right). \quad \text{(A.12)}$$

Proof. Let $(\mathbb{P}_n \in \mathcal{P},\, n \geq 1)$ be some sequence of probability measures whose limit is the supremum. Define $\mathbb{Q}_n(A) \equiv \mathbb{P}_n(A)$ for any

$$A \subseteq \left((\hat{u}^{(0)}, \hat{u}^{(0)} + \frac{\epsilon'}{n}] \cup (\hat{u}^{(1)}, \hat{u}^{(1)} + \frac{\epsilon'}{n}] \cup \cdots \cup (\hat{u}^{(N)}, \hat{u}^{(N)} + \frac{\epsilon'}{n}] \right).$$

Furthermore, let

$$\begin{aligned}
&\mathbb{Q}_n(\tilde{u} = \hat{u}^{(j)} + \frac{\epsilon'}{n}) \equiv \mathbb{P}_n(\hat{u}^{(j)} + \frac{\epsilon'}{n} \leq \tilde{u} < \hat{u}^{(j+1)}), \quad j = 0, \dots, N-1, \\
&\mathbb{Q}_n(\tilde{u} = \hat{u}^{(N)} + \frac{\epsilon'}{n}) \equiv \mathbb{P}_n(\hat{u}^{(N)} + \frac{\epsilon'}{n} \leq \tilde{u} \leq \hat{u}^{(N+1)}).
\end{aligned}$$

Intuitively, for each j, the measure $\mathbb{Q}_n$ collapses all the mass that $\mathbb{P}_n$ assigns to $[\hat{u}^{(j)} + \frac{\epsilon'}{n}, \hat{u}^{(j+1)})$ onto $\hat{u}^{(j)} + \frac{\epsilon'}{n}$, but leaves $\mathbb{P}_n$ otherwise untouched. (See Fig. A-1).

One can check that $\mathbb{Q}_n$ has the required support and $\mathbb{Q}_n \in \overline{\mathcal{P}}$. Moreover, since

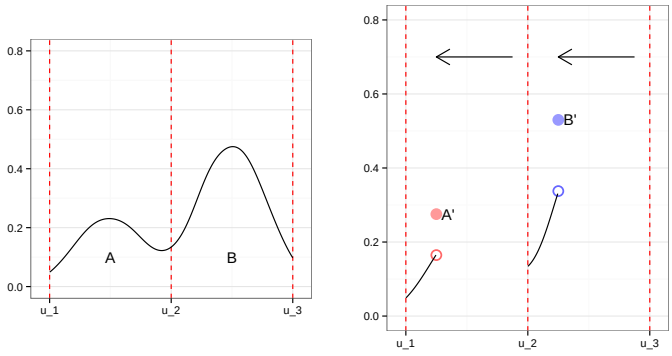


Figure A-1: The left panel shows the original distribution. Redistributing the probability mass of $\mathbb{P}$ on the open interval $(\hat{u}^{(j-1)}, \hat{u}^{(j)})$ does not affect its feasibility. Consequently, when $g(u)$ is non-increasing, it will accumulate on an interval arbitrarily close to the left hand endpoint of each interval (right panel).

$g(u)$ is non-increasing, $\mathbb{E}^{\mathbb{P}_n}[g(\tilde{u})] \leq \mathbb{E}^{\mathbb{Q}_n}[g(\tilde{u})]$. It follows that

$$\sup_{\mathbb{P}\in\overline{\mathcal{P}}} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})] = \lim_{n\to\infty} \mathbb{E}^{\mathbb{P}_n}[g(\tilde{u})] \leq \lim_{n\to\infty} \mathbb{E}^{\mathbb{Q}_n}[g(\tilde{u})] \leq \sup_{\mathbb{P}\in\overline{\mathcal{P}}} \mathbb{E}^{\mathbb{P}}[g(\tilde{u})],$$

whereby we have equality everywhere. □

We can now prove the theorem.

Proof of Thm. A.1. We first prove Eq. (A.9). From an identical argument to that in the proof of Thm. 2.4 we can show that the given supremum is equivalent to the following linear optimization problem and its dual:

$$\text{Primal:} \begin{cases} \max\limits_{\mathbf{p},\Gamma_1,\Gamma_2} & \sum\limits_{k=0}^{N+1} p_k g(\hat{u}^{(k)}) \\ \text{s.t.} & \mathbf{p} \geq \mathbf{0}, \quad \mathbf{e}^T\mathbf{p} = 1, \\ & \sum\limits_{k=0}^{j} p_k \geq \dfrac{j}{N} - \Gamma_1, \quad j = 1,\ldots,N, \\ & \sum\limits_{k=j}^{N+1} p_k \geq \dfrac{N-j+1}{N} - \Gamma_2, \quad j = 1,\ldots,N, \\ & \Gamma_1 + \Gamma_2 \leq \Gamma^K, \end{cases}$$

$$\text{Dual:} \begin{cases} \min\limits_{t,\mathbf{x},\mathbf{y},\theta} & t - \sum\limits_{j=1}^{N} \dfrac{j}{N} x_j - \sum\limits_{j=1}^{N} \dfrac{N-j+1}{N} y_j + \theta\Gamma_k, \\ \text{s.t.} & t - \sum\limits_{k \le j \le N} x_j - \sum\limits_{1 \le j \le k} y_j \ge g(\hat{u}^{(k)}), \quad k = 0, \ldots, N+1, \\ & \theta \ge \sum\limits_{j=1}^{N} x_j, \quad \theta \ge \sum\limits_{j=1}^{N} y_j, \\ & \mathbf{x}, \mathbf{y}, \theta \ge 0. \end{cases}$$

One can check that if $g(u)$ is non-decreasing, then the primal solution $\mathbf{q}^R(\Gamma^K)$ and dual solution $\mathbf{y} = \mathbf{0}, t = g(\hat{u}^{(N+1)}), \theta = \sum_{j=1}^{N} x_j$ and

$$x_j = \begin{cases} g(\hat{u}^{(j+1)}) - g(\hat{u}^{(j)}), & \text{for } N - j^* \le j \le N, \\ 0 & \text{otherwise,} \end{cases}$$

are an optimal pair. Similarly, if $g(u)$ is non-increasing, then the primal soution $\mathbf{q}^L(\Gamma^K)$ and dual solution $\mathbf{x} = \mathbf{0}, t = g(\hat{u}^{(0)}), \theta = \sum_{j=1}^{N} y_j$ and

$$y_j = \begin{cases} g(\hat{u}_{j-1}) - g(\hat{u}_j), & \text{for } 1 \le j \le j^* + 1, \\ 0 & \text{otherwise,} \end{cases}$$

are an optimal pair. The remainder of the proof is identical to the proof of Thm. 2.4.

We next prove Eq. (A.10). First observe that given any $\mathbf{p}^R, \mathbf{p}^L$ which is feasible in the right-hand side maximization, we can construct a sequence of probability measures $\mathbb{P}_n$ defined by

$$\mathbb{P}_n(\tilde{\mathbf{u}} = \hat{u}^{(j)}) \equiv p_j^R, \quad j = 0, \ldots, N+1,$$
$$\mathbb{P}_n(\tilde{\mathbf{u}} = \hat{u}^{(j)} + \frac{\epsilon'}{n}) \equiv p_j^L, \quad j = 0, \ldots, N+1.$$

By construction, $\mathbb{P}_n \in \overline{\mathcal{P}}^{CvM}$ and $\lim_{n\to\infty} \mathbb{E}^{\mathbb{P}_n}[g(\tilde{u})] = \sum_{j=0}^{N+1} (p_j^R + p_j^L) g(\hat{u}^{(j)})$. It follows that Eq. (A.10) holds with "=" replaced by "$\ge$".

For the reverse inequality, we have two cases. First suppose that $g(u)$ is non-

decreasing. Then for any $\mathbb{P} \in \overline{\mathcal{P}}^{CvM}$, define $\mathbb{Q}$ according to Eq. (2.17). Observe $\mathbb{Q} \in \overline{\mathcal{P}}^{CvM}$ and $\mathbb{E}^{\mathbb{P}}[g(\tilde{u})] \leq \mathbb{E}^{\mathbb{Q}}[g(\tilde{u})]$. It follows that the measure attaining the supremum in Eq. (A.10) has discrete support on $\hat{u}^{(1)}, \ldots, \hat{u}^{(N+1)}$. Thus, the supremum is equivalent to the following optimization problem:

$$\max_{\mathbf{p} \in \mathcal{P}^{CvMR}} \quad \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}),$$

where $\mathcal{P}^{CvMR} \equiv \{\mathbf{p} : \exists \mathbf{F} \in \mathcal{F}_1^{CvM},\ F_j = \sum_{k=0}^{j} p_k, \quad j = 0, \ldots, N+1\}$. Notice this optimization problem corresponds to the right-hand maximization in Eq. (A.10) when $\theta = 1$. This proves the reverse inequality holds for Eq. (A.10) when $g(u)$ is non-decreasing.

On the other hand, suppose $g(u)$ is non-increasing and right-continuous. From Proposition A.2, the limit of the sequence of measures attaining the supremum in Eq. (A.10) is also discrete and supported on points arbitrarily close, but to the right of $\hat{u}^{(0)}, \ldots, \hat{u}^{(N+1)}$. Since $g(u)$ is right-continuous, the supremum is equivalent to

$$\max_{\mathbf{p} \in \mathcal{P}^{CvML}} \quad \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}),$$

where $\mathcal{P}^{CvML} \equiv \{\mathbf{p} : \exists \mathbf{F} \in \mathcal{F}_1^{CvM},\ F_j = \sum_{k=0}^{j-1} p_k, \quad j = 0, \ldots, N+1\}$. Notice this optimization problem corresponds to the right-hand maximization in Eq. (A.10) when $\theta = 0$. This implies the reverse inequality holds for Eq. (A.10) when $g(u)$ is non-increasing and right continuous. This completes the proof of the second statement in the theorem.

The third statement of the theorem is proven identically to the second statement. Finally, the last the statement is a direct consequence of the constructions above. □

Equipped with Thm. A.1, we can tractably evaluate worst-case expectations of continuous, monotonic functions over each of our confidence regions. Specifically, for

K: We can evaluate worst-case expectations in closed-form just as for the KS test.

CvM or W: We evaluate worst-case expectations by solving a second order cone

problem. This follows from the form of $\mathcal{F}_\theta^{CvM}$ and $\mathcal{F}_\theta^{W}$. The former is clearly second order cone representable. The quadratic constraint in the latter can be written as $\mathbf{F}^T(\mathbf{I} - \frac{1}{N}\mathbf{e}\mathbf{e}^T)\mathbf{F} + (\mathbf{e} - 2\mathbf{c})^T\mathbf{F} \leq \theta\Gamma^W + \frac{N}{4} - \mathbf{c}^T\mathbf{c}$, where $\mathbf{I}$ is the identity matrix and $c_j = \frac{2j-1}{2N}$. The matrix $\mathbf{I} - \frac{1}{N}\mathbf{e}\mathbf{e}^T$ is diagonally dominant, and, therefore, positive semidefinite. Thus, the resulting constraint is convex and representable using the second order cone.

AD: We can evaluate worst-case expectations by solving a simple convex optimization problem. The problem can also be reformulated as an exponential cone optimization problem or a geometric program and can be solved with a variety of approaches. See online supplement A.4.1.

The constructions of the analogues to $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ is nearly complete. Leveraging the above characterization, the reader can check that the constructions and proofs for $\mathcal{U}^I$ and $\mathcal{U}^{FB}$ presented in the text remain valid if we replace $\overline{\mathcal{P}}^{KS}$ with one of $\overline{\mathcal{P}}^{K}, \overline{\mathcal{P}}^{CvM}, \overline{\mathcal{P}}^{W}$, or $\overline{\mathcal{P}}^{AD}$ and $\mathcal{P}^{KS}$ with one of $\mathcal{P}^{K}, \mathcal{P}^{CvM}, \mathcal{P}^{W}, \mathcal{P}^{AD}$ respectively. This observation proves the following theorems which generalize Thms. 2.4 and Thm. 2.6

Theorem A.2. *Suppose that $\mathbb{P}^*$ has independent components and supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Let $\mathcal{P}_i$ be the finite dimensional analogue of a $\sqrt[d]{1-\delta}$ confidence region for the i^{th} marginal distribution for any of the above EDF tests.*

i) With probability $1-\delta$ over the sample $\mathcal{S}$, the set

$$\mathcal{U}^I = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \ \exists \mathbf{q}_i \in \Delta_{N+2}, \quad \hat{\mathbf{u}}_i^T \mathbf{q}_i = u_i, \quad \mathbf{p}_i \in \mathcal{P}_i, \quad i = 1, \ldots, d, \quad \sum_{i=1}^{d} D(\mathbf{q}_i, \mathbf{p}_i) \leq \log(1/\epsilon) \Bigg\} \tag{A.13}$$

implies a probabilistic guarantee at level ϵ for $\mathbb{P}^*$.

ii) Define

$$
\begin{aligned}
m_{fi} &= \max_{\mathbf{p}^i \in \mathcal{P}_i} \sum_{j=0}^{N+1} p_j^i \hat{u}_{ij}, \\
\sigma_{fi}^2 &= \sup_{x>0} -\frac{2m_{fi}}{x} + \frac{2}{x^2} \log\left(\max_{\mathbf{p} \in \mathcal{P}_i} \sum_{j=0}^{N+1} p_j^i e^{x_i \hat{u}_{ij}} \right), \\
m_{bi} &= \min_{\mathbf{p}^i \in \mathcal{P}_i} \sum_{j=0}^{N+1} p_j^i \hat{u}_{ij}, \\
\sigma_{bi}^2 &= \sup_{x>0} \frac{2m_{bi}}{x} + \frac{2}{x^2} \log\left(\max_{\mathbf{p} \in \mathcal{P}_i} \sum_{j=0}^{N+1} p_j^i e^{-x_i \hat{u}_{ij}} \right).
\end{aligned}
$$

The set

$$
\begin{aligned}
\mathcal{U}^{FB} = \Bigg\{ \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_3 : \mathbf{y}_2, \mathbf{y}_2 \in \mathbb{R}_+^d, \quad & \sum_{i=1}^{d} \frac{y_{2i}^2}{2\sigma_{fi}^2} + \frac{y_{3i}^2}{2\sigma_{bi}^2} \leq \log(1/\epsilon), \\
& m_{bi} \leq y_{1i} \leq m_{fi}, \quad i = 1, \ldots, d, \Bigg\}
\end{aligned}
$$

implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ.

A.3 Uncertainty Sets for Independent, Identically Distributed Marginals

In this section, we consider the case where $\mathbb{P}^*$ has i.i.d. components. To simplify notation, we will assume that $\mathcal{S} = \{\hat{u}_1, \ldots, \hat{u}_N\}$ is a sample from the common marginal distribution of the components $\mathbb{P}^*$ and that $\mathcal{S}$ is ordered so that $\hat{u}_j = \hat{u}^{(j)}$ for all j.

Our approach is based on the 2-sample Kolmogorov-Smirnov goodness-of-fit test. Let $\mathcal{S}_1 = \{\hat{u}_1, \ldots, \hat{u}_{N_1}\}$ and $\mathcal{S}_2 = \{u_1, \ldots, u_{N_2}\}$ be two i.i.d. samples. The 2-sample KS test compares the hypotheses:

$$
\begin{aligned}
H_0 &: \ \mathcal{S}_1 \text{ and } \mathcal{S}_2 \text{ were drawn from the same distribution,} \\
H_A &: \ \mathcal{S}_1 \text{ and } \mathcal{S}_2 \text{ were drawn from different distributions,}
\end{aligned}
$$

but does not specify the form of the distribution under H_0. It rejects the null hypothesis if

$$\max_{t\in\mathbb{R}} \left|\hat{\mathbb{P}}_{\mathcal{S}_1}(\tilde{u} \leq t) - \hat{\mathbb{P}}_{\mathcal{S}_2}(\tilde{u} \leq t)\right| > K_{1-\delta}\sqrt{\frac{N_1+N_2}{N_1N_2}},$$

where $\hat{\mathbb{P}}_{\mathcal{S}_1}, \hat{\mathbb{P}}_{\mathcal{S}_2}$ are the empirical measures for each sample.

Let $\mathcal{S}_1 = \mathcal{S}$ and let $N_2 = d$. The ϵ-confidence region of the sample $\mathcal{S}_2$ is the set of samples that would pass the 2-sample KS test at level ϵ, i.e.,

$$\mathcal{U}^{2KS} = \left\{u_1, \ldots, u_d \in \mathbb{R} : \left|\hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \leq t) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \leq t)\right| \leq \Gamma^{2KS} \quad \forall t \in \mathbb{R}\right\}, \quad \text{(A.14)}$$

where $\Gamma^{2KS} = K_{1-\delta}\sqrt{\frac{N+d}{Nd}}$. Notice that $\mathcal{U}^{2KS}$ is a set of realizations, not a set of probability measures. Define

$$\phi_{2KS}(\mathbf{x}, \mathcal{S}) \equiv \max_{\mathbf{u}\in\mathcal{U}^{2KS}} \mathbf{x}^T\mathbf{u} = \max_{\mathbf{u}\in conv(\mathcal{U}^{2KS})} \mathbf{x}^T\mathbf{u}.$$

We claim that ϕ_{2KS} is an upper bound to the $\text{VaR}_\epsilon^{\mathbb{P}^*}$ almost surely. Indeed, suppose this were not the case. Then, $\text{VaR}_\epsilon^{\mathbb{P}^*}(\mathbf{x}) > \phi_{2KS}(\mathbf{x}, \mathcal{S})$ implies that

$$\mathbb{P}^*(\mathbf{x}^T\tilde{\mathbf{u}} \leq \phi_{2KS}(\mathbf{x}, \mathcal{S})) < 1 - \epsilon. \quad \text{(A.15)}$$

On the other hand, the vector $\tilde{\mathbf{u}}$ represents an i.i.d. draw of d samples from $\mathbb{P}^*$. By construction of the test, then,

$$1 - \epsilon \leq \mathbb{P}(\tilde{\mathbf{u}} \in \mathcal{U}^{2KS}) \leq \mathbb{P}(\mathbf{x}^T\tilde{\mathbf{u}} \leq \phi_{2KS}(\mathbf{x})),$$

which contradicts Eq. (A.15). This proves the claim.

By our schema, the set $\mathcal{U}^{2KS}$ implies a probabilistic guarantee at level ϵ. Unfortunately, this set is awkwardly defined in terms of the empirical measure. In the remainder of the section, we show that we can rewrite $\mathcal{U}^{2KS}$ in the simpler form given in the main text.

Recall that $u^{(j)}$ is the j^{th} largest value of the sample $\{u_1, \ldots, u_d\}$.

Proposition A.3. *The set $\mathcal{U}^{2KS}$ can be rewritten as:*

$$\begin{aligned}\mathcal{U}^{2KS} = \Big\{\mathbf{u} \in \mathbb{R}^d \ : \hat{u}^{(0)} \le u^{(i)} \le \hat{u}^{(N+1)}, i = 1 \ldots, d, \\ u^{(\lceil \frac{dk}{N} - d\Gamma^{2KS} \rceil)} \le \hat{u}_k, \quad k = \lfloor N\Gamma^{2KS} \rfloor + 1, \ldots, N, \\ u^{(\lfloor \frac{d(k-1)}{N} + d\Gamma^{2KS} \rfloor)} \ge \hat{u}_k, \quad k = 1, \ldots, \lceil N(1-\Gamma^{2KS}) \rceil - 1\Big\}.\end{aligned} \tag{A.16}$$

Proof. We first show that $\mathcal{U}^{2KS}$ can be rewritten as

$$\begin{aligned}\Big\{\mathbf{u} \in [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] \ : \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le \hat{u}_k) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le \hat{u}_k) \le \Gamma^{2KS}, \quad k = \lfloor N\Gamma^{2KS} \rfloor + 1, \ldots, N, \\ \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} < \hat{u}_k) - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} < \hat{u}_k) \le \Gamma^{2KS}, \quad k = 1, \ldots, \lceil N(1-\Gamma^{2KS}) \rceil - 1\Big\}.\end{aligned} \tag{A.17}$$

Indeed, the first set of constraints in Eq. (A.17) are a subset of the constraints in Eq. (A.14) where $t = \hat{u}_k$. The second set of constraints are also implied by Eq. (A.14) since

$$\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} < \hat{u}_k) - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} < \hat{u}_k) = \lim_{\epsilon' \to 0} \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le \hat{u}_k - \epsilon') - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le \hat{u}_k - \epsilon').$$

Thus, the set defined by Eq. (A.14) is a subset of the set defined by Eq. (A.17).

To show the reverse inclusion, suppose $\mathbf{u}$ is not an element of (A.14). We have two cases:

Case 1: $\hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le t) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le t) > \Gamma^{2KS}$ for some t. Since $\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le t) \ge 0$, it must be that $t \ge \hat{u}_{\lfloor N\Gamma^{2KS} \rfloor + 1}$. Let k be such that $t \in [\hat{u}_k, \hat{u}_{k+1})$ with $k = \lfloor N\Gamma^{2KS} \rfloor + 1, \ldots, N$. Then,

$$\Gamma^{2KS} < \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le t) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le t) \le \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le \hat{u}_k) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le \hat{u}_k),$$

since $\hat{\mathbb{P}}_{\mathcal{S}}$ is constant on the interval $[\hat{u}_k, \hat{u}_{k+1})$ and $\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}$ is non-decreasing. This shows that $\mathbf{u}$ is not an element of the set defined by Eq. (A.17).

Case 2: $\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le t) - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \le t) > \Gamma^{2KS}$ for some t. Since $\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \le t) \le 1$, it must be that $t \le \hat{u}_{\lceil N(1-\Gamma^{2KS}) \rceil - 1}$. Let k be such that $t \in [\hat{u}_{k-1}, \hat{u}_k)$ with

$k = 1, \ldots, \lceil N(1 - \Gamma^{2KS}) \rceil - 1$. Then, for any ϵ' sufficiently small,

$$\Gamma^{2KS} > \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \leq t) - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \leq t) \geq \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \leq \hat{u}_k - \epsilon') - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \leq \hat{u}_k - \epsilon'),$$

since $\hat{\mathbb{P}}_{\mathcal{S}}$ is constant on the interval $[\hat{u}_k, \hat{u}_{k+1})$ and $\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}$ is non-decreasing. Taking the limit as $\epsilon' \to 0$ shows that $\mathbf{u}$ is not an element of set Eq. (A.17). This proves that the sets defined by Eq. (A.14) and Eq. (A.17) are equal.

Now, consider the first set of constraints in Eq. (A.17). For any k, we have that

$$\begin{aligned}
\hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} \leq \hat{u}_k) - \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \leq \hat{u}_k) \leq \Gamma^{2KS} &\iff \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} \leq \hat{u}_k) \geq \frac{k}{N} - \Gamma^{2KS} \\
&\iff \sum_{i=1}^{d} \mathbb{I}(u_i \leq \hat{u}_k) \geq \left\lceil \frac{kd}{N} - d\Gamma^{2KS} \right\rceil \\
&\iff u^{(\lceil \frac{kd}{N} - \Gamma d \rceil)} \leq \hat{u}_k.
\end{aligned}$$

This is the first set of constraints in Eq. (A.16). Similarly, from the second set of constraints in Eq. (A.17), we have for any k that

$$\begin{aligned}
\hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} < \hat{u}_k) - \hat{\mathbb{P}}_{\mathcal{S}}(\tilde{u} < \hat{u}_k) \leq \Gamma^{2KS} &\iff \hat{\mathbb{P}}_{\{u_1,\ldots,u_d\}}(\tilde{u} < \hat{u}_k) \leq \frac{k-1}{N} + \Gamma^{2KS} \\
&\iff \sum_{i=1}^{d} \mathbb{I}(u_i < \hat{u}_k) \leq \left\lfloor \frac{d(k-1)}{N} + d\Gamma^{2KS} \right\rfloor \\
&\iff u^{(\lfloor \frac{d(k-1)}{N} + d\Gamma^{2KS} \rfloor)} \geq \hat{u}_k.
\end{aligned}$$

This completes the proof. □

Observe that the proof relies critically on the structure of the 2-sample KS test. Indeed, all of the EDF tests we have considered (K, CvM, W, AD) admit 2 sample variants, but it is not clear how to extend the above proof to these cases.

A straightforward manipulation of the indices of Eq. (A.16) yields the formulation Eq. (2.28) given in the main text. Finally, we are in a position to prove Theorem 2.7.

Proof of Theorem 2.7. We have already shown that ϕ_{2KS} is a valid upper-bound for $\mathrm{VaR}_{\epsilon}^{\mathbb{P}^*}$. This proves the first part of the theorem. For the second part, given $\mathbf{x} \in \mathbb{R}^d$,

let σ be the permutation that orders the components of $\mathbf{x}$, i.e., $x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(d)}$. Then the $\mathbf{u}^*$ which achieves the maximum in $\max_{\mathbf{u}\in\mathcal{U}^{2KS}} \mathbf{u}^T\mathbf{x}$ must have the same ordering. It follows that

$$\begin{aligned}\phi_{2KS}(\mathbf{x}) = \max_{\mathbf{u}} \quad & \mathbf{u}^T\mathbf{x} \\ \text{s.t} \quad & \hat{u}_{\underline{k}(i)} \leq u_{\sigma(i)} \leq \hat{u}_{\overline{k}(i)}, \quad i = 1, \ldots, d, \\ & u_{\sigma(i)} \leq u_{\sigma(i+1)} \quad i = 1, \ldots, d-1.\end{aligned} \tag{A.18}$$

Neglecting the ordering constraints, we have the solution

$$u_{\sigma(i)} = \begin{cases} \hat{u}_{\overline{k}(i)} \text{ if } x_{\sigma(i)} \geq 0, \\ \hat{u}_{\underline{k}(i)} \text{ if } x_{\sigma(i)} < 0, \end{cases} \tag{A.19}$$

for $i = 1, \ldots, d$. This solution, however, also satisfies the ordering constraints since $\overline{k}(i)$ and $\underline{k}(i)$ are non-decreasing in i. □

A.4 Specialized Algorithms

A.4.1 Optimizing over $\mathcal{P}^{AD}$

In this section, we address evaluating $\max_{\mathbf{p}\in\mathcal{P}^{AD}} \sum_{j=0}^{N+1} g(\hat{u}^{(j)})p_j$ when $g(u)$ is either non-decreasing or else non-increasing and right-continuous. We first consider the non-decreasing case. From Thm. A.1, we know there exists a solution to this optimization problem in which $\theta = 1$. Consequently, by eliminating the variables $\mathbf{p}^R, \mathbf{p}^L$, we can

rewrite this optimization problem as

$$\begin{aligned}
\max_{\mathbf{F},\overline{\mathbf{F}}} \quad & \mathbf{c}^T\mathbf{F} \\
\text{s.t.} \quad & \mathbf{F} \geq \mathbf{0},\ F_{N+1} = 1, \\
& F_j \leq F_{j+1}, \quad j = 0,\ldots,N, \\
& \overline{F}_j = 1 - F_{N+1_j}, \quad j = 1,\ldots,N, \\
& \sum_{j=1}^{N} \frac{2j-1}{N}[\log(F_j) + \log(\overline{F}_j)] \geq \Gamma^{AD},
\end{aligned} \tag{A.20}$$

where $c_0 = -g(\hat{u}^{(1)})$, $c_j = g(\hat{u}^{(j)} - g(\hat{u}^{(j+1)})$ for $j = 1,\ldots,N$, and $c_{N+1} = g(\hat{u}^{(N+1)})$. By introducing auxiliary variables $\overline{F}_j$, we can rewrite this optimization as a geometric program:

$$\begin{aligned}
\sum_{j:c_j<0} c_j + \max_{\mathbf{F},\overline{\mathbf{F}}} \quad & \sum_{j:c_j\geq 0} c_j F_j + \sum_{j:c_j<0} |c_j|\overline{F}_j \\
\text{s.t.} \quad & \mathbf{F} \geq \mathbf{0},\ F_{N+1} = 1, \\
& F_j F_{j+1}^{-1} \leq 1, \quad j = 0,\ldots,N, \\
& \overline{F}_j + F_{N+1-j} \leq 1, \quad j = 1,\ldots,N, \\
& \exp(\Gamma^{AD}) \prod_{j=1}^{N} F_j^{-\frac{2j-1}{N}} \overline{F}_j^{-\frac{2j-1}{N}} \leq 1.
\end{aligned}$$

This formulation has the benefit of being in a standard form recognizable by existing geometric programming solvers. Our computational experiments, however, suggest that it is in fact more efficient to solve the dual of Eq. (A.20) directly. The

Lagrangian dual optimization problem is

$$\begin{aligned}
\min_{\boldsymbol{\lambda},\mathbf{y},t} \quad & \sum_{j=1}^{N} \frac{t(2j-1)}{N}\left(-2-\log\left(\frac{N(y_{N+1-j}-\lambda_{j+1}-\lambda_j-c_j)}{t(2j-1)}\right)-\log\left(\frac{Ny_j}{t(2j-1)}\right)\right) \\
\text{s.t.} \quad & \lambda_{N+1}-y_0+c_{N+1}=0, \quad \lambda_0-\lambda_1+c_0=0, \\
& y_{N+1-j}-\lambda_{j+1}-\lambda_j-c_j\geq 0, \quad j=1,\ldots,N, \\
& y_j\geq 0, \quad j=1,\ldots,N, \\
& t,\boldsymbol{\lambda}\geq 0.
\end{aligned}$$

This is a minimization of a smooth, convex function over linear constraints. There are a number of algorithms for solving this problem, including barrier methods and optimal first-order methods. In our experiments, we use the software IpOpt [111] to solve this problem. Computational experience suggests this formulation is superior to the geometric programming formulation for large N.

The case of non-increasing g is very similar to the above. We omit it.

Appendix B

Supplement to Chapt. 4

B.1 Omitted Proofs

B.1.1 Proof of Theorem 4.4

Proof. Let $\mathbf{f}^* = (f_1^*, \ldots, f_n^*)$ be any solution. We will construct a new solution with potentially lower cost with the required representation. We do this iteratively beginning with f_1^*.

Consider the subspace $\mathcal{T} \subset \mathcal{H}_1$ defined by $\mathcal{T} = \text{span}(k_{\mathbf{x}_1}, \ldots, k_{\mathbf{x}_N})$, and let $\mathcal{T}^\perp$ be its orthogonal complement. It follows that f_1^* decomposes uniquely into $f_1^* = f_0 + f_0^\perp$ with $f_0 \in \mathcal{T}$ and $f_0^\perp \in \mathcal{T}^\perp$. Consequently,

$$\begin{aligned} f_1^*(\mathbf{x}_j) &= \langle k_{\mathbf{x}_j}, f_1^* \rangle, && \text{(by (4.20))} \\ &= \langle k_{\mathbf{x}_j}, f_0 \rangle + \langle k_{\mathbf{x}_j}, f_0^\perp \rangle \\ &= \langle k_{\mathbf{x}_j}, f_0 \rangle && \text{(since } f_0^\perp \in \mathcal{T}^\perp\text{)} \\ &= f_0(\mathbf{x}_j) && \text{(by (4.20)).} \end{aligned}$$

Thus, the solution $\mathbf{f} = (f_0, f_2^*, \ldots, f_n^*)$ is feasible to (4.22). Furthermore, by orthogonality $\|f_1^*\|_{\mathcal{H}_1} = \|f_0\|_{\mathcal{H}_1} + \|f_0^\perp\|_{\mathcal{H}_1} \geq \|f_0\|_{\mathcal{H}_1}$. Since the objective is non-decreasing in $\|f_1\|_{\mathcal{H}}$, $\mathbf{f}$ has an objective value which is no worse than $\mathbf{f}^*$. We can now proceed iteratively, considering each coordinate in turn. After at most n steps, we have

constructed a solution with the required representation. □ □

B.1.2 Proof of Theorem 4.5

Proof. Suppose Problem (4.24) is feasible and let $\boldsymbol{\alpha}$ be a feasible solution. Define $\mathbf{f}$ via eq. (4.23). It is straightforward to check that $\mathbf{f}$ is feasible in Problem (4.22) with the same objective value.

On the other hand, let $\mathbf{f}$ be some feasible solution to Problem (4.22). By Theorem 4.4, there exists $\boldsymbol{\alpha}$ such that $f_i(\mathbf{x}_j) = \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j$, and $\|f_i\|_{\mathcal{H}}^2 = \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha}^T \mathbf{e}_i$. It straightforward to check that such $\boldsymbol{\alpha}$ is feasible in Problem (4.24) and that they yield the same objective value. Thus, Problem (4.22) is feasible if and only if Problem (4.24) is feasible, and we can construct an optimal solution to Problem (4.22) from an optimal solution to Problem (4.24) via (4.23). □ □

B.1.3 Proof of Theorem 4.6

Proof. As mentioned in the text, the key idea in the proof is to relate (4.12) with a randomized uncertain convex program. To this end, notice that if $z_N, \boldsymbol{\theta}_N$ are an optimal solution to (4.12) with the ℓ_∞-norm, then $(z_N, \boldsymbol{\theta}_N) \in \bigcap_{j=1}^N \mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ where

$$\mathcal{X}(\mathbf{x}, \mathbf{A}, \mathbf{b}, C) = \left\{ z, \boldsymbol{\theta} \in \Theta : \exists \mathbf{y} \in \mathbb{R}^m \text{ s.t. } \mathbf{A}^T \mathbf{y} \leq \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \;\; \mathbf{x}^T \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{b}^T \mathbf{y} \leq z \right\}.$$

The sets $\mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ are convex. Consider then the problem

$$\min_{z \geq 0, \boldsymbol{\theta}} \quad z \;\; \text{s.t.} \quad (z, \boldsymbol{\theta}) \in \bigcap_{j=1}^N \mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j).$$

This is exactly of the form Eq. 2.1 in [44]. Applying Theorem 2.4 of that work shows that with probability $\beta(\alpha)$ with respect to the sampling, the "violation probability" of the pair (z_N, θ_N) is a most α. In our context, the probability of violation is exactly the probability that $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ is not a z_N approximate equilibria. This proves the

theorem. □

Observe that the original proof in [44] requires that the solution $\boldsymbol{\theta}_N$ be unique almost surely. However, as mentioned on pg. 7 discussion point 5 of that text, it suffices to pick a tie-breaking rule for the $\boldsymbol{\theta}_N$ in the case of multiple solutions. The tie-breaking rule discussed in the main text is one possible example.

B.1.4 Proof of Theorem 4.7

We require auxiliary results. Our treatment closely follows [11]. Let $\zeta_1, \ldots, \zeta_N$ be i.i.d. For any class of functions $\mathcal{S}$, define the empirical Rademacher complexity $\mathcal{R}_N(\mathcal{S})$ by

$$\mathcal{R}_N(\mathcal{S}) = \mathbb{E}\left[\sup_{f\in\mathcal{S}} \frac{2}{N}\left|\sum_{i=1}^{N} \sigma_i f(\zeta_i)\right| \middle| \zeta_1, \ldots, \zeta_N\right],$$

where σ_i are independent uniform $\{\pm 1\}$-valued random variables. Notice this quantity is random, because it depends on the data $\zeta_1, \ldots, \zeta_N$.

Our interest in Rademacher complexity stems from the following lemma.

Lemma B.1. *Let $\mathcal{S}$ be a class of functions whose range is contained in $[0, M]$. Then, for any N, and any $0 < \beta < 1$, with probability at least $1 - \beta$ with respect to $\mathbb{P}$, every $f \in \mathcal{F}$ simultaneously satisfies*

$$\mathbb{E}[f(\zeta)] \leq \frac{1}{N}\sum_{i=1}^{N} f(\zeta_i) + \mathcal{R}_N(\mathcal{S}) + \sqrt{\frac{8M\log(2/\beta)}{N}} \tag{B.1}$$

Proof. The result follows by specializing Theorem 8 of [11]. Namely, using the notation of that work, let $\phi(y, a) = \mathcal{L}(y, a) = a/M$, $\delta = \beta$ and then apply the theorem. Multiply the resulting inequality by M and use Theorem 12, part 3 of the same work to conclude that $M\mathcal{R}_N(M^{-1}\mathcal{S}) = \mathcal{R}_N(\mathcal{S})$ to finish the proof. □

Remark B.1. The constants in the above lemma are not tight. Indeed, modifying the proof of Theorem 8 in [11] to exclude the centering of ϕ to $\tilde{\phi}$, one can reduce the constant 8 in the above bound to 2. For simplicity in what follows, we will not be concerned with improvements at constant order.

Remark B.2. Lemma B.1 relates the empirical expectation of a function to its true expectation. If $f \in \mathcal{S}$ were fixed a priori, stronger statements can be proven more simply by invoking the weak law of large numbers. The importance of Lemma B.1 is that it asserts the inequality holds uniformly for all $f \in \mathcal{S}$. This is important since in what follows, we will be identifying the relevant function f by an optimization, and hence it will not be known to us a priori, but will instead depend on the data.

Our goal is to use Lemma B.1 to bound the $\mathbb{E}[\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})]$. To do so, we must compute an upper-bound on the Rademacher complexity of a suitable class of functions. As a preliminary step,

Lemma B.2. *For any* $\mathbf{f}$ *which is feasible in* (4.12) *or* (4.28)*, we have*

$$\tilde{\epsilon}(\mathbf{f}, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}) \leq \overline{B} \quad \text{a.s.} \tag{B.2}$$

Proof. Using strong duality as in Theorem 4.2,

$$\tilde{\epsilon}(\mathbf{f}, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}) = \max_{\mathbf{x} \in \tilde{\mathcal{F}}} (\tilde{\mathbf{x}} - \mathbf{x})^T \mathbf{f}(\tilde{\mathbf{x}}) \leq 2R \sup_{\tilde{\mathbf{x}} \in \tilde{\mathcal{F}}} \|\mathbf{f}(\tilde{\mathbf{x}})\|_2, \tag{B.3}$$

by **A6**. For Problem (4.12), the result follows from the definition of $\overline{B}$. For Problem (4.28), observe that for any $\tilde{\mathbf{x}} \in \tilde{\mathcal{F}}$,

$$|f_i(\tilde{\mathbf{x}})|^2 = \langle f_i, k_{\tilde{\mathbf{x}}} \rangle^2 \leq \|f_i\|_{\mathcal{H}}^2 \sup_{\|\mathbf{x}\|_2 \leq R} k(\mathbf{x}, \mathbf{x}) = \|f_i\|_{\mathcal{H}}^2 \overline{K}^2 \leq \kappa_i^2 \overline{K}^2, \tag{B.4}$$

where the middle inequality follows from Cauchy-Schwartz. Plugging this into Eq. (B.3) and using the definition of $\overline{B}$ yields the result. □

Now consider the class of functions

$$F = \begin{cases} \left\{ (\mathbf{x}, \mathbf{A}, \mathbf{b}, C) \mapsto \epsilon(\mathbf{f}, \mathbf{x}, \mathbf{A}, \mathbf{b}, C) : \mathbf{f} = \mathbf{f}(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \right\} & \text{for Problem (4.12)} \\ \left\{ (\mathbf{x}, \mathbf{A}, \mathbf{b}, C) \mapsto \epsilon(\mathbf{f}, \mathbf{x}, \mathbf{A}, \mathbf{b}, C) : f_i \in \mathcal{H}, \ \ \|f_i\|_{\mathcal{H}} \leq \kappa_i \ \ i = 1, \ldots, N \right\} & \text{for Problem (4.28).} \end{cases}$$

Lemma B.3.

$$\mathcal{R}_N(F) \leq \frac{2\overline{B}}{\sqrt{N}}$$

Proof. We prove the lemma for Problem (4.12). The proof in the other case is identical. Let $\mathcal{S} = \{\mathbf{f}(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Then,

$$\begin{aligned}
\mathcal{R}_N(F) &= \frac{2}{N}\mathbb{E}\left[\sup_{f\in\mathcal{S}}\left|\sum_{j=1}^{N}\sigma_j\epsilon(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)\right|\right| (\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)_{j=1}^{N}\right] \\
&\leq \frac{2\overline{B}}{N}\mathbb{E}[(\sum_{j=1}^{N}\sigma_j^2)^{\frac{1}{2}}] && \text{(using (B.2))} \\
&\leq \frac{2\overline{B}}{N}\sqrt{\mathbb{E}[\sum_{j=1}^{N}\sigma_j^2]} && \text{(Jensen's inequality)} \\
&= \frac{2\overline{B}}{\sqrt{N}} && (\sigma_j^2 = 1 \text{ a.s.}).
\end{aligned}$$

□

We are now in a position to prove the theorem.

Theorem 4.7 . Observe that $z_N = \frac{1}{N}\sum_{j=1}^{N}(\epsilon(\mathbf{f}_N, \mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j))^p$. Next, the function $\phi(z) = z^p$ satisfies $\phi(0) = 0$ and is Lipschitz with constant $L_\phi = p\overline{B}^{p-1}$ on the interval $[0, \overline{B}]$. Consequently, from Theorem 12 part 4 of [11],

$$\begin{aligned}
\mathcal{R}_N(\phi \circ F) &\leq 2L_\phi\mathcal{R}_N(F) \\
&\leq 2p\overline{B}^{p-1}\frac{2\overline{B}}{\sqrt{N}} \\
&= \frac{4p\overline{B}^p}{\sqrt{N}}.
\end{aligned}$$

Now applying Lemma B.1 with $\zeta \to (\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$, $f(\cdot) \to \epsilon(\cdot)^p$, and $M = \overline{B}^p$ yields the first part of the theorem.

For the second part of the theorem, observe that, conditional on the sample, the event $\tilde{\mathbf{x}}$ is not a $z_N + \alpha$-approximate equilibrium is equivalent to the event that $\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{C}}) > z_N + \alpha$. Now use Markov's inequality and apply the first part of the theorem. □

B.1.5 Proof of Theorem 4.8

Proof. Consider the first part of the theorem.

By construction, $\hat{\mathbf{x}}$ solves $\text{VI}(\mathbf{f}(\cdot, \theta_N), \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$. The theorem, then, claims that $\mathbf{x}_{N+1}$ is $\delta' \equiv \sqrt{\frac{z_N}{\gamma}}$ near a solution to this VI. From Theorem 4.1, if $\mathbf{x}_{N+1}$ were not δ' near a solution, then it must be the case that

$$\epsilon(\mathbf{f}(\cdot, \theta_N), \mathbf{x}_{N+1}, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1}) > z_N.$$

By Theorem 4.6, this happens only with probability $\beta(\alpha)$.

The second part is similar to the first with Theorem 4.6 replaced by Theorem 4.7.

□

B.2 Casting Structural Estimation as an Inverse Variational Inequality

In this appendix, we study how some structural estimation techniques can be viewed through the lens of inverse variational inequalities. Thus, in the spirit of structural estimation, assume there exists a *true* $\boldsymbol{\theta}^* \in \Theta$ that generates solutions $\mathbf{x}_j^*$ to $\text{VI}(\mathbf{f}(\cdot, \theta^*), \mathbf{A}_j^*, \mathbf{b}_j^*, C_j^*)$. We observe $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ which are noisy versions of these true parameters. We additionally are given a precise mechanism for the noise, e.g., that

$$\mathbf{x}_j = \mathbf{x}_j^* + \Delta\mathbf{x}_j, \quad \mathbf{A}_j = \mathbf{A}_j^* + \Delta\mathbf{A}_j, \quad \mathbf{b}_j = \mathbf{b}_j^* + \Delta\mathbf{b}_j, \quad C_j = C_j^*,$$

where $(\Delta\mathbf{x}_j, \Delta\mathbf{A}_j, \Delta\mathbf{b}_j)$ are i.i.d realizations of the random triplet $(\tilde{\Delta\mathbf{x}}, \tilde{\Delta\mathbf{A}}, \tilde{\Delta\mathbf{b}})$ and $\tilde{\Delta\mathbf{x}}, \tilde{\Delta\mathbf{A}}, \tilde{\Delta\mathbf{b}}$ are mutually uncorrelated.

We use Theorem 4.2 to estimate $\boldsymbol{\theta}$ under these assumptions by solving

$$\begin{aligned}\min_{\mathbf{y}\geq\mathbf{0},\boldsymbol{\theta}\in\Theta,\Delta\mathbf{x},\Delta\mathbf{A},\Delta\mathbf{b}} \quad & \left\| \begin{pmatrix} \Delta\tilde{\mathbf{x}}_j \\ \Delta\tilde{\mathbf{A}}_k \\ \Delta\tilde{\mathbf{b}}_j \end{pmatrix}_{j=1,\ldots,N} \right\| \\ \text{s.t.} \quad & (\mathbf{A}_j - \Delta\mathbf{A}_j)^T\mathbf{y}_j \leq_{C_j} \mathbf{f}(\mathbf{x}_j - \Delta\mathbf{x}_j, \boldsymbol{\theta}),\ j = 1,\ldots,N, \\ & (\mathbf{x}_j - \Delta\mathbf{x}_j)^T\mathbf{f}(\mathbf{x}_j - \Delta\mathbf{x}_j, \boldsymbol{\theta}) = \mathbf{b}_j^T\mathbf{y}_j, j = 1,\ldots,N, \end{aligned} \tag{B.5}$$

where $\|\cdot\|$ refers to some norm. Notice this formulation also supports the case where potentially some of the components of $\mathbf{x}$ are unobserved; simply replace them as optimization variables in the above. In words, this formulation assumes that the "de-noised" data constitute a perfect equilibrium with respect to the fitted $\boldsymbol{\theta}$.

We next claim that if we assume all equilibria occur on the strict interior of the feasible region, Problem (B.5) is equivalent to a least-squares approximate solution to the equations $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$. Specifically, when $\mathbf{x}^*$ occurs on the interior of $\mathcal{F}$, the VI condition Eq. (4.1) is equivalent to the equations $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$. At the same time, by Theorem 4.2, Eq. (4.1) is equivalent to the system (4.8), (4.9) with $\epsilon = 0$ which motivated the constraints in Problem (B.5). Thus, Problem (B.5) is equivalent to finding a minimal (with respect to the given norm) perturbation which satisfies the structural equations.

We can relate this weighted least-squares problem to some structural estimation techniques. Indeed, [53] and [107] observed that many structural estimation techniques can be reinterpreted as a constrained optimization problem which minimizes the size of the perturbation necessary to make the observed data satisfy the structural equations, and, additionally, satisfy constraints motivated by orthogonality conditions and the generalized method of moments (GMM). In light of our previous comments, if we augment Problem (B.5) with the same orthogonality constraints, and all equilibria occur on the strict interior of the feasible region, the solutions to this problem will coincide with traditional estimators.

Of course, some structural estimation techniques incorporate even more sophisti-

cated adaptations. They may also pre-process the data (e.g., 2 stage least squares technique in econometrics) incorporate additional constraints (e.g. orthogonality of instruments approach), or tune the choice of norm in the least-squares computation (two-stage GMM estimation). These application-specific adaptations improve the statistical properties of the estimator given certain assumptions about the data generating process. What we would like to stress is that, provided we make the same adaptations to Problem (B.5) – i.e., preprocess the data, incorporate orthogonality of instruments, and tune the choice of norm – and provided that all equilibria occur on the interior, the solution to Problem (B.5) must coincide exactly with these techniques. Thus, they necessarily inherit all of the same statistical properties.

Recasting (at least some) structural estimation techniques in our framework facilitates a number of comparisons to our proposed approach based on Problem (4.12). First, it is clear how our perspective on data alters the formulation. Problem (B.5) seeks minimal perturbations so that the observed data are exact equilibria with respect to $\boldsymbol{\theta}$, while Problem (4.12) seeks a $\boldsymbol{\theta}$ that makes the observed data approximate equilibria and minimizes the size of the approximation. Secondly, the complexity of the proposed optimization problems differs greatly. The complexity of Problem (B.5) depends on the dependence of $\mathbf{f}$ on $\mathbf{x}$ and $\boldsymbol{\theta}$ (as opposed to just $\boldsymbol{\theta}$ for (4.12)), and there are unavoidable non-convex, bilinear terms like $\Delta\mathbf{A}_j^T\mathbf{y}_j$. These terms are well-known to cause difficulties for numerical solvers. Thus, we expect that solving this optimization to be significantly more difficult than solving Problem (4.12). Finally, as seen in Sec. 4.5, Problem (4.12) generalizes naturally to a nonparametric setting.

B.3 Omitted Formulations

B.3.1 Formulation from Section 4.8.1

Let ξ^{med} be the median value of ξ over the dataset. Breaking ties arbitrarily, ξ^{med} occurs for some observation $j = j^{med}$. Let $p_1^{med}, p_2^{med}, \xi_1^{med}, \xi_2^{med}$ be the corresponding prices and demand shocks at time j^{med}. (Recall that in this section $\xi = \xi_1 = \xi_2$.)

These definitions make precise what we mean in the main text by "fixing other variables to the median observation. Denote by $\underline{p}_1, \underline{p}_2$ the minimum prices observed over the data set.

Our parametric formation in Sec. 4.8.1 is

$$\min_{\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} \quad \|\boldsymbol{\epsilon}\|_\infty \tag{B.6a}$$

$$\text{s.t.} \quad \mathbf{y}^j \geq \mathbf{0}, \quad j = 1, \ldots, N,$$

$$y_i^j \geq M_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i), \quad i = 1, 2, \ j = 1, \ldots, N,$$

$$\sum_{i=1}^{2} \overline{p}^j y_i^j - (p_i^j) M_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) \leq \epsilon_j, \quad j = 1, \ldots, N,$$

$$M_1(p_1^j, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1) \geq M_1(p_1^k, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1), \quad \forall 1 \leq j, k \leq N \text{ s.t. } p_1^j \leq p_1^k, \tag{B.6b}$$

$$M_2(p_1^{med}, p_2^j, \xi^{med}; \boldsymbol{\theta}_2) \geq M_2(p_1^{med}, p_2^k, \xi^{med}; \boldsymbol{\theta}_2), \quad \forall 1 \leq j, k \leq N \text{ s.t. } p_2^j \leq p_2^k, \tag{B.6c}$$

$$M_1(\underline{p}_1, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1) = M_1^*(\underline{p}_1, p_2^{med}, \xi_1^{med}; \boldsymbol{\theta}_1^*) \tag{B.6d}$$

$$M_2(p_1^{med}, \underline{p}_2, \xi^{med}; \boldsymbol{\theta}_2) = M_2^*(p_1^{med}, \underline{p}_2, \xi_2^{med}; \boldsymbol{\theta}_2^*) \tag{B.6e}$$

Here M_1 and M_2 are given by Eq. (4.32). Notice, for this choice, the optimization is a linear optimization problem.

Eqs. (B.6b) and (B.6c) constrain the fitted function to be non-decreasing in the firm's own price. Eqs. (B.6d) and (B.6e) are normalization conditions. We have chosen to normalize the functions to be equal to the true functions at this one point to make the visual comparisons easier. In principle, any suitable normalization can be used.

Our nonparametric formulation is similar to the above, but we replace

- The parametric $M_1(\cdot, \boldsymbol{\theta}_1), M_2(\cdot, \boldsymbol{\theta}_2)$ with nonparametric $M_1(\cdot), M_2(\cdot) \in \mathcal{H}$
- The objective by $\|\boldsymbol{\epsilon}\|_1 + \lambda(\|M_1\|_{\mathcal{H}} + \|M_2\|_{\mathcal{H}})$.

By Theorem 4.4 and the discussion in Section 4.6, we can rewrite this optimization as a convex quadratic program.

B.3.2 Formulation from Section 4.8.2

Our parametric formulation is nearly identical to the parametric formulation in Appendix B.3.1, with the following changes:

- Replace Eq. (B.6a) by $\|\boldsymbol{\epsilon}\|_\infty + \lambda(\|\boldsymbol{\theta}_1\|_1 + \|\boldsymbol{\theta}_2\|_1)$
- Replace the definition of M_1, M_2 by Eq. (4.33).

Our nonparametric formulation is identical to the nonparametric formulation of the previous section.

Bibliography

[1] T. Abrahamsson. "Estimation of origin-destination matrices using traffic counts : A literature survey". In: *International Institute for Applied Systems Analysis: Interim report* (1998). URL: www.iiasa.ac.at/Admin/PUB/Documents/IR-98-021.pdf.

[2] C. Acerbi and D. Tasche. "On the coherence of expected shortfall". In: *Journal of Banking & Finance* 26.7 (2002), pp. 1487–1503.

[3] E. Adida and G. Perakis. "A robust optimization approach to dynamic pricing and inventory control with no backorders". In: *Mathematical Programming* 107.1 (2006), pp. 97–129.

[4] M. Aghassi, D. Bertsimas, and G. Perakis. "Solving asymmetric variational inequalities via convex optimization". In: *Operations Research Letters* 34.5 (2006), pp. 481 –490. ISSN: 0167-6377. DOI: 10.1016/j.orl.2005.09.006. URL: http://www.sciencedirect.com/science/article/pii/S0167637705001124.

[5] R.K. Ahuja and J.B. Orlin. "Inverse Optimization". In: *Operations Research* 49.5 (2001), pp. 771–783.

[6] G. Allon, A. Federgruen, and M. Pierson. "How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods". In: *Manufacturing & Service Operations Management* 13.4 (2011), pp. 489–507.

[7] P. Artzner et al. "Coherent measures of risk". In: *Mathematical Finance* 9.3 (1999), pp. 203–228.

[8] P. Bajari, C.L. Benkard, and J. Levin. "Estimating dynamic models of imperfect competition". In: *Econometrica* 75.5 (2007), pp. 1331–1370.

[9] C. Bandi and D. Bertsimas. "Tractable stochastic analysis in high dimensions via robust optimization". In: *Mathematical Programming* 134.1 (2012), pp. 23–70.

[10] C. Bandi, D. Bertsimas, and N. Youssef. "Robust Queueing Theory". Submitted for publication to Operations Research. 2012.

[11] P.L. Bartlett and S. Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *The Journal of Machine Learning Research* 3 (2003), pp. 463–482.

[12] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[13] A. Ben-Tal and A. Nemirovski. "On polyhedral approximations of the second-order cone". In: *Mathematics of Operations Research* 26.2 (2001), pp. 193–205.

[14] A. Ben-Tal and A. Nemirovski. "Robust solutions of linear programming problems contaminated with uncertain data". In: *Mathematical Programming* 88.3 (2000), pp. 411–424.

[15] A. Ben-Tal et al. "Adjustable robust solutions of uncertain linear programs". In: *Mathematical Programming* 99.2 (2004), pp. 351–376.

[16] A. Ben-Tal et al. "Retailer-supplier flexible commitments contracts: a robust optimization approach". In: *Manufacturing & Service Operations Management* 7.3 (2005), pp. 248–271.

[17] A. Ben-Tal et al. "Robust solutions of optimization problems affected by uncertain probabilities". In: *Management Science* (2012).

[18] A. Ben-Tal et al. "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities". In: *Management Science* 59.2 (2013), pp. 341–357.

[19] S. Berry, J. Levinsohn, and A. Pakes. "Automobile prices in market equilibrium". In: *Econometrica: Journal of the Econometric Society* (1995), pp. 841–890.

[20] S.T. Berry. "Estimating discrete-choice models of product differentiation". In: *The RAND Journal of Economics* (1994), pp. 242–262.

[21] S.T. Berry and P.A. Haile. *Nonparametric identification of multinomial choice demand models with heterogeneous consumers*. Tech. rep. National Bureau of Economic Research, 2009.

[22] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA USA., 1999.

[23] D.P. Bertsekas, A. Nedi, A.E. Ozdaglar, et al. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[24] D. Bertsimas and D.B. Brown. "Constructing uncertainty sets for robust linear optimization". In: *Operations Research* 57.6 (2009), pp. 1483–1495.

[25] D. Bertsimas, I. Dunning, and M. Lubin. *Reformulations versus cutting planes for robust optimization: A computational and machine learning perspective*. Tech. rep. Sloan School of Business, Massachusetts Institute of Technology., In preparation.

[26] D. Bertsimas and V. Goyal. "On the power and limitations of affine policies in two-stage adaptive optimization". In: *Mathematical Programming* 134.2 (2012), pp. 491–531.

[27] D. Bertsimas and V. Gupta. *Data-Driven Uncertainty Sets for the Robust Unit Commitment Problem*. Tech. rep. Sloan School of Business, Massachusetts Institute of Technology., In preparation.

[28] D. Bertsimas, V. Gupta, and N. Kallus. "Robust Sample Average Approximation". In: *Technical report* (2013). In preparation.

[29] D. Bertsimas, V. Gupta, and I. Ch Paschalidis. "Inverse optimization: a new perspective on the Black-Litterman model". In: *Operations Research* 60.6 (2012), pp. 1389–1403.

[30] D. Bertsimas, D.A. Iancu, and P.A. Parrilo. "Optimality of affine policies in multistage robust optimization". In: *Mathematics of Operations Research* 35.2 (2010), pp. 363–394.

[31] D. Bertsimas, D. Pachamanova, and M. Sim. "Robust linear optimization under general norms". In: *Operations Research Letters* 32.6 (2004), pp. 510–516.

[32] D. Bertsimas and I. Popescu. "Optimal inequalities in probability theory: A convex optimization approach". In: *SIAM Journal on Optimization* 15.3 (2005), pp. 780–804.

[33] D. Bertsimas and M. Sim. "The price of robustness". In: *Operations Research* 52.1 (2004), pp. 35–53.

[34] D. Bertsimas and J.N. Tsitsiklis. "Introduction to linear optimization". In: (1997).

[35] D. Bertsimas et al. "Adaptive robust optimization for the security constrained unit commitment problem". In: *IEEE Transactions on Power Systems* (2011), pp. 52–63.

[36] D. Bertsimas et al. "Adaptive robust optimization for the security constrained unit commitment problem". In: *Power Systems, IEEE Transactions on* 28.1 (2013), pp. 52–63.

[37] G. Biau. "Analysis of a random forests model". In: *The Journal of Machine Learning Research* 98888.1 (2012), pp. 1063–1095.

[38] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[39] D. Branston. "Link capacity functions: A review". In: *Transportation Research* 10.4 (1976), pp. 223–236.

[40] L. Breiman and et al. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (2001), pp. 199–231.

[41] P. Brunette. *Overview of ISO NE*. ISO NE. Oct. 2, 2013. URL: {http://www.iso-ne.com/support/training/courses/isone_101/01_welcome_iso_overview.pdf}.

[42] G.C. Calafiore and L. El Ghaoui. "On distributionally robust chance-constrained linear programs". In: *Journal of Optimization Theory and Applications* 130.1 (2006), pp. 1–22.

[43] M.C Campi and A. Carè. "Random Convex Programs with L_1-Regularization: Sparsity and Generalization". In: *SIAM Journal on Control and Optimization* 51.5 (2013), pp. 3532–3557.

[44] M.C. Campi and S. Garatti. "The exact feasibility of randomized solutions of uncertain convex programs". In: *SIAM Journal on Optimization* 19.3 (2008), pp. 1211–1230.

[45] C. Chatfield. *The analysis of time series: an introduction*. CRC press, 2013.

[46] C. Chatfield. *Time-series forecasting*. CRC Press, 2000.

[47] W. Chen et al. "From CVaR to uncertainty set: Implications in joint chance-constrained optimization". In: *Operations Research* 58.2 (2010), pp. 470–485.

[48] X. Chen, M. Sim, and P. Sun. "A robust optimization perspective on stochastic programming". In: *Operations Research* 55.6 (2007), pp. 1058–1071.

[49] T.W.S. Chow and C.T. Leung. "Neural network based short-term load forecasting using weather compensation". In: *Power Systems, IEEE Transactions on* 11.4 (1996), pp. 1736–1742.

[50] S. Dafermos and A. Nagurney. "A network formulation of market equilibrium problems and variational inequalities". In: *Operations Research Letters* 3.5 (1984), pp. 247–250.

[51] H.A. David and H.N. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[52] E. Delage and Y. Ye. "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". In: *Operations Research* (2010). DOI: 10.1287/opre.1090.0741. eprint: http://or.journal.informs.org/content/early/2010/01/28/opre.1090.0741.full.pdf+html. URL: http://or.journal.informs.org/content/early/2010/01/28/opre.1090.0741.abstract.

[53] J.P. Dubé, J.T. Fox, and C.L. Su. "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation". In: *Econometrica* 80.5 (2012), pp. 2231–2267.

[54] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Vol. 57. CRC press, 1993.

[55] L. El Ghaoui, M. Oks, and F. Oustry. "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach". In: *Operations Research* 51.4 (2003), pp. 543–556.

[56] P. Embrechts, A. Höing, and A. Juri. "Using copulae to bound the Value-at-Risk for functions of dependent risks". English. In: *Finance and Stochastics* 7.2 (2003), pp. 145–167. ISSN: 0949-2984. DOI: 10.1007/s007800200085. URL: http://dx.doi.org/10.1007/s007800200085.

[57] M. Espinoza et al. “Electric load forecasting”. In: *Control Systems, IEEE* 27.5 (2007), pp. 43–57.

[58] T. Evgeniou, M. Pontil, and T. Poggio. “Regularization networks and support vector machines”. In: *Advances in Computational Mathematics* 13.1 (2000), pp. 1–50.

[59] J.H. Friedman. “On Multivariate Goodness–of–Fit and Two–Sample Testing”. In: *Proceedings of Phystat2003, http://www. slac. stanford. edu/econf C* 30908 (2004).

[60] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

[61] G. Gallego et al. “Price competition with the attraction demand model: Existence of unique equilibrium and its stability”. In: *Manufacturing & Service Operations Management* 8.4 (2006), pp. 359–375.

[62] F. Girosi, M. Jones, and T. Poggio. *Priors stabilizers and basis functions: From regularization to radial, tensor and additive splines*. Massachusetts Institute of Technology, Artificial Intelligence Library, 1993. URL: `http://dspace.mit.edu/bitstream/handle/1721.1/7212/AIM-1430.pdf?sequence=2`.

[63] *Global Energy Forecasting Competition 2012 - Load Forecasting*. Accessed March 2014. Kaggle, IEE Power, and Energy Systems. URL: `http://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting`.

[64] J. Goh and M. Sim. “Distributionally robust optimization and its tractable approximations”. In: *Operations Research* 58.4-part-1 (2010), pp. 902–917.

[65] D. Goldfarb and G. Iyengar. “Robust portfolio selection problems”. In: *Mathematics of Operations Research* 28.1 (2003), pp. 1–38.

[66] B.L. Gorissen and D. Den Hertog. “Robust counterparts of inequalities containing sums of maxima of linear functions”. In: *European Journal of Operational Research* 227.1 (2013), pp. 30–43.

[67] A. Gretton et al. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13 (2012), pp. 723–773.

[68] E. Guerre, I. Perrigne, and Q. Vuong. “Optimal Nonparametric Estimation of First-price Auctions”. In: *Econometrica* 68.3 (2000), pp. 525–574.

[69] Inc. Gurobi Optimization. *Gurobi Optimizer Reference Manual*. 2014. URL: `http://www.gurobi.com`.

[70] P.T. Harker and J.S. Pang. “Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications”. In: *Mathematical Programming* 48.1 (1990), pp. 161–220.

[71] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Vol. 2. 1. Springer, 2009.

[72] C. Heuberger. “Inverse Combinatorial Optimization: A Survey on Problems, Methods, and Results”. In: *Journal of Combinatorial Optimization* 8.3 (2004), pp. 329–361.

[73] B.G. Hillel. *Transportation Test Problems*. Accessed April 2014. URL: http://www.bgu.ac.il/~bargera/tntp/.

[74] H.S. Hippert, C.E. Pedreira, and R.C. Souza. "Neural networks for short-term load forecasting: A review and evaluation". In: *Power Systems, IEEE Transactions on* 16.1 (2001), pp. 44–55.

[75] *Hourly Zonal Information*. Accessed March 2014. ISO NE. URL: http://www.iso-ne.org/markets/hstdata/znl_info/hourly/index.html.

[76] S.J. Huang and K.R. Shih. "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations". In: *Power Systems, IEEE Transactions on* 18.2 (2003), pp. 673–679.

[77] D.A. Iancu, M. Sharma, and M. Sviridenko. "Supermodularity and affine policies in dynamic robust optimization". In: *Operations Research* 61.4 (2013), pp. 941–956.

[78] G. Iyengar and W. Kang. "Inverse conic programming with applications". In: *Operations Research Letters* 33.3 (2005), p. 319.

[79] R. Jiang and Y. Guan. *Data-driven chance constrained stochastic program*. Tech. rep. University of Florida., 2013. URL: www.optimization-online.org.

[80] J.F.C. Kingman. "Some inequalities for the queue GI/G/1". In: *Biometrika* 49.3/4 (1962), pp. 315–324.

[81] D. Klabjan, D. Simchi-Levi, and M. Song. "Robust Stochastic Lot-Sizing by Means of Histograms". In: *Production and Operations Management* (2013), pp. 691–710.

[82] L.J. LeBlanc, E.K. Morlok, and W.P. Pierskalla. "An efficient approach to solving the road network equilibrium traffic assignment problem". In: *Transportation Research* 9.5 (1975), pp. 309–318.

[83] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics, 2010.

[84] D.V. Lindley. "The theory of queues with a single server". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 48. 02. Cambridge University Press. 1952, pp. 277–289.

[85] M.S. Lobo et al. "Applications of second-order cone programming". In: *Linear Algebra and its Applications* 284.1 (1998), pp. 193–228.

[86] M. Lubin and I. Dunning. "Computing in Operations Research using Julia". In: *ArXiv preprint arXiv:1312.1431* (2013). URL: http://arxiv.org/abs/1312.1431.

[87] D.G. Luenberger. *Optimization by Vector Space Methods*. Wiley-Interscience, 1997.

[88] S. Nakayama, R. Connors, and D. Watling. "A Method of Estimating Parameters on Transportation Equilibrium Models: Toward Integration Analysis on Both Demand and Supply Sides". In: *Transportation Research Board Annual Meeting 2007* (2007).

[89] K. Natarajan, P. Dessislava, and M. Sim. "Incorporating asymmetric distributional information in robust value-at-risk optimization". In: *Management Science* 54.3 (2008), pp. 573–585.

[90] K. Natarajan, D. Pachamanova, and M. Sim. "Constructing risk measures from uncertainty sets". In: *Operations Research* 57.5 (2009), pp. 1129–1141.

[91] A. Nemirovski and A. Shapiro. "Convex approximations of chance constrained programs". In: *SIAM Journal on Optimization* 17.4 (2006), pp. 969–996.

[92] A. Nevo. "Measuring market power in the ready-to-eat cereal industry". In: *Econometrica* 69.2 (2001), pp. 307–342.

[93] N.P. Padhy. "Unit commitment-a bibliographical survey". In: *Power Systems, IEEE Transactions on* 19.2 (2004), pp. 1196–1205.

[94] J.S. Pang. "A posteriori error bounds for the linearly-constrained variational inequality problem". In: *Mathematics of Operations Research* (1987), pp. 474–484.

[95] D.C. Park et al. "Electric load forecasting using an artificial neural network". In: *Power Systems, IEEE Transactions on* 6.2 (1991), pp. 442–449.

[96] G. Perakis and G. Roels. "An analytical model for traffic delays and the dynamic user equilibrium problem". In: *Operations Research* 54.6 (2006), p. 1151.

[97] I. Popescu. "A semidefinite programming approach to optimal-moment bounds for convex classes of distributions". In: *Mathematics of Operations Research* 30.3 (2005), pp. 632–657.

[98] Bureau of Public Roads, ed. *Traffic Assignment Manual*. US Department of Commerce, Urban Planning Division. 1964.

[99] J. Rice. *Mathematical Statistics and Data Analysis*. Duxbury press, 2007.

[100] R.T. Rockafellar and S. Uryasev. "Optimization of conditional value-at-risk". In: *Journal of Risk* 2 (2000), pp. 21–42.

[101] P. Rusmevichientong and H. Topaloglu. "Robust assortment optimization in revenue management under the multinomial logit choice model". In: *Operations Research* 60.4 (2012), pp. 865–882.

[102] J. Rust. "Structural estimation of Markov decision processes". In: *Handbook of Econometrics* 4 (1994), pp. 3081–3143.

[103] C.T. See and M. Sim. "Robust approximation to multiperiod inventory management". In: *Operations Research* 58.3 (2010), pp. 583–594.

[104] J. Shawe-Taylor and N. Cristianini. *Estimating the moments of a random vector with applications*. 2003. URL: `http://eprints.soton.ac.uk/260372/1/EstimatingTheMomentsOfARandomVectorWithApplications.pdf`.

[105] A.J. Smola and B. Schölkopf. *Learning with Kernels*. MIT press, 1998.

[106] M.A. Stephens. "EDF statistics for goodness of fit and some comparisons". In: *Journal of the American Statistical Association* 69.347 (1974), pp. 730–737.

[107] C.L. Su and K.L. Judd. "Constrained optimization approaches to estimation of structural models". In: *Econometrica* 80.5 (2012), pp. 2213–2230.

[108] J.W. Taylor and R. Buizza. "Neural network load forecasting with weather ensemble predictions". In: *Power Systems, IEEE Transactions on* 17.3 (2002), pp. 626–632.

[109] O. Thas. *Comparing distributions*. Springer, 2010.

[110] H. Trevor, T. Robert, and F. Jerome. "The Elements of Statistical Learning: Data Mining, Inference and Prediction". In: *New York: Springer-Verlag* 1.8 (2001), pp. 371–406.

[111] A. Wächter and L. Biegler. "Line Search Filter Methods for Nonlinear Programming: Local Convergence". In: *SIAM Journal on Optimization* 16.1 (2005), pp. 32–48. DOI: 10.1137/S1052623403426544. eprint: http://epubs.siam.org/doi/pdf/10.1137/S1052623403426544. URL: http://epubs.siam.org/doi/abs/10.1137/S1052623403426544.

[112] G. Wahba. *Spline models for observational data*. Vol. 59. Society for Industrial Mathematics, 1990.

[113] A.J. Wood and B.F. Wollenberg. *Power generation, operation, and control*. John Wiley & Sons, 2012.

[114] H. Yang et al. "Estimation of origin-destination matrices from link traffic counts on congested networks". In: *Transportation Research Part B: Methodological* 26.6 (1992), pp. 417–434.

[115] L. Zhao and S. Dafermos. "General economic equilibrium and variational inequalities". In: *Operations Research Letters* 10.7 (1991), pp. 369–376.

[116] L. Zhao and B. Zeng. "Robust unit commitment problem with demand response and wind energy". In: *Power and Energy Society General Meeting, 2012 IEEE*. IEEE. 2012, pp. 1–8.

[117] S. Zhu and M. Fukushima. "Worst-case conditional value-at-risk with application to robust portfolio management". In: *Operations Research* 57.5 (2009), pp. 1155–1168.