From Data to Decisions Through New Interfaces Between Optimization and Statistics

by

Nathan Kallus

B.S., University of California, Berkeley (2009)
B.A., University of California, Berkeley (2009)

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 2015

Author .....
Sloan School of Management
May 15, 2015

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

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

From Data to Decisions Through New Interfaces Between Optimization and Statistics

by
Nathan Kallus

Submitted to the Sloan School of Management
on May 15, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The growing availability of data is creating opportunities for making better decisions, but in many circumstances it is yet unknown how to correctly leverage this data in systematic and optimal ways. In this thesis, we investigate new modes of data-driven decision making, enabled by novel connections we uncover between optimization and statistics. We pursue fundamental theory, specific methodologies, and revealing applications that advance data analytics from a tool of understanding to a decision-making engine.

In part I, we focus on the interface between predictive and prescriptive analytics. In the first half, we combine ideas from machine learning and operations research to prescribe optimal decisions given historical data and auxiliary, predictive observations. We develop theory on tractability, asymptotic optimality, and performance metrics and apply our methods to leverage large-scale web data to drive a real-world inventory-management system. In the second half, we study the problem of data-driven pricing and show that a naive but common predictive approach leaves money on the table whereas a theoretically-sound prescriptive approach we propose performs well in practice, demonstrated by a novel statistical test applied to data from a loan provider.

In part II, we focus on the interface between statistical hypothesis testing and optimization under uncertainty. In the first half, we propose a novel method for data-driven stochastic optimization that combines finite-sample guarantees with large-sample convergence by leveraging new theory linking distributionally-robust optimization and statistical hypothesis testing. In the second half, we develop data-driven uncertainty sets for robust optimization and demonstrate that, when data is available, our sets outperform conventional sets when used in their place in existing applications of robust optimization.

In part III, we focus on the interface between controlled experimentation and modern optimization. In the first half, we propose an optimization-based approach to constructing experimental groups with discrepancies in covariate data that are orders-of-magnitude smaller than any randomization-based approach. In the second half, we develop a unified theory of designs that balance covariate data and their optimality. We show no notion of balance exists without structure on outcomes' functional form, whereas with structure expressed using normed spaces, various existing designs

emerge as optimal and new designs arise that prove successful in practice.

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

Acknowledgments

First and foremost, I would like thank Dimitris Bertsimas and acknowledge his efforts and support that made this thesis possible. A discerning advisor, Dimitris was always there to guide me along the way. Dimitris went beyond the call of duty. His support was unwavering and he has always been my champion even when I did not listen to his prudent advice. I am eternally grateful for this. Dimitris inspired me to be the best academic I can be and to work on problems that matter. I have learned from him to choose research projects that I am passionate about and to pursue these projects tirelessly. I have been extremely fortunate to have Dimitris as my advisor.

Being at MIT, I have also been fortunate to have had the chance to talk with and enjoy the company of some of the world's most brilliant academics and all-around incredible persons. I would like to thank Vivek Farias and David Gamarnik for their guidance and support and for many interesting conversations, and, of course, for serving on my thesis committee. I would like to thank all the other faculty who have been my teachers, counsels, advocates, and friends – thank you Itai Ashalgi, Rob Freund, Michel Goemans, Patrick Jaillet, Retsef Levi, Jim Orlin, Asu Ozdaglar, Pablo Parrillo, Georgia Perakis, Andreas Schulz, David Simchi-Levi, John Tsitsiklis, Juan Pablo Vielma, and Roy Welsch.

The Operations Research Center (ORC) is an awe-inspiring place (once you look past the decor) with some of the world's most amazing people and every day I was there I felt lucky to be able count myself among them. I will dearly miss that feeling. My fellow students made the ORC the incredibly warm, supportive, and intellectually stimulating place it is. They are my friends, support, confidants, colleagues, and collaborators without whom I would have failed miserably. I am thankful to every single one of them and specifically would like to thank Ross Anderson, Chaitanya Bandi, Fernanda Bravo, Maxime Cohen, Adam Elmachtoub, Michael Frankovich, Paul Grigas, Swati Gupta, Vishal Gupta, Kris Johnson, Phillip Keller, Angie King, Will Ma, Yaron Shaposhnik, Stefano Tracá, Alexander Weinstein, Joline Uichanco, Yehua Wei, Nataly Youssef, and Daisy Zhuo. I would like to thank our committed co-directors Dimitris and Patrick for leading the ORC to greatness and keeping it that way. I would like to thank Laura Rose, the heart of the ORC, for accepting so many inexcusably late forms from me. I would like to thank Andrew Carvalho for helping me track down Dimitris when he was nowhere to be found at our scheduled meeting times.

I would like to thank my family for their love and support. I would like to thank my brother Yoav for so many interesting conversations about math and beyond and my sister Niva for always looking out for me. I would like to thank my parents Ron and Rachel for raising me to be who I am and for always believing in me.

This thesis is dedicated to my loving husband Mads Hvas Jensen who has stood by me through thick and thin. I would be no one without Mads – certainly there would be no thesis. Mads has always been supportive of my ambitions and every day encourages me to be the best person I can be in all aspects of life. He is my rock and the love of my life.

Thank you.

Contents

IV Appendices

List of Figures

List of Tables

Chapter 1

Introduction: Data and Decisions, Optimization and Statistics

The explosion in the availability and accessibility of machine-readable data in many applications of operations research and management science (OR/MS) is creating new and exciting opportunities for better decision making. This thesis explores how one can seize these opportunities by artful combination of optimization and statistics. The ambition of this thesis is to advance data analytics from a tool of understanding to a decision-making engine.

OR/MS has traditionally focused on prescribing decisions, usually via optimization and often under uncertainty about key quantities affecting objectives. Statistics, from its foundation, has been a tool of understanding, focusing on describing and predicting via such procedures as estimation, learning, and testing. Whereas the decision and its optimization has been the protagonist of OR/MS, data and its understanding has been the protagonist of statistics. If we are interested in making data-driven decisions, we must consider the combination of the two. This thesis combines ideas form the two and considers thoughtfully the complete process from data collection to decision making in order to come up with theory, methods, and applications of data-driven decision making that are principled, systematic, and optimal yet also succeed in practice. We do this by leveraging new interfaces we uncover between optimization and statistics – interfaces that allow us to trace the process from the statistical realm of data analysis to the operational realm of optimal decisions. We identify three particular such points of contact where new connections enable new modes of data-driven decision making.

1.1 The Interface Between Predictive and Prescriptive Analytics

Part I of this thesis focuses on the interface of predictive and prescriptive analytics and how to bridge the gap between the analysis of large-scale data and the making of relevant decisions in operations contexts. With the explosive growth of data, it is no wonder that machine learning (ML) and data mining have grown in importance in fa-

cilitating descriptive analyses (e.g. clustering) and predictive analyses (e.g. regression and classification) of such data leading to valuable insights including, for example, predicting movie earnings (Asur and Huberman 2010) and book sales (Gruhl et al. 2005) based on social media chatter. But such quantities being predicted are often of key interest in decision making. For example capacity allocation, facility location, shipment planning, and inventory management are all relevant decision-making problems that concern the quantities predicted in the examples given. Unfortunately, the powerful ML methods that have taken hold of data science (Hastie et al. 2001) do not address these decision-making questions, which at the end of the day are of primary interest to most analysts and managers. Applying these methods to make a prediction and basing one's decision solely on this prediction leads to woefully inadequate performance.

This leads us to consider the problem of learning how to make a decision under uncertainty and using predictive observations from large-scale historical data in Part I's Chapter 2. We combine ideas from ML and OR/MS to develop a framework and specific solutions to this problem. Accounting for the full process from data to decisions, we present theory that accounts for the statistical behavior of the decisions we make and the tractability of computing these. We demonstrate the power of this new approach in a real-world context by applying it to leverage large-scale web data, including web search trends, to drive the inventory-management system of an entertainment media distributor that includes a network of over 50,000 retailers.

In Part I's Chapter 3, we consider the more intricate problem of making a decision with an unknown effect on the objective and based on non-experimental data, focusing on data-driven pricing in particular. A naive but common approach to data-driven pricing involves constructing a predictive model by regressing demand on price and optimizing revenues implied by predicted demand, but such a predictive model may bear no relationship to the demand induced by prescribing a particular price and such a pricing strategy can leave money on the table. Therefore, we develop a direct prescriptive approach that considers the prescriptive effect that setting a price control may have on demand. We present both non-parametric and parametric approaches to prescriptive data-driven pricing. But at the end of day, the soundness of one's model is irrelevant insofar as revenues generated by a particular approach cannot be distinguished from optimal to a statistically significant degree. For this reason, extending recent work (Besbes et al. 2010), we develop a statistical test for revenue optimality of a particular prescription. Applying this test to real pricing problem and using data from a loan provider, we show, nonetheless, that predictive approaches to data-driven pricing fail in practice. On the other hand, we find that parametric approaches to data-driven pricing often suffice, but only when they take into account the prescriptive nature of the problem.

1.2 The Interface Between Hypothesis Testing and Optimization Under Uncertainty

Part II of this thesis focuses on the interface of statistical hypothesis testing and optimization under uncertainty. Decision-making in OR/MS is often framed as a problem of optimizing a control where there is uncertainty in key quantities that affect the optimization problem. There are a variety of paradigms for optimization under uncertainty including stochastic optimization (Shapiro and Andrzej 2003, Birge and Louveaux 2011) and robust optimization (Bertsimas and Sim 2004, Ben-Tal et al. 2009). Traditionally, corresponding models of the uncertainty – such as probability distributions or uncertainty sets – are derived from a priori assumptions or, sometimes, some estimation from data but with little consideration of the ensuing decision-making process. Statistical hypothesis testing is the process of assessing the validity of hypothesis about unknown distributions based on data. Thus, it allows us to assess our models. But it is not clear how to directly and correctly incorporate such a procedure into a decision-making process, what is the potential effect on decisions made by such a process, and how to make the complete data-to-decision process computationally tractable. These are the questions we address in this part of the thesis.

In Part II's Chapter 4, we propose a new approach to data-driven stochastic optimization, which we term *Robust SAA*, where SAA stands for sample average approximation. SAA is a popular approach to data-driven stochastic optimization whereby unknown true distributions are replaced by empirical distributions, which are their maximum-likelihood estimates. Because the true distributions in the nominal stochastic optimization problem are replaced with something else, the relationship between the corresponding optimal decisions is not always clear, except when sample sizes go to infinity where empirical distributions converge to true ones. Therefore, under mild assumptions, SAA is both computationally tractable and enjoys strong asymptotic performance guarantees, but similar guarantees do not typically hold in finite samples. The method we propose, Robust SAA, makes a decision that optimizes expected costs and revenues with respect to the worst-case distribution among those that pass a statistical goodness-of-fit test against the observed data. Robust SAA enjoys the tractability and favorable asymptotic behavior of SAA while also providing finite-sample performance guarantees for the decision it recommends. The key to Robust SAA is a novel connection between statistical hypothesis testing, SAA, and optimization that allows us to link properties of a data-driven *optimization* problem, such as finite-sample and asymptotic performance with respect to an objective, to *statistical* properties of an associated goodness-of-fit hypothesis test, such as statistical significance and consistency. As a theoretical consequence, we can describe the finite-sample and asymptotic performance of Robust SAA and some existing data-driven optimization methodologies. As a practical consequence, this connection sheds light on which data-driven formulations are likely to perform well in particular applications and enables us to leverage powerful applied statistical tools like bootstrapping to improve their performance in practice.

In Part II's Chapter 5, we propose a new way to construct uncertainty sets for robust optimization based directly on data. Robust optimization is a popular approach to optimization under uncertainty with a proven track record in practice where the key idea is to define an uncertainty set of possible realizations of uncertain parameters and then optimize against worst-case realizations within this set (for a review see Ben-Tal and Nemirovski 2002, Bertsimas et al. 2011b). The choice of uncertainty set is crucial for making effective decisions with robust optimization – too small and the decision may be too sensitive to unforeseen realizations of the parameters, too large and it may be far too conservative, not carefully shaped and the optimization problem may be computationally intractable to solve. Whereas there are a variety of proposals for uncertainty sets that are theoretically motivated and experimentally validated (Ben-Tal and Nemirovski 2000, Bertsimas and Sim 2004, Ben-Tal et al. 2009, Bandi and Bertsimas 2012), they all share a common paradigm of relying on a priori assumptions without direct deference to data to motivate the set and prove theoretical guarantees enjoyed by the sets. Building on the previous success of robust optimization, the question we address in this chapter is how to transform robust optimization into a data-driven methodology so to seize upon the opportunity for better decision making offered by the wide availability of data in applications. Toward this end, we propose a novel and general schema for designing uncertainty sets for robust optimization from data by leveraging statistical hypothesis tests. The approach is flexible and widely applicable, and robust optimization problems built from our new sets are computationally tractable, both theoretically and practically. Furthermore, optimal solutions to these problems enjoy a strong, finite-sample probabilistic guarantee. The approach can be used in the vast array of existing applications of robust optimization and our numerical experiments confirm that, when data is available, using data-driven uncertainty sets improves the performance of robust optimization decisions.

1.3 The Interface Between Controlled Experimentation and Modern Optimization

Part III of this thesis focuses on the interface of integer optimization and the statistics of controlled experiments. The sort of data often referred to as "Big Data" is powerful because it is cheap and plentiful – it is often a proxy to more structured information that is expensive or impossible to gather directly. But some information cannot be gleaned by observation alone, like the effect of a new drug or intervention. These must be gleaned through experiments. Field experiments have also recently grown in popularity in empirical operations management as a complement to observational studies. However, such experiments can often be prohibitively expensive and necessarily small. But large-scale observational data can help here too. A multitude of prognostic baseline covariates are today routinely recorded on each individual experimental unit such as past click behavior for experiments on the efficacy of new web ads, demographics for experiments on the efficacy of new social programs, or genetic and other biological characteristics for experiments on the efficacy of new pharmaceutical

drugs. The best use of such data, especially when high-dimensional, for improving power and precision of experiments is not always clear and many methods leave room for improvement and are not well motivated.

In Part III's Chapter 6, we investigate the impact of integer optimization on this statistical problem. Most approaches to assigning test subjects to experimental groups involve a great deal of randomization – whether it be complete randomization, blocked randomization, pair-matched randomization, or re-randomization. However, the goal is always to come up with experimental groups that are well-matched so to make comparison possible. With this objective in mind, we propose a new optimization-based approach to designing experimental groups that directly minimizes the imbalances in group means and variances of baseline covariates using mixed-integer optimization. We demonstrate that imbalances, when fully minimized, are orders-of-magnitude smaller than can ever be achieved by any randomization-based approach. At the same time, both hidden covariates and moments of observed covariates that are not directly incorporated into the optimization are no worse matched than under any other method. A new bootstrap-based hypothesis test allows for valid statistical inference with this new optimization-based approach and applying it in an example of an oncological study shows that the approach offers large gains in power. Where every subject can cost tens of thousands of dollars, optimal use of the data is important.

In Part III's Chapter 7, we take a step back and consider a contemplative view of the problem through the lens of optimization and functional analysis. Many designs, including all aforementioned designs as well as our approach above, attempt to balance baseline covariates a priori by assigning subjects before applying treatments (as compared to a regression adjustment after the fact). Each has an implicit metric for balance that it is addressing and trying to reduce (perhaps optimally). But each such balancing metric is different and it is not clear which one is correct. First, we establish a no-free-lunch theorem of causal inference that dictates that, without structural information on the functional form of how outcomes are associated with baseline covariates, there cannot be any notion of balance and complete randomization, which pays no attention to baseline covariates, is an optimal design from the point of view of minimax variance. On the other hand, imposing structural constraints in the form of normed vector spaces of functions gives rise to various balancing metrics as equal to the objective of worst-case variance and hence to optimal designs that minimize this objective. A mild restrictions such as Lipschitz continuity gives rise to pairwise matching, which had no obvious relationship to post-treatment variance previously. Taking a leaf from machine learning and restricting unknown functions in reproducing kernel Hilbert spaces, which can be non-parametric and dense in continuous functions, gives rise to new and powerful designs that can be achieved by solving integer optimization problems or their semidefinite relaxations. Theoretical results show that these designs achieve linear convergence (inverse exponential) in the part of the variance that can potentially be reduced by a priori balance as the number of subjects grows, whereas classical designs achieve only logarithmic rates (inverse polynomial). In particular, this means that only modest sample sizes are needed in order to balance high-dimensional prognostic covariates.

1.4 Organization and Conventions

For the sake of accessibility each chapter is made as self-contained as possible. Mathematical notation and other conventions are established independently in each chapter and may vary between chapters. For the sake of readability, some details such as overly technical proofs are omitted in certain places where so noted, in which case the omissions are included in the appendices.

Part I

The Interface Between Predictive and Prescriptive Analytics

Chapter 2

From Predictive to Prescriptive Analytics

In this chapter, we combine ideas from machine learning (ML) and operations research and management science (OR/MS) in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems. In a departure from other work on data-driven optimization and reflecting our practical experience with the data available in applications of OR/MS, we consider data consisting, not only of observations of quantities with direct effect on costs/revenues, such as demand or returns, but predominantly of observations of associated auxiliary quantities. The main problem of interest is a conditional stochastic optimization problem, given imperfect observations, where the joint probability distributions that specify the problem are unknown. We demonstrate that our proposed solution methods, which are inspired by ML methods such as local regression (LOESS), classification and regression trees (CART), and random forests (RF), are generally applicable to a wide range of decision problems. We prove that they are computationally tractable and asymptotically optimal under mild conditions even when data is not independent and identically distributed (iid) and even for censored observations. As an analogue to the coefficient of determination R^2, we develop a metric P termed the coefficient of prescriptiveness to measure the prescriptive content of data and the efficacy of a policy from an operations perspective. To demonstrate the power of our approach in a real-world setting we study an inventory management problem faced by the distribution arm of an international media conglomerate, which ships an average of 1 billion units per year. We leverage both internal data and public online data harvested from IMDb, Rotten Tomatoes, and Google to prescribe operational decisions that outperform baseline measures. Specifically, the data we collect, leveraged by our methods, accounts for an 88% improvement as measured by our coefficient of prescriptiveness.

2.1 Introduction

In today's data-rich world, many problems of operations research and management science (OR/MS) can be characterized by three primitives:

a) Data $\left\{y^1, \ldots, y^N\right\}$ on uncertain quantities of interest $Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$ such as simultaneous demands.

b) Auxiliary data $\left\{x^1, \ldots, x^N\right\}$ on associated covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ such as recent sale figures, volume of Google searches for a products or company, news coverage, or user reviews, where x^i is concurrently observed with y^i.

c) A decision z constrained in $\mathcal{Z} \subset \mathbb{R}^{d_z}$ made after some observation $X = x$ with the objective of minimizing the *uncertain* costs $c(z; Y)$.

Traditionally, decision-making under uncertainty in OR/MS has largely focused on the problem

$$v^{\text{stoch}} = \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right], \qquad z^{\text{stoch}} \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right] \tag{2.1}$$

and its multi-period generalizations and addressed its solution under a priori assumptions about the distribution μ_Y of Y (cf. Birge and Louveaux (2011)), or, at times, in the presence of data $\{y^1, \ldots, y^n\}$ in the assumed form of independent and identically distributed (iid) observations drawn from μ_Y (cf. Shapiro (2003), Shapiro and Nemirovski (2005), Kleywegt et al. (2002a)). (We will discuss examples of (2.1) in Section 2.1.1.) By and large, auxiliary data $\left\{x^1, \ldots, x^N\right\}$ has not been extensively incorporated into OR/MS modeling, despite its growing influence in practice.

From its foundation, machine learning (ML), on the other hand, has largely focused on supervised learning, or the prediction of a quantity Y (usually univariate) as a function of X, based on data $\left\{(x^1, y^1), \ldots, (x^N, y^N)\right\}$. By and large, ML does not address optimal decision-making under uncertainty that is appropriate for OR/MS problems.

At the same time, an explosion in the availability and accessibility of data and advances in ML have enabled applications that predict, for example, consumer demand for video games (Y) based on online web-search queries (X) (Choi and Varian (2012)) or box-office ticket sales (Y) based on Twitter chatter (X) (Asur and Huberman (2010)). There are many other applications of ML that proceed in a similar manner: use large-scale auxiliary data to generate predictions of a quantity that is of interest to OR/MS applications (Goel et al. (2010), Da et al. (2011), Gruhl et al. (2005, 2004), Kallus (2014a)). However, it is not clear how to go from a good prediction to a good decision. A good decision must take into account uncertainty wherever present. For example, in the absence of auxiliary data, solving (2.1) based on data $\{y^1, \ldots, y^n\}$ but using only the sample mean $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^i \approx \mathbb{E}\left[Y\right]$ and ignoring all other aspects of the data would generally lead to inadequate solutions to (2.1) and an unacceptable waste of good data.

In this chapter, we combine ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS

problems that leverage auxiliary observations. Specifically, the problem of interest is

$$v^*(x) = \min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X = x\right], \qquad z^*(x) \in \mathcal{Z}^*(x) = \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X = x\right], \tag{2.2}$$

where the underlying distributions are unknown and only data S_N is available, where

$$S_N = \left\{(x^1, y^1), \ldots, (x^N, y^N)\right\}.$$

The solution $z^*(x)$ to (2.2) represents the full-information optimal decision, which, via full knowledge of the unknown joint distribution $\mu_{X,Y}$ of (X, Y), leverages the observation $X = x$ to the fullest possible extent in minimizing costs. We use the term *predictive prescription* for any function $z(x)$ that prescribes a decision in anticipation of the future given the observation $X = x$. Our task is to use S_N to construct a data-driven predictive prescription $\hat{z}_N(x)$. Our aim is that its performance in practice, $\mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X = x\right]$, is close to that of the full-information optimum, $v^*(x)$.

Our key contributions include:

a) We propose various ways for constructing predictive prescriptions $\hat{z}_N(x)$ The focus of the chapter is predictive prescriptions that have the form

$$\hat{z}_N(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x) c(z; y^i), \tag{2.3}$$

where $w_{N,i}(x)$ are weight functions derived from the data. We motivate specific constructions inspired by a great variety of predictive ML methods, including for example and random forests (RF; Breiman (2001)). We briefly summarize a selection of these constructions that we find the most effective below.

b) We also consider a construction motivated by the traditional empirical risk minimization (ERM) approach to ML. This construction has the form

$$\hat{z}_N(\cdot) \in \arg\min_{z(\cdot) \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} c(z(x^i); y^i), \tag{2.4}$$

where $\mathcal{F}$ is some class of functions. We extend the standard ML theory of out-of-sample guarantees for ERM to the case of multivariate-valued decisions encountered in OR/MS problems. We find, however, that in the specific context of OR/MS problems, the construction (2.4) suffers from some limitations that do not plague the predictive prescriptions derived from (2.3).

c) We show that that our proposals are computationally tractable under mild conditions.

d) We study the asymptotics of our proposals under sampling assumptions more general than iid. Under appropriate conditions and for certain predictive prescriptions $\hat{z}_N(x)$ we show that costs with respect to the true distributions con-

verge to the full information optimum, i.e.,

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right] = v^*(x),$$

and that the limit points of the decision itself are optimizers of the full information problem (2.2), i.e.,

$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \mathcal{Z}^*(x),$$

both for almost everywhere x and almost surely. We also extend our results to the case of censored data (such as observing demand via sales).

e) We introduce a new metric P, termed *the coefficient of prescriptiveness*, in order to measure the efficacy of a predictive prescription and to assess the prescriptive content of covariates X, that is, the extent to which observing X is helpful in reducing costs. An analogue to the coefficient of determination R^2 of predictive analytics, P is a unitless quantity that is (eventually) bounded between 0 (not prescriptive) and 1 (highly prescriptive).

f) We demonstrate in a real-world setting the power of our approach. We study an inventory management problem faced by the distribution arm of an international media conglomerate. This entity manages over 0.5 million unique items at some 50,000 retail locations around the world, with which it has vendor-managed inventory (VMI) and scan-based trading (SBT) agreements. On average it ships about 1 billion units a year. We leverage both internal company data and, in the spirit of the aforementioned ML applications, large-scale public data harvested from online sources, including IMDb, Rotten Tomatoes, and Google Trends. These data combined, leveraged by our approach, lead to large improvements in comparison to baseline measures, in particular accounting for an 88% improvement toward the deterministic perfect-foresight counterpart.

Of our proposed constructions of predictive prescriptions $\hat{z}_N(x)$, the ones that we find to be generally the most broadly and practically effective are the following:

a) Motivated by k-nearest-neighbors regression (kNN; Altman (1992)),

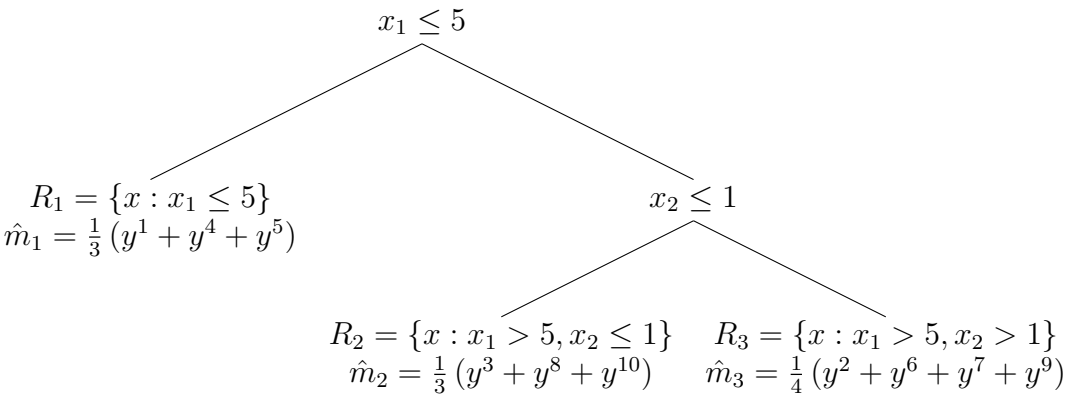
$$\hat{z}_N^{k\text{NN}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i\in\mathcal{N}_k(x)} c(z;y^i), \tag{2.5}$$

where $\mathcal{N}_k(x) = \left\{i = 1, \ldots, N : \sum_{j=1}^N \mathbb{I}\left[||x-x_i|| \geq ||x-x_j||\right] \leq k\right\}$ is the neighborhood of the k data points that are closest to x.

b) Motivated by local linear regression (LOESS; Cleveland and Devlin (1988)),

$$\hat{z}_N^{\text{LOESS}}(x) \in \arg\min_{z\in\mathcal{Z}} \sum_{i=1}^n k_i(x)\left(1 - \sum_{j=1}^n k_j(x)(x^j-x)^T\Xi(x)^{-1}(x^i-x)\right)c(z;y^i), \tag{2.6}$$

Figure 2-1: An Example of a Regression Tree and the Implicit Binning Rule $R(x)$



Implicit binning rule: $R(x) = (j \text{ s.t. } x \in R_j)$

Note: The regression tree is trained on data $\{(x^1, y^1), \ldots, (x^{10}, y^{10})\}$ and partitions the X data into regions defined by the leaves. The Y prediction $\hat{m}(x)$ is $\hat{m}_j$, the average of Y data at the leaf in which $X = x$ ends up. The implicit binning rule is $R(x)$, which maps x to the identity of the leaf in which it ends up.

where $k_i(x) = \left(1 - \left(||x^i - x|| / h_N(x)\right)^3\right)^3 \mathbb{I}\left[||x^i - x|| \leq h_N(x)\right]$ and the matrix $\Xi(x) = \sum_{i=1}^n k_i(x)(x^i - x)(x^i - x)^T$ and $h_N(x) > 0$ is the distance to the k-nearest point from x. Although this form may seem complicated, it corresponds to the simple idea of approximating $\mathbb{E}\left[c(z;Y)\middle|X = x\right]$ locally by a linear function in x, which we will discuss at greater length in Section 2.2.

c) Motivated by classification and regression trees (CART; Breiman et al. (1984)),

$$\hat{z}_N^{\text{CART}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i:R(x^i)=R(x)} c(z; y^i), \tag{2.7}$$

where $R(x)$ is the binning rule implied by a regression tree trained on the data S_N as shown in an example in Figure 2-1.

d) Motivated by random forests (RF; Breiman (2001)),

$$\hat{z}_N^{\text{RF}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{t=1}^{T} \frac{1}{|\{j : R^t(x^j) = R^t(x)\}|} \sum_{i:R^t(x^i)=R^t(x)} c(z; y^i), \tag{2.8}$$

where where $R^t(x)$ is the binning rule implied by the t^{th} tree in a random forest trained on the data S_N.

Further detail and other constructions are given in Sections 2.2 and 2.6.

In this chapter, we focus on the single-period problem (2.2), in which uncertain quantities are realized at one point in the problem. Such problems include, for example, two-stage decision problems where one set of decisions is made before uncertain quantities are realized and another set of decisions, known as the recourse, after. We study the more general and more intricate multi-period extensions to this problem, where uncertain quantities are realized at several points and in between subsequent decisions, in the multi-period extension to the present chapter, Bertsimas and Kallus (2015a).

2.1.1 Two Illustrative Examples

In this section, we illustrate various possible approaches to data-driven decision making and the value of auxiliary data in two examples. We illustrate this with synthetic data but, in Section 2.5, we study a real-world problem and use real-world data.

We first consider the mean-conditional-value-at-risk portfolio allocation problem. Here, our decision is how we would split our budget among each of d_y securities of which our portfolio may consist. The uncertain quantities of interest are $Y \in \mathbb{R}^{d_y}$, the returns of each of the securities, and the decisions are $z \in \mathbb{R}^{d_y}$, the allocation of the budget to each security. We are interested in maximizing the mean return while minimizing the conditional value at risk at level ϵ (CVaR_ϵ) of losses (negative return), which is the conditional expectation above the $1-\epsilon$ quantile. Following the reformulation of CVaR_ϵ due to Rockafellar and Uryasev (2000) and using an exchange rate λ between CVaR_ϵ and mean return, we can write this problem using an extra decision variable $\beta \in \mathbb{R}$, the following cost function for a realization y of returns

$$c((z,\beta);y) = \beta + \frac{1}{\epsilon}\max\left\{-z^T y - \beta, 0\right\} - \lambda z^T y,$$

and the feasible set

$$\mathcal{Z} = \left\{(z,\beta) \in \mathbb{R}^{d_y \times 1} \;:\; \beta \in \mathbb{R},\, z \geq 0,\, \sum_{i=1}^{d_y} z_i = 1\right\}.$$

The second example we consider is a two-stage shipment planning problem. Here we have a network of d_z warehouses that we use in order to satisfy the demand for a product at d_y locations. We consider two stages of the problem. In the first stage, some time in advance, we choose amounts $z_i \geq 0$ of units of product to produce and store at each warehouse i, at a cost of $p_1 > 0$ per unit produced. In the second stage, demand $Y \in \mathbb{R}^{d_y}$ realizes at the locations and we must ship units to satisfy it. We can ship from warehouse i to location j at a cost of c_{ij} per unit shipped (recourse variable $s_{ij} \geq 0$) and we have the option of using last-minute production at a cost of

$p_2 > p_1$ per unit (recourse variable t_i). The overall problem has the cost function

$$\begin{aligned} c(z;y) = p_1 \sum_{i=1}^{d_z} z_i \quad + \quad \min \quad & \left(p_2 \sum_{i=1}^{d_z} t_i + \sum_{i=1}^{d_z} \sum_{j=1}^{d_y} c_{ij} s_{ij} \right) \\ \text{s.t. } & t_i \geq 0 && \forall i \\ & s_{ij} \geq 0 && \forall i,\, j \\ & \sum_{i=1}^{d_z} s_{ij} \geq y_j && \forall j \\ & \sum_{j=1}^{d_y} s_{ij} \leq z_i + t_i && \forall i \end{aligned}$$

and the feasible set

$$\mathcal{Z} = \left\{ z \in \mathbb{R}^{d_z} \ : \ z \geq 0 \right\}.$$

The key in each problem is that we do not know Y or its distribution. We consider the situation where we only have data $S_N = ((x^1, y^1), \ldots, (x^N, y^N))$ consisting of observations of Y along with concurrent observations of some auxiliary quantities X that may be associated with the future value of Y. For example, in the portfolio allocation problem, X may include past security returns, behavior of underlying securities, analyst ratings, or volume of Google searches for a company together with keywords like "merger." In the shipment planning problem, X may include, for example, past product sale figures at each of the different retail locations, weather forecasts at the locations, or volume of Google searches for a product to measure consumer attention.

Let us consider two possible extant data-driven approaches to leveraging such data for making a decision. One approach is the sample average approximation of stochastic optimization (SAA, for short). SAA only concerns itself with the marginal distribution of Y, thus ignoring data on X, and solves the following data-driven optimization problem

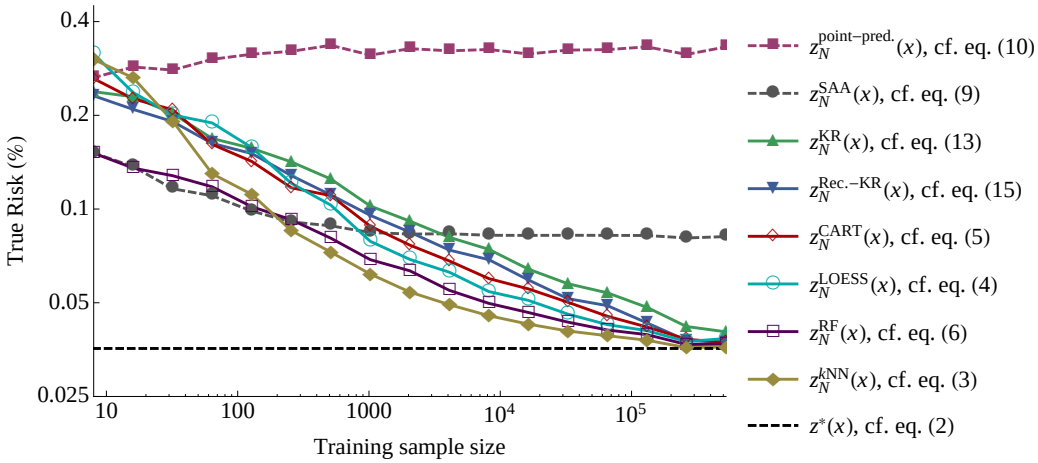
$$\hat{z}_N^{\text{SAA}} \in \arg\min_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} c(z; y^i). \tag{2.9}$$

The objective approximates $\mathbb{E}\left[c(z;Y)\right]$.

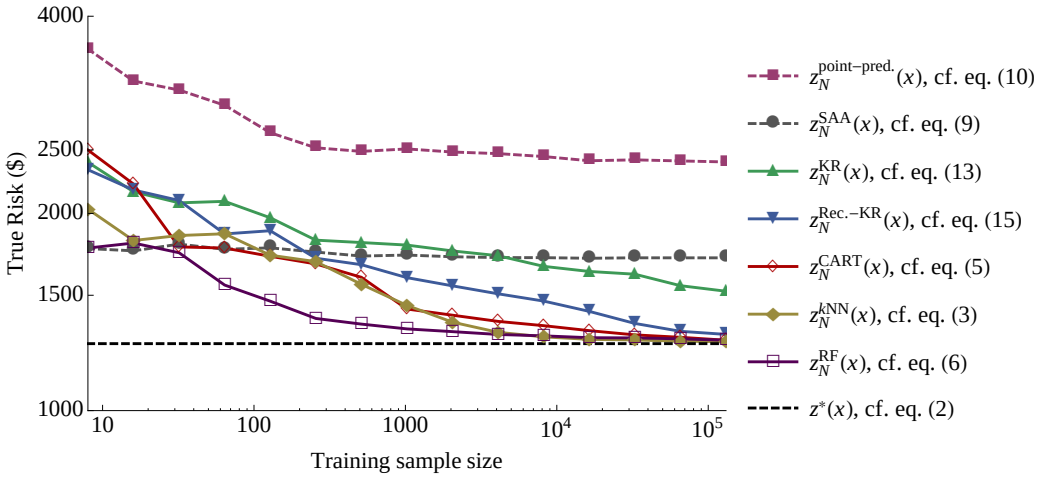
Machine learning, on the other hand, leverages the data on X as it tries to predict Y given observations $X = x$. Consider for example a random forest trained on the data S_N. It provides a point prediction $\hat{m}_N(x)$ for the value of Y when $X = x$. Given this prediction, one possibility is to consider the approximation of the random variable Y by our best-guess value $\hat{m}_N(x)$ and solve the corresponding optimization problem,

$$\hat{z}_N^{\text{point-pred}} \in \arg\min_{z \in \mathcal{Z}} c(z; \hat{m}_N(x)). \tag{2.10}$$

Figure 2-2: Comparison of Out-of-Sample Performance of Various Prescriptions



(a) Mean-CVaR portfolio allocation



(b) Two-stage shipment planning

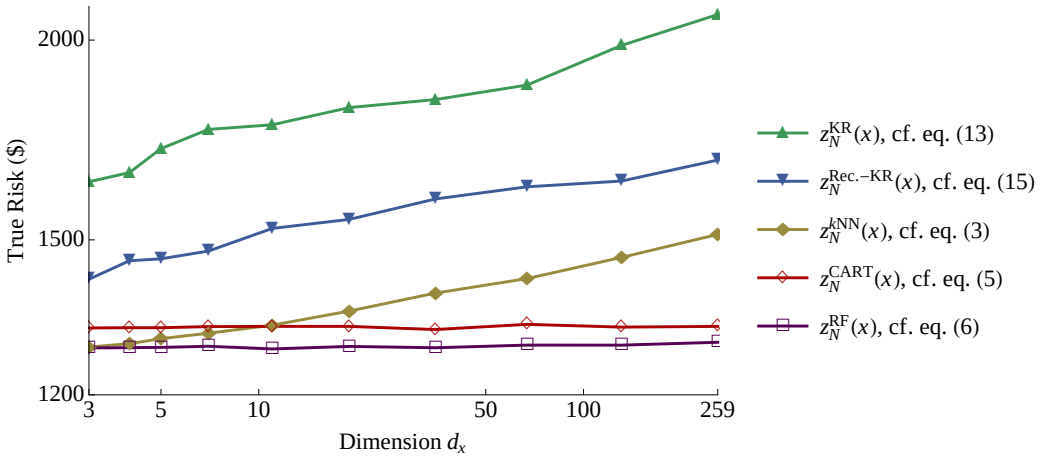
Note: Out-of-sample performance is averaged over data samples and new observations x with respect to true distributions. Lower is better. The horizontal and vertical are on log scales.

The objective approximates $c\left(z; \mathbb{E}\left[Y \middle| X = x\right]\right)$. We call (2.10) a point-prediction-driven decision.

If we knew the full joint distribution of Y and X, then the optimal decision having observed $X = x$ is given by (2.2). Let us compare SAA and the point-prediction-driven decision (using a random forest) to this optimal decision in the two decision problems presented. Let us also consider our proposals (2.5)-(2.8) and others that will be introduced in Section 2.2.

We consider a particular instance of the mean-CVaR portfolio allocation problem with $d_y = 12$ securities, where we observe some predictive market factors X before

Figure 2-3: The Dependence of Performance on the Dimension d_x in the Two-Stage Shipment Planning Example



Note: $N = 16384$.

making our investment. We also consider a particular instance of the two-stage shipment planning problem with $d_z = 5$ warehouses and $d_y = 12$ locations, where we observe some features predictive of demand. In both cases we consider $d_x = 3$ and data S_N that, instead of iid, is sampled from a multidimensional evolving process in order to simulate real-world data collection. We give the particular parameters of the problems in the supplementary Section A.4. In Figure 2-2, we report the average performance of the various solutions with respect to the *true* distributions.

The full-information optimum clearly does the best with respect to the true distributions, as expected. The SAA and point-prediction-driven decisions have performances that quickly converge to suboptimal values. The former because it does not use observations on X and the latter because it does not take into account the remaining uncertainty after observing $X = x$.[1] In comparison, we find that our proposals converge upon the full-information optimum given sufficient data. In Section 2.4.2, we study the general asymptotics of our proposals and prove that the convergence observed here empirically is generally guaranteed under only mild conditions.

Inspecting the figures further, it seems that ignoring X and using only the data on Y, as SAA does, is appropriate when there is very little data; in both examples, SAA outperforms other data-driven approaches for N smaller than $\sim$64. Past that point, our constructions of predictive prescriptions, in particular (2.5)-(2.8), leverage the auxiliary data effectively and achieve better, and eventually optimal, performance. The predictive prescription motivated by RF is notable in particular for performing no worse than SAA in the small N regime, and better in the large N regime.

In both examples, the dimension d_x of the observations x was relatively small at

[1]Note that the uncertainty of the point prediction in estimating the conditional expectation, gleaned e.g. via the bootstrap, is the wrong uncertainty to take into account, in particular because it shrinks to zero as $N \to \infty$.

$d_x = 3$. In many practical problems, this dimension may well be bigger, potentially inhibiting performance. E.g., in our real-world application in Section 2.5, we have $d_x = 91$. To study the effect of the dimension of x on the performance of our proposals, we consider polluting x with additional dimensions of uninformative components distributed as independent normals. The results, shown in Figure 2-3, show that while some of the predictive prescriptions show deteriorating performance with growing dimension d_x, the predictive prescriptions based on CART and RF are largely unaffected, seemingly able to detect the 3-dimensional subset of features that truly matter.

These examples serve as an illustration of the problems and data we tackle, existing approaches, and the gaps filled by our approach. In Section 2.5, we study an application of our approach to real-world problem and real – not synthetic – data.

2.1.2 Relevant Literature

Stochastic optimization as in (2.1) has long been the focus of decision making under uncertainty in OR/MS problems (cf. Birge and Louveaux (2011)) as has its multi-period generalization known commonly as dynamic programming (cf. Bertsekas (1995)). The solution of stochastic optimization problems as in (2.1) in the presence of data $\{y^1, \ldots, y^N\}$ on the quantity of interest is a topic of active research. The traditional approach is the sample average approximation (SAA) where the true distribution is replaced by the empirical one (cf. Shapiro (2003), Shapiro and Nemirovski (2005), Kleywegt et al. (2002a)). Other approaches include stochastic approximation (cf. Robbins and Monro (1951), Nemirovski et al. (2009)), robust SAA (cf. Bertsimas et al. (2014b)), and data-driven mean-variance distributionally-robust optimization (cf. Delage and Ye (2010), Calafiore and El Ghaoui (2006)). A notable alternative approach to decision making under uncertainty in OR/MS problems is robust optimization (cf. Ben-Tal et al. (2009), Bertsimas et al. (2011b)) and its data-driven variants (cf. Bertsimas et al. (2013), Calafiore and Campi (2005)). There is also a vast literature on the tradeoff between the collection of data and optimization as informed by data collected so far (cf. Robbins (1952), Lai and Robbins (1985), Besbes and Zeevi (2009)). In all of these methods for data-driven decision making under uncertainty, the focus is on data in the assumed form of iid observations of the parameter of interest Y. On the other hand, ML has attached great importance to the problem of supervised learning wherein the conditional expectation (regression) or mode (classification) of target quantities Y given auxiliary observations $X = x$ is of interest (cf. Hastie et al. (2001), Mohri et al. (2012)).

Statistical decision theory is generally concerned with the optimal selection of statistical estimators (cf. Berger (1985), Lehmann and Casella (1998)). Following the early work of Wald (1949), a loss function such as sum of squared errors or of absolute deviations is specified and the corresponding admissibility, minimax-optimality, or Bayes-optimality are of main interest. Statistical decision theory and ML intersect most profoundly in the realm of regression via empirical risk minimization (ERM), where a regression model is selected on the criterion of minimizing empirical average of loss. A range of ML methods arise from ERM applied to certain function classes

and extensive theory on function-class complexity has been developed to analyze these (cf. Bartlett and Mendelson (2003), Vapnik (2000, 1992)). Such ML methods include ordinary linear regression, ridge regression, the LASSO of Tibshirani (1996), quantile regression, and ℓ_1-regularized quantile regression of Belloni and Chernozhukov (2011).

In certain OR/MS decision problems, one can employ ERM to select a decision policy, conceiving of the loss as costs. Indeed, the loss function used in quantile regression is exactly equal to the cost function of the newsvendor problem of inventory management. Rudin and Vahn (2014) consider this loss function and the selection of a univariate-valued linear function with coefficients restricted in ℓ_1-norm in order to solve a newsvendor problem with auxiliary data, resulting in a method similar to Belloni and Chernozhukov (2011). Kao et al. (2009) study finding a convex combination of two ERM solutions, the least-cost decision and the least-squares predictor, which they find to be useful when costs are quadratic. In more general OR/MS problems where decisions are constrained, we show in Section 2.6 that ERM is not applicable. Even when it is, a linear decision rule may be inappropriate as we show by example. For the limited problems where ERM is applicable, we generalize the standard function-class complexity theory and out-of-sample guarantees to multivariate decision rules since most OR/MS problems involve multivariate decisions.

Instead of ERM, we are motivated more by a strain of non-parametric ML methods based on local learning, where predictions are made based on the mean or mode of past observations that are in some way similar to the one at hand. The most basic such method is kNN (cf. Altman (1992)), which define the prediction as a locally constant function depending on which k data points lie closest. A related method is Nadaraya-Watson kernel regression (KR) (cf. Nadaraya (1964), Watson (1964)), which is notable for being highly amenable to theoretical analysis but sees less use in practice. KR weighting for solving conditional stochastic optimization problems as in (2.2) has been considered in Hanasusanto and Kuhn (2013), Hannah et al. (2010) but these have not considered the more general connection to a great variety of ML methods used in practice and neither have they considered asymptotic optimality rigorously. A more widely used local learning regression method than KR is local regression (Cameron and Trivedi (2005) pg. 311) and in particular the LOESS method of Cleveland and Devlin (1988). Even more widely used are recursive partitioning methods, most often in the form of trees and most notably CART of Breiman et al. (1984). Ensembles of trees, most notably RF of Breiman (2001), are known to be very flexible and have competitive performance in a great range of prediction problems. The former averages locally over a partition designed based on the data (the leaves of a tree) and the latter combines many such averages. While there are many tree-based methods and ensemble methods, we focus on CART and RF because of their popularity and effectiveness in practice.

2.2 From Data to Predictive Prescriptions

Recall that we are interested in the conditional-stochastic optimization problem (2.2) of minimizing uncertain costs $c(z;Y)$ after observing $X = x$. The key difficulty

is that the true joint distribution $\mu_{X,Y}$, which specifies problem (2.2), is unknown and only data S_N is available. One approach may be to approximate $\mu_{X,Y}$ by the empirical distribution $\hat{\mu}_N$ over the data S_N where each datapoint (x^i, y^i) is assigned mass $1/N$. This, however, will in general fail unless X has small and finite support; otherwise, either $X = x$ has not been observed and the conditional expectation is undefined with respect to $\hat{\mu}_N$ or it has been observed, $X = x = x^i$ for some i, and the conditional distribution is a degenerate distribution with a single atom at y^i without any uncertainty. Therefore, we require some way to generalize the data to reasonably estimate the conditional expected costs for any x. In some ways this is similar to, but more intricate than, the prediction problem where $\mathbb{E}[Y|X = x]$ is estimated from data for any possible $x \in \mathcal{X}$. We are therefore motivated to consider predictive methods and their adaptation to our cause.

In the next subsections we propose a selection of constructions of predictive prescriptions $\hat{z}_N(x)$, each motivated by a local-learning predictive methodology. All the constructions in this section will take the common form of defining some data-driven weights

$$w_{N,i}(x) \quad : \quad \text{the weight associated with } y^i \text{ when observing } X = x, \tag{2.11}$$

and optimizing the decision $\hat{z}_N$ against a re-weighting of the data, as in (2.3):

$$\hat{z}_N^{\text{local}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x) c(z; y^i).$$

In some cases the weights are nonnegative and can be understood to correspond to an estimated conditional distribution of Y given $X = x$. But, in other cases, some of the weights may be negative and this interpretation breaks down.

2.2.1 *k*NN

Motivated by k-nearest-neighbor regression we propose

$$w_{N,i}^{k\text{NN}}(x) = \begin{cases} 1/k, & \text{if } x^i \text{ is a } k\text{NN of } x, \\ 0, & \text{otherwise,} \end{cases} \tag{2.12}$$

giving rise to the predictive prescription (2.5). Ties among equidistant data points are broken either randomly or by a lower-index-first rule. Finding the kNNs of x without pre-computation can clearly be done in $O(Nd)$ time. Data-structures that speed up the process at query time at the cost of pre-computation have been developed (cf. Bentley (1975)) and there are also approximate schemes that can significantly speed up queries (c.f. Arya et al. (1998)).

A variation of nearest neighbor regression is the radius-weighted k-nearest neighbors where observations in the neighborhood are weighted by a decreasing function

f in their distance:

$$w_{N,i}^{\text{radius-}k\text{NN}}(x) = \frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^{N} \tilde{w}_{N,j}(x)}, \quad \tilde{w}_{N,i}(x) = \begin{cases} f(||x^i - x||), & \text{if } x^i \text{ is a } k\text{NN of } x, \\ 0, & \text{otherwise.} \end{cases}$$

2.2.2 Kernel Methods

The Nadaraya-Watson kernel regression (KR; cf. Nadaraya (1964), Watson (1964)) estimates $m(x) = \mathbb{E}[Y|X = x]$ by

$$\hat{m}_N(x) = \frac{\sum_{i=1}^{N} y^i K\left((x^i - x)/h_N\right)}{\sum_{i=1}^{N} K\left((x^i - x)/h_N\right)},$$

where $K : \mathbb{R}^d \to \mathbb{R}$, known as the kernel, satisfies $\int K < \infty$ (and often unitary invariance) and $h_N > 0$, known as the bandwidth. For nonnegative kernels, KR is the result of the conditional distribution estimate that arises from the Parzen-window density estimates (cf. Parzen (1962)) of $\mu_{X,Y}$ and μ_X (i.e., their ratio). In particular, using the same conditional distribution estimate, the following weights lead to a predictive prescription as in (2.3):

$$w_{N,i}^{\text{KR}}(x) = \frac{K\left((x^i - x)/h_N\right)}{\sum_{j=1}^{N} K\left((x^j - x)/h_N\right)}. \tag{2.13}$$

Some common choices of nonnegative kernels are:

a) Naïve: $K(x) = \mathbb{I}\left[||x|| \leq 1\right]$.

b) Epanechnikov: $K(x) = (1 - ||x||^2)\mathbb{I}\left[||x|| \leq 1\right]$.

c) Tri-cubic: $K(x) = \left(1 - ||x||^3\right)^3 \mathbb{I}\left[||x|| \leq 1\right]$.

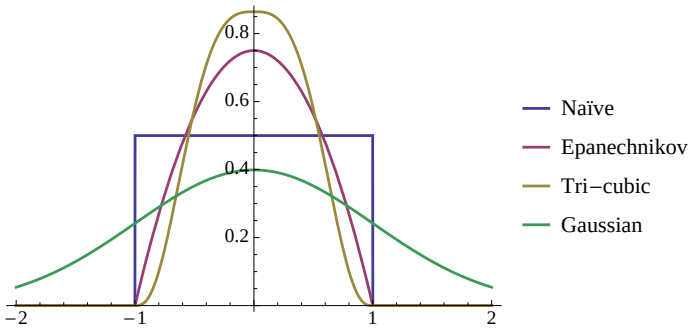
d) Gaussian: $K(x) = \exp\left(-||x||^2/2\right)$.

(2.14)

Note that the naïve kernel with bandwidth h_N corresponds directly to uniformly weighting all neighbors of x that are within a radius h_N. A comparison of different kernels is shown in Figure 2-4.

It is these weights (2.13) that are used in Hanasusanto and Kuhn (2013), Hannah et al. (2010) (without formally considering asymptotic optimality of $\hat{z}_N(x)$). A problem with KR is that it can be very biased in high dimensions, especially at the boundaries of the data (estimates will tend toward the outliers at the boundaries). While KR is particularly amenable to theoretical analysis due to its simplicity, it is not widely used in practice. We will next consider local regression, which is a related, more widely used approach. Before we proceed we first introduce a recursive modification to (2.13) that is motivated by an alternative kernel regressor introduced by

Figure 2-4: Comparison of the Different Kernels



Note: The kernels were normalized so they all integrate to 1.

Devroye and Wagner (1980):

$$w_{N,i}^{\text{recursive-KR}}(x) = \frac{K\left((x^i - x)/h_i\right)}{\sum_{j=1}^{N} K\left((x^j - x)/h_j\right)}, \tag{2.15}$$

where now the bandwidths h_i are selected per-data-point and independent of N. The benefits of (2.15) include the simplicity of an update when accumulating additional data since all previous weights remain unchanged as well as smaller variance under certain conditions (cf. Roussas (1992)). Moreover, from a theoretical point of view, much weaker conditions are necessary to ensure good asymptotic behavior of (2.15) compared to (2.13), as we will see in the next section.

2.2.3 Local Linear Methods

An alternative interpretation of KR predictions is they solves the locally-weighted least squares problem for a constant predictor:

$$\hat{m}_N(x) = \arg\min_{\beta_0} \sum_{i=1}^{N} k_i(x) \left(y^i - \beta_0\right)^2.$$

One can instead consider a predictive method that solves a similar locally-weighted least squares problem for a linear predictor:

$$\hat{m}_N(x) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{N} k_i(x) \left(y^i - \beta_0 - \beta_1^T (x^i - x)\right)^2.$$

In prediction, local linear methods are known to be preferable over KR (cf. Fan (1993)). Combined with a particular choice of $k_i(x)$, this results in the linear version of the LOESS variant of local regression developed in Cleveland and Devlin (1988). If, instead, we use this idea to locally approximate the conditional costs $\mathbb{E}\left[c(z;Y)\middle|X = x\right]$ by a linear function we will arrive at a functional estimate and a

predictive prescription as in (2.3) with the weights

$$w_{N,i}^{\text{LOESS}}(x) = \frac{\tilde{w}_{N,i}(x)}{\sum_{j=1}^{N} \tilde{w}_{N,j}(x)}, \tag{2.16}$$
$$\tilde{w}_{N,i}(x) = k_i(x) \left(1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^T \Xi(x)^{-1}(x^i - x)\right),$$

where $\Xi(x) = \sum_{i=1}^{n} k_i(x)(x^i - x)(x^i - x)^T$ and $k_i(x) = K\left((x^i - x)/h_N(x)\right)$. In the LOESS method, K is the tri-cubic kernel and $h_N(x)$ is not fixed, as it is in (2.13), but chosen to vary with x so that at each query point x the same number of data points taken into consideration; in particular, $h_N(x)$ is chosen to be the distance to x's k-nearest neighbor where k, in turn, is fixed. These choices lead to the form of $\hat{z}_N^{\text{LOESS}}(x)$ presented in Section 2.1.

2.2.4 Trees

Tree-based methods recursively split the sample S_N into regions in $\mathcal{X}$ so to gain reduction in "impurity" of the response variable Y within each region. The most well known tree-based predictive method is CART developed in Breiman et al. (1984). There are different definitions of "impurity," such as Gini or entropy for classification and variance reduction for regression, and different heuristics to choose the best split, different combinations resulting in different algorithms. Multivariate impurity measures are usually the component-wise average of univariate impurities. Splits are usually restricted to axis-aligned half-spaces. Such splits combined with the Gini impurity for classification and variance reduction for regression results in the original CART algorithm of Breiman et al. (1984). Once a tree is constructed, the value of $\mathbb{E}[Y|X = x]$ (or, the most likely class) is then estimated by the average (or, the mode) of y^i's associated with the x^i's that reside in the same region as x. The recursive splitting is most often represented as a tree with each non-leaf node representing an intermediate region in the algorithm (see Figure 2-1). With axis-aligned splits, the tree can be represented as subsequent inquiries about whether a particular component of the vector x is larger or smaller than a value. For a thorough review of tree-based methods and their computation see §9.2 of Hastie et al. (2001).

Regardless of the particular method chosen, the final partition can generally be represented as a binning rule identifying points in $\mathcal{X}$ with the disjoint regions, $\mathcal{R} : \mathcal{X} \to \{1, \ldots, r\}$. The partition is then the disjoint union $\mathcal{R}^{-1}(1) \sqcup \cdots \sqcup \mathcal{R}^{-1}(r) = \mathcal{X}$. The tree regression and classification estimates correspond directly to taking averages or modes over the uniform distribution of the data points residing in the region $R(x)$.

For our prescription problem, we propose to use the binning rule to construct weights as follows for a predictive prescription of the form (2.3):

$$w_{N,i}^{\text{CART}}(x) = \frac{\mathbb{I}\left[\mathcal{R}(x) = \mathcal{R}(x^i)\right]}{\left|\left\{j : R(x^j) = R(x)\right\}\right|}. \tag{2.17}$$

Notice that the weights (2.17) are piecewise constant over the partitions and therefore the recommended optimal decision $\hat{z}_N(x)$ is also piecewise constant. Therefore, solving r optimization problems after the recursive partitioning process, the resulting predictive prescription can be fully compiled into a decision tree, with the decisions that are truly decisions. This also retains CART's lauded interpretability. Apart from being interpretable, tree-based methods are also known to be useful in learning complex interactions and to perform well with large datasets.[2]

2.2.5 Ensembles

A random forest, developed in Breiman (2001), is an ensemble of trees each trained on a random subset of components of X and a random subsample of the data. This makes them more uncorrelated and therefore their average have lower variance. Random forests are one of the most flexible tools of ML and is extensively used in predictive applications. For a thorough review of random forests and their computation see §15 of Hastie et al. (2001).

After training such a random forest of trees, we can extract the partition rules $\mathcal{R}_t$ $t = 1, \ldots, T$, one for each tree in the forest. We propose to use these to construct the following weights as follows for a predictive prescription of the form (2.3):

$$w_{N,i}^{\mathrm{RF}}(x) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{I}\left[\mathcal{R}^t(x) = \mathcal{R}^t(x^i)\right]}{\left|\left\{j : R^t(x^j) = R^t(x)\right\}\right|}. \tag{2.18}$$

There are also other tree-ensembles methods. RF essentially combines the ideas from bagged (bootstrap-aggregated) forests (cf. Breiman (1996)) and random-subspace forests (cf. Ho (1998)). Other forest ensembles include extremely randomized trees developed in Geurts et al. (2006). The weights extracted from alternative tree ensembles would have the same form as (2.18).

In practice, RF is known as a flexible prediction algorithm that can perform competitively in almost any problem instance (cf. Breiman et al. (2001)). For our prescription problem, in Section 2.1.1 we saw that our predictive prescription based on RF, given in eq. (2.8), performed well overall in two different problems, for a range of sample sizes, and for a range of dimensions d_x. Based on this evidence of flexible performance, we choose our predictive prescription based on RF for our real-world application, which we study in Section 2.5.

[2] A more direct application of tree methods to the prescription problem would have us consider the impurities being minimized in each split to be equal to the mean cost $c(z; y)$ of taking the best constant decision z in each side of the split. However, since we must consider splitting on each variable and at each data point to find the best split (cf. pg. 307 of Hastie et al. (2001)), this can be overly computationally burdensome for all but the simplest problems that admit a closed form solution such as least sum of squares or the newsvendor problem.

2.3 Metrics of Prescriptiveness

In this section, we develop a relative, unitless measure of the efficacy of a predictive prescription. An absolute measure of efficacy is marginal expected costs,

$$R(\hat{z}_N) = \mathbb{E}\left[\mathbb{E}\left[c\left(\hat{z}_N(X);Y\right)\middle|X\right]\right] = \mathbb{E}\left[c\left(\hat{z}_N(X);Y\right)\right].$$

Given a validation data set $\tilde{S}_{N_v} = \left((\tilde{x}^1,\tilde{y}^1),\cdots,(\tilde{x}^{N_v},\tilde{y}^{N_v})\right)$, we can estimate $R(\hat{z}_N)$ as the sample average:

$$\hat{R}_{N_v}(\hat{z}_N) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\tilde{x}^i);\tilde{y}^i\right).$$

If $\tilde{S}_{N_v} = S_N$ then this an in-sample estimate, which biases in favor of overfitting, and if $\tilde{S}_{N_v}$ is disjoint, then this is an out-of-sample estimate that provides an unbiased estimate of $R(\hat{z}_N)$.

While an absolute measure allows one to compare two predictive prescriptions for the same problem and data, a relative measure can quantify the overall prescriptive content of the data and the efficacy of a prescription on a universal scale. For example, in predictive analytics, the coefficient of determination R^2 – rather than the absolute root-mean-squared error – is a unitless quantity used to quantify the overall quality of a prediction and the predictive content of data X. R^2 measures the fraction of variance of Y reduced, or "explained," by the prediction based on X. Another way of interpreting R^2 is as the fraction of the way that X and a particular predictive model take us from a data-poor prediction (the sample average) to a perfect-foresight prediction that knows Y in advance.

We define an analogous quantity for the predictive prescription problem, which we term *the coefficient of prescriptiveness*. It involves three quantities. First,

$$\hat{R}_{N_v}\left(\hat{z}_N(x)\right) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}_N(\tilde{x}^i);\tilde{y}^i\right)$$

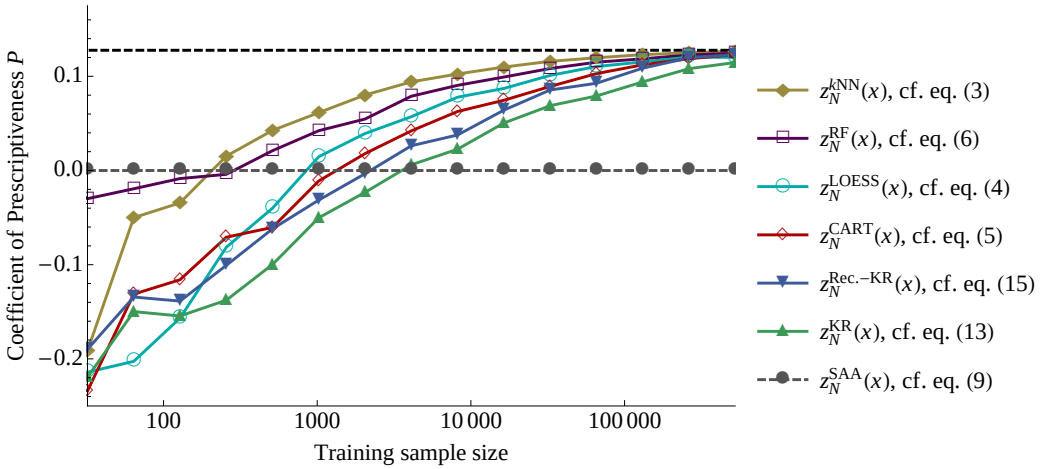
is the estimated expected costs due to our predictive prescription. Second,

$$\hat{R}^*_{N_v} = \frac{1}{N_v}\sum_{i=1}^{N_v} \min_{z\in\mathcal{Z}} c\left(z;\tilde{y}^i\right)$$

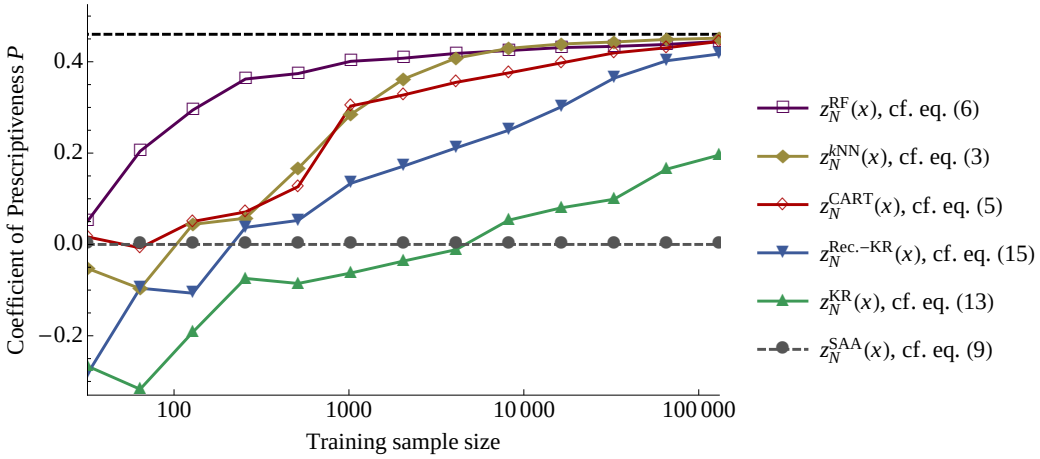
is the estimated expected costs in the deterministic perfect-foresight counterpart problem, in which one has foreknowledge of Y without any uncertainty (note the difference to the full-information optimum, which does have uncertainty). Third,

$$\hat{R}_{N_v}\left(\hat{z}^{\text{SAA}}_N\right) = \frac{1}{N_v}\sum_{i=1}^{N_v} c\left(\hat{z}^{\text{SAA}}_N;\tilde{y}^i\right) \quad \text{where} \quad \hat{z}^{\text{SAA}}_N \in \arg\min_{z\in\mathcal{Z}} \frac{1}{N}\sum_{i=1}^{N} c\left(z;y^i\right)$$

Figure 2-5: The Coefficient of Prescriptiveness P



(a) Mean-CVaR portfolio allocation



(b) Two-stage shipment planning

Note: The examples follow Section 2.1.1. Here P is measured out of sample. The black horizontal line denotes the theoretical limit.

is the estimated expected costs of a data-driven prescription that is data poor, based only on Y data. This is the SAA solution to the prescription problem, which serves as the analogue to the sample average as a data-poor solution to the prediction problem. Using these three quantities, we define the coefficient of prescriptiveness P as follows:

$$P = 1 - \frac{\hat{R}_{N_v}\left(\hat{z}_N(x)\right) - \hat{R}^*_{N_v}}{\hat{R}_{N_v}\left(\hat{z}^{\text{SAA}}_N\right) - \hat{R}^*_{N_v}} \tag{2.19}$$

When $\tilde{S}_{N_v} = S_N$ (in-sample) we can write

$$P = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} c\left(\hat{z}_N(x^i); y^i\right) - \frac{1}{N}\sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c\left(z; y^i\right)}{\min_{z\in\mathcal{Z}} \frac{1}{N}\sum_{i=1}^{N} c\left(z; y^i\right) - \frac{1}{N}\sum_{i=1}^{N} \min_{z\in\mathcal{Z}} c\left(z; y^i\right)}.$$

The coefficient of prescriptiveness P is a unitless quantity bounded above by 1. A low P denotes that X provides little useful information for the purpose of prescribing an optimal decision in the particular problem at hand or that $\hat{z}_N(x)$ is ineffective in leveraging the information in X. A high P denotes that taking X into consideration has a significant impact on reducing costs and that $\hat{z}_N$ is effective in leveraging X for this purpose.

Let us consider the coefficient of prescriptiveness in the two examples from Section 2.1.1. For each of our predictive prescriptions and for each N, we measure the out of sample P on a validation set of size $N_v = 200$ and plot the results in Figure 2-5. Notice that even when we converge to the full-information optimum, P does not approach 1 as N grows. Instead we see that for the same methods that converged to the full-information optimum in Figure 2-2, we have a P that approaches 0.13 in the portfolio allocation example and 0.46 in the shipment planning example. This number represents the extent of the potential that X has to reduce costs in this particular problem. It is the fraction of the way that knowledge of X, leveraged correctly, takes us from making a decision under full uncertainty about the value of Y to making a decision in a completely deterministic setting. As is the case with R^2, what magnitude of P denotes a successful application depends on the context. In our real-world application in Section 2.5, we find an out-of-sample P of 0.88.

2.4 Properties of Local Predictive Prescriptions

In this section, we study two important properties of local predictive prescriptions: computational tractability and asymptotic optimality. All proofs are given in the E-companion.

2.4.1 Tractability

In Section 2.2, we considered a variety of predictive prescriptions $\hat{z}_N(x)$ that are computed by solving the optimization problem (2.3). An important question is then when is this optimization problem computationally tractable to solve. As an optimization problem, problem (2.3) differs from the problem solved by the standard SAA approach (2.9) only in the weights given to different observations. Therefore, it is similar in its computational complexity and we can defer to computational studies of SAA such as Shapiro and Nemirovski (2005) to study the complexity of solving problem (2.3). For completeness, we develop sufficient conditions for problem (2.3) to be solvable in

polynomial time.

Theorem 2.1. *Fix x and weights $w_{N,i}(x) \geq 0$. Suppose $\mathcal{Z}$ is a closed convex set and let a separation oracle for it be given. Suppose also that $c(z; y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z. Then for any x we can find an ϵ-optimal solution to (2.3) in time and oracle calls polynomial in N_0, d, $\log(1/\epsilon)$ where $N_0 = \sum_{i=1}^{N} \mathbb{I}\left[w_{N,i}(x) > 0\right] \leq N$ is the effective sample size.*

Note that all of the weights presented in Section 2.2 have been necessarily all nonnegative with the exception of local regression (2.16). As the spans $h_N(x)$ shrink and the number of data points increases, these weights will always become nonnegative. However, for a fixed problem the weights $w_{N,i}^{\mathrm{LOESS}}(x)$ may be negative for some i and x, in which case the optimization problem (2.3) may not be polynomially solvable. In particular, in the case of the portfolio example presented in Section 2.1.1, if some weights are negative, we formulate the corresponding optimization problem as a mixed integer-linear optimization problem.

2.4.2 Asymptotic Optimality

In Section 2.1.1, we saw that our predictive prescriptions $\hat{z}_N(x)$ converged to the full-information optimum as the sample size N grew. Next, we show that this anecdotal evidence is supported by mathematics and that such convergence is guarantees under only mild conditions. We define *asymptotic optimality* as the desirable asymptotic behavior for $\hat{z}_N(x)$.

Definition 2.2. We say that $\hat{z}_N(x)$ is *asymptotically optimal* if, with probability 1, we have that for μ_X-almost-everywhere $x \in \mathcal{X}$, as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x); Y)\middle| X = x\right] = v^*(x),$$
$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \mathcal{Z}^*(x),$$

where $L(A)$ denotes the limit points of A.

Asymptotic optimality depends on our choice of $\hat{z}_N(x)$, the structure of the decision problem (cost function and feasible set), and on how we accumulate our data S_N. The traditional assumption on data collection is that it constitutes an iid process. This is a strong assumption and is often only a modeling approximation. The velocity and variety of modern data collection often means that historical observations do not generally constitute an iid sample in any real-world application. We are therefore motivated to consider an alternative model for data collection, that of mixing processes. These encompass such processes as ARMA, GARCH, and Markov chains, which can correspond to sampling from evolving systems like prices in a market, daily product demands, or the volume of Google searches on a topic. While many of our results extend to such settings, we present only the iid case in the main text to avoid cumbersome exposition and defer these extensions to the supplemental Section A.1.2. For the rest of the section let us assume that S_N is generated by iid sampling.

As mentioned, asymptotic optimality also depends on the structure of the decision problem. Therefore, we will also require the following conditions.

Assumption 2.3 (Existence)**.** The full-information problem (2.2) is well defined: $\mathbb{E}\left[|c(z;Y)|\right] < \infty$ for every $z \in \mathcal{Z}$ and $\mathcal{Z}^*(x) \neq \varnothing$ for almost every x.

Assumption 2.4 (Continuity)**.** $c(z;y)$ is equicontinuous in z: for any $z \in \mathcal{Z}$ and $\epsilon > 0$ there exists $\delta > 0$ such that $|c(z;y) - c(z';y)| \leq \epsilon$ for all z' with $||z - z'|| \leq \delta$ and $y \in \mathcal{Y}$.

Assumption 2.5 (Regularity)**.** $\mathcal{Z}$ is closed and nonempty and in addition either

1. $\mathcal{Z}$ is bounded,

2. $\mathcal{Z}$ is convex and $c(z;y)$ is convex in z for every $y \in \mathcal{Y}$, or

3. $\liminf_{||z||\to\infty} \inf_{y\in\mathcal{Y}} c(z;y) > -\infty$ and for every $x \in \mathcal{X}$, there exists $D_x \subset \mathcal{Y}$ such that $\lim_{||z||\to\infty} c(z;y) \to \infty$ uniformly over $y \in D_x$ and $\mathbb{P}\left(y \in D_x \middle| X = x\right) > 0$.

Under these conditions, we have the following sufficient conditions for asymptotic optimality.

Theorem 2.6 (kNN)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold. Let $w_{N,i}(x)$ be as in* (2.12) *with $k = \min\left\{\lceil CN^\delta \rceil, N-1\right\}$ for some $C > 0$, $0 < \delta < 1$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal.*

Theorem 2.7 (Kernel Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold and that $\mathbb{E}\left[|c(z;Y)| \max\left\{\log|c(z;Y)|, 0\right\}\right] < \infty$ for each z. Let $w_{N,i}(x)$ be as in* (2.13) *with K being any of the kernels in* (2.14) *and $h_N = CN^{-\delta}$ for $C > 0$, $0 < \delta < 1/d_x$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal.*

Theorem 2.8 (Recursive Kernel Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold. Let $w_{N,i}(x)$ be as in* (2.15) *with K being the naïve kernel and $h_i = Ci^{-\delta}$ for some $C > 0$, $0 < \delta < 1/(2d_x)$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal.*

Theorem 2.9 (Local Linear Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X, and that costs are bounded over y for each z (i.e., $|c(z;y)| \leq g(z)$). Let $w_{N,i}(x)$ be as in* (2.16) *with K being any of the kernels in* (2.14) *and with $h_N = CN^{-\delta}$ for some $C > 0$, $0 < \delta < 1/d_x$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal.*

Although we do not have firm theoretical results on the asymptotic optimality of the predictive prescriptions based on CART (eq. (2.7)) and RF (eq. (2.8)), we have observed them to converge empirically in Section 2.1.1.

2.5 A Real-World Application

In this section, we apply our approach to a real-world problem faced by the distribution arm of an international media conglomerate (the vendor) and demonstrate that our approach, combined with extensive data collection, leads to significant advantages. The vendor has asked us to keep its identity confidential as well as data on sale figures and specific retail locations. Some figures are therefore shown on relative scales.

2.5.1 Problem Statement

The vendor sells over 0.5 million entertainment media titles on CD, DVD, and BluRay at over 50,000 retailers across the US and Europe. On average they ship 1 billion units in a year. The retailers range from electronic home goods stores to supermarkets, gas stations, and convenience stores. These have vendor-managed inventory (VMI) and scan-based trading (SBT) agreements with the vendor. VMI means that the inventory is managed by the vendor, including replenishment (which they perform weekly) and planogramming. SBT means that the vendor owns all inventory until sold to the consumer. Only at the point of sale does the retailer buy the unit from the vendor and sell it to the consumer. This means that retailers have no cost of capital in holding the vendor's inventory.

The cost of a unit of entertainment media consists mainly of the cost of production of the content. Media-manufacturing and delivery costs are secondary in effect. Therefore, the primary objective of the vendor is simply to sell as many units as possible and the main limiting factor is inventory capacity at the retail locations. For example, at many of these locations, shelf space for the vendor's entertainment media is limited to an aisle endcap display and no back-of-the-store storage is available. Thus, the main loss incurred in over-stocking a particular product lies in the loss of potential sales of another product that sold out but could have sold more. In studying this problem, we will restrict our attention to the replenishment and sale of video media only and to retailers in Europe.

Apart from the limited shelf space the other main reason for the difficulty of the problem is the particularly high uncertainty inherent in the initial demand for new releases. Whereas items that have been sold for at least one period have a somewhat predictable decay in demand, determining where demand for a new release will start is a much less trivial task. At the same time, new releases present the greatest opportunity for high demand and many sales.

We now formulate the full-information problem. Let $r = 1, \ldots, R$ index the locations, $t = 1, \ldots, T$ index the replenishment periods, and $j = 1, \ldots, d$ index the products. Denote by z_j the order quantity decision for product j, by Y_j the uncertain demand for product j, and by K_r the overall inventory capacity at location r. Considering only the *main* effects on revenues and costs as discussed in the previous paragraph, the problem decomposes on a per-replenishment-period, per-location

basis. We therefore wish to solve, for each t and r, the following problem:

$$v^*(x_{tr}) = \max \quad \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j,\, z_j\} \,\middle|\, X = x_{tr}\right] = \sum_{j=1}^{d} \mathbb{E}\left[\min\{Y_j,\, z_j\} \,\middle|\, X_j = x_{tr}\right] \tag{2.20}$$

$$\text{s.t.} \quad \sum_{j=1}^{d} z_j \leq K_r$$
$$z_j \geq 0 \qquad\qquad \forall j = 1, \ldots, d,$$

where x_{tr} denotes auxiliary data available at the beginning of period t in the $(t,r)^{\text{th}}$ problem.

Note that had there been no capacity constraint in problem (2.20) and a per-unit ordering cost were added, the problem would decompose into d separate newsvendor problems, the solution to each being exactly a quantile regression on the regressors x_{tr}. As it is, the problem is coupled, but, fixing x_{tr}, the capacity constraint can be replaced with an equivalent per-unit ordering cost λ via Lagrangian duality and the optimal solution is attained by setting each z_j to the λ^{th} conditional quantile of Y_j. However, the reduction to quantile regression does not hold since the dual optimal value of λ depends *simultaneously* on all of the conditional distributions of Y_j for $j = 1, \ldots, d$.

2.5.2 Applying Predictive Prescriptions to Censored Data

In applying our approach to problem (2.20), we face the issue that we have data on sales, not demand. That is, our data on the quantity of interest Y is right-censored. In this section, we develop a modification of our approach to correct for this. The results in this section apply generally.

Suppose that instead of data $\{y^1, \ldots, y^N\}$ on Y, we have data $\{u^1, \ldots, u^N\}$ on $U = \min\{Y, V\}$ where V is an observable random threshold, data on which we summarize via $\delta = \mathbb{I}[U < V]$. For example, in our application, V is the on-hand inventory level at the beginning of the period. Overall, our data consists of $\tilde{S}_N = \left\{(x^1, u^1, \delta^1), \ldots, (x^N, u^N, \delta^N)\right\}$.

In order to correct for the fact that our observations are in fact censored, we develop a conditional variant of the Kaplan-Meier method (cf. Kaplan and Meier (1958), Huh et al. (2011)) to transform our weights appropriately. Let (i) denote the ordering $u^{(1)} \leq \cdots \leq u^{(N)}$. Given the weights $w_{N,i}(x)$ generated based on the naïve

assumption that $y^i = u^i$, we transform these into the weights

$$w^{\mathrm{KM}}_{N,(i)}(x) = \begin{cases} \left(\dfrac{w_{N,(i)}(x)}{\sum_{\ell=i}^{N} w_{N,(\ell)}(x)}\right) \displaystyle\prod_{k \le i-1\,:\,\delta^{(k)}=1} \left(\dfrac{\sum_{\ell=k+1}^{N} w_{N,(\ell)}(x)}{\sum_{\ell=k}^{N} w_{N,(\ell)}(x)}\right), & \text{if } \delta^{(i)} = 1, \\ 0, & \text{if } \delta^{(i)} = 0. \end{cases} \tag{2.21}$$

We next show that the transformation (2.21) preserves asymptotic optimality under certain conditions. The proof is in the E-companion.

Theorem 2.10. *Suppose that Y and V are conditionally independent given X, that Y and V share no atoms, that for every $x \in \mathcal{X}$ the upper support of V given $X = x$ is greater than the upper support of Y given $X = x$, and that costs are bounded over y for each z (i.e., $|c(z; y)| \le g(z)$). Let $w_{N,i}(x)$ be as in* (2.12), (2.13), (2.15), *or* (2.16) *and suppose the corresponding assumptions of Theorem 2.6, 2.7, 2.8, or 2.9 apply. Let $\hat{z}_N(x)$ be as in* (2.3) *but using the transformed weights* (2.21). *Then $\hat{z}_N(x)$ is asymptotically optimal.*

The assumption that Y and V share no atoms (which holds in particular if either is continuous) provides that $\delta \overset{a.s.}{=} \mathbb{I}[Y \le V]$ so that the event of censorship is observable. In applying this to problem (2.20), the assumption that Y and V are conditionally independent given X will hold if X captures at least all of the information that past stocking decisions, which are made before Y is realized, may have been based on. The assumption on bounded costs applies to problem (2.20) because the cost (negative of the objective) is bounded in $[-K_r, 0]$.

2.5.3 Data

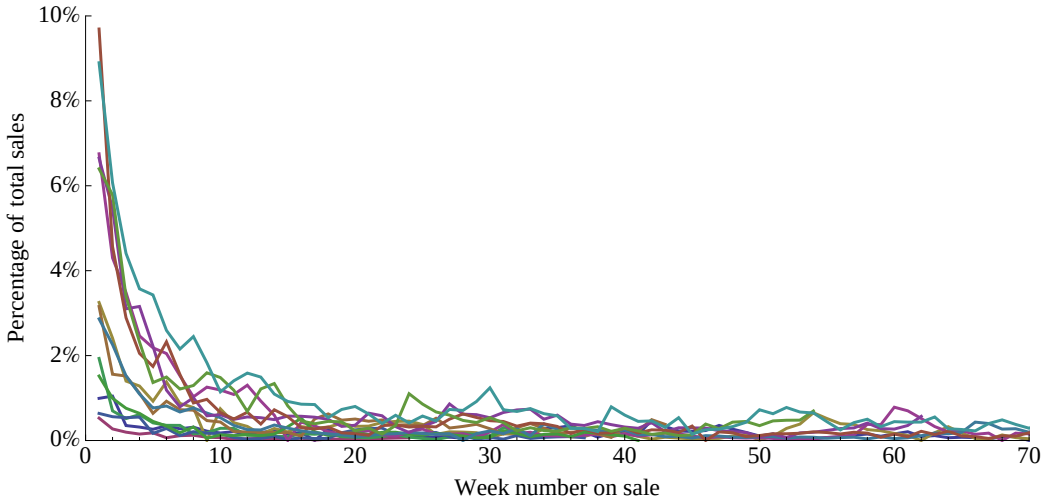
In this section, we describe the data collected. To get at the best data-driven predictive prescription, we combine both internal company data and public data harvested from online sources. The predictive power of such public data has been extensively documented in the literature (cf. Asur and Huberman (2010), Choi and Varian (2012), Goel et al. (2010), Da et al. (2011), Gruhl et al. (2005, 2004), Kallus (2014a)). Here we study its prescriptive power.

Internal Data.

The internal company data consists of 4 years of sale and inventory records across the network of retailers, information about each of the locations, and information about each of the items.

We aggregate the sales data by week (the replenishment period of interest) for each feasible combination of location and item. As discussed above, these sales-per-week data constitute a right-censored observation of weekly demand, where censorship

Figure 2-6: Percentage of All Sales in the German State of Berlin by Title



Note: Each line corresponds to one of 13 selected titles and starts from the point of release of the title to HE sales.

occurs when an item is sold out. We developed the transformed weights (2.21) to tackle this issue exactly. Figure 2-6 shows the sales life cycle of a selection of titles in terms of their marketshare when they are released to home entertainment (HE) sales and onwards. Since new releases can attract up to almost 10% of sales in their first week of release, they pose a great sales opportunity, but at the same time significant demand uncertainty.

Information about retail locations includes to which chain a location belongs and the address of the location. To parse the address and obtain a precise position of the location, including country and subdivision, we used the Google Geocoding API (Application Programming Interface).[3]

Information about items include the medium (e.g. DVD or BluRay) and an item "title." The title is a short descriptor composed by a local marketing team in charge of distribution and sales in a particular region and may often include information beyond the title of the underlying content. For example, a hypothetical film titled *The Film* sold in France may be given the item title "THE FILM DVD + LIVRET - EDITION FR", implying that the product is a French edition of the film, sold on a DVD, and accompanied by a booklet (*livret*), whereas the same film sold in Germany on BluRay may be given the item title "FILM, THE (2012) - BR SINGLE", indicating it is sold on a single BluRay disc.

Public Data: Item Metadata, Box Office, and Reviews.

We sought to collect additional data to characterize the items and how desirable they may be to consumers. For this we turned to the Internet Movie Database (IMDb;

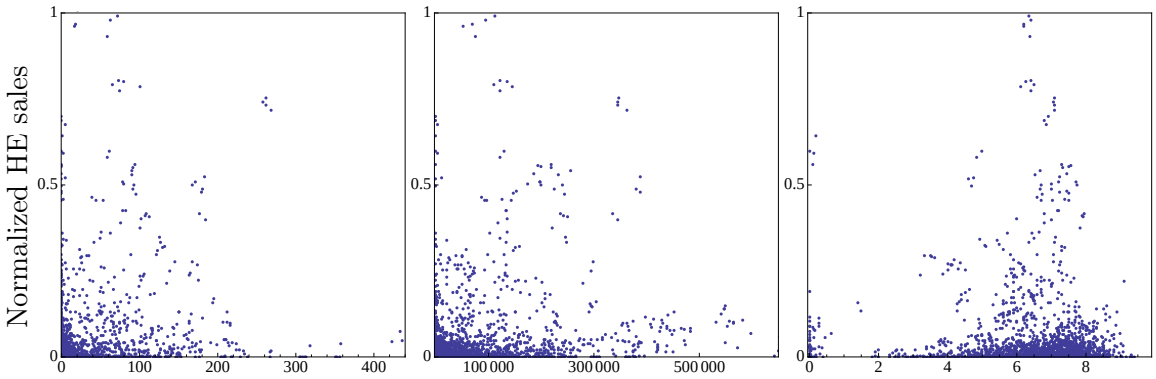
[3]See `https://developers.google.com/maps/documentation/geocoding` for details.

Figure 2-7: Screen Shots from IMDb and Rotten Tomatoes

(a) IMDb (b) Rotten Tomatoes

Note: The screen shots are of the details for 2012 Bond movie *Skyfall*. Meta-data reported includes release date, user rating, number of user rating votes, plot summary, first-billed actors, MPAA rating, and aggregate reviews.

Figure 2-8: IMDB and RT Data and Sales



(a) Box office gross ($M) $\rho = 0.32$ (b) Number IMDb votes $\rho = 0.25$ (c) Average IMDb rating $\rho = 0.02$

Note: In the scatter plots, horizontal axes are various data from IMDb and RT and the vertical axes are total European sales during first week of HE release (rescaled to anonymize). The corresponding coefficients of correlation are reported as ρ.

`www.imdb.com`) and Rotten Tomatoes (RT; `www.rottentomatoes.com`). IMDb is an online database of information on films and TV series. RT is a website that aggregates professional reviews from newspapers and online media, along with user ratings, of films and TV series. Example screen shots from IMDb and RT showing details about the 2012 movie *Skyfall* are shown in Figure 2-7.

In order to harvest information from these sources on the items being sold by the vendor, we first had to disambiguate the item entities and extract original content titles from the item titles. Having done so, we extract the following information from

IMDb:

1. type (film, TV, other/unknown);
2. US original release date of content (e.g. in theaters);
3. average IMDb user rating (0-10);
4. number of IMDb users voting on rating;
5. number of awards (e.g. Oscars for films, Emmys for TV) won and number nominated for;
6. the main actors (i.e., first-billed);
7. plot summary (30-50 words);
8. genre(s) (of 26; can be multiple); and
9. MPAA rating (e.g. PG-13, NC-17) if applicable.

And the following information from RT:

10. professional reviewers' aggregate score;
11. RT user aggregate rating;
12. number of RT users voting on rating; and
13. if item is a film, then American box office gross when shown in theaters.

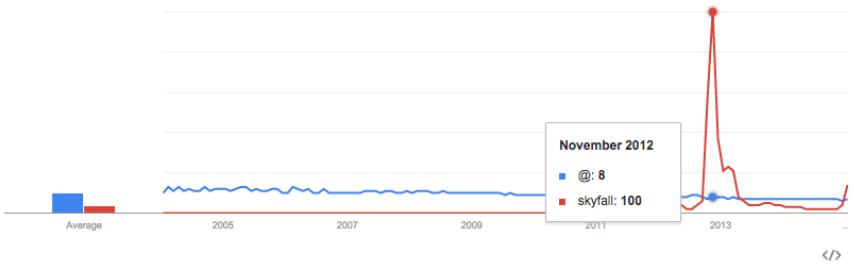
In Figure 2-8, we provide scatter plots of some of these attributes against sale figures in the first week of HE release. Notice that the number of users voting on the rating of a title is much more indicative of HE sales than the quality of a title as reported in the aggregate score of these votes.

Public Data: Search Engine Attention.

In the above, we saw that box office gross is reasonably informative about future HE sale figures. The box office gross we are able to access, however, is for the American market and is also missing for various European titles. We therefore would like additional data to quantify the attention being given to different titles and to understand the local nature of such attention. For this we turned to Google Trends (GT; `www.google.com/trends`).[4] GT provides data on the volume of Google searches for a given search term by time and geographic location. An example screen shot from GT is seen in Figure 2-9. GT does not provide absolute search volume figures, only volume time series given in terms relative to itself or to another series and in whole-number precision between 0 and 100. Therefore, to compare different such series, we

[4]While GT is available publicly online, access to massive-scale querying and week-level trends data is not public. See acknowledgements.

Figure 2-9: Screen Shot from Google Trends



Note: The screen shot displays a comparison of searches for "Skyfall" (red) and "@" (blue) in all of Germany.

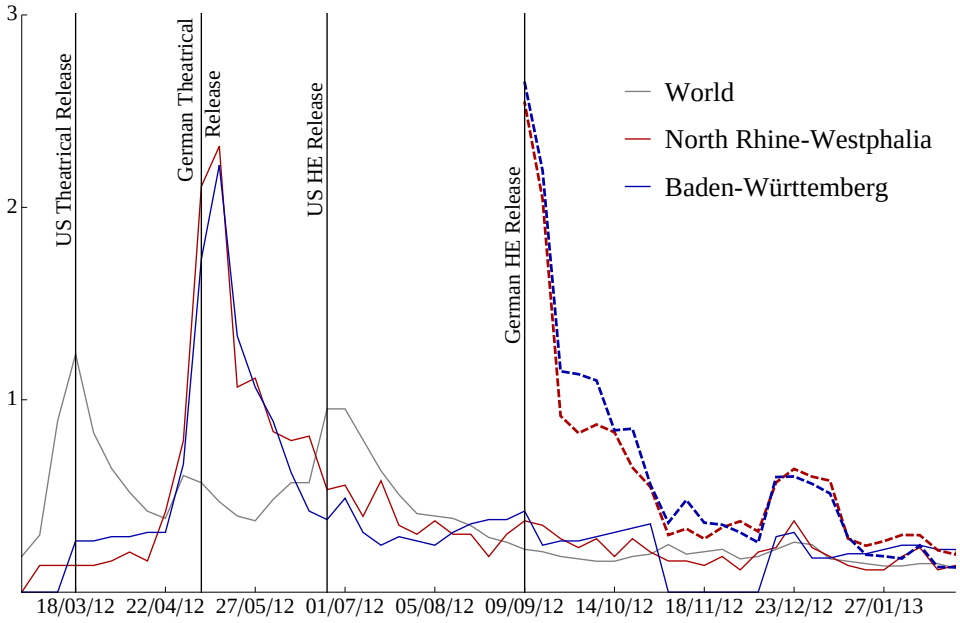
establish as a baseline the search volume for the query "@", which works well because it has a fairly constant volume across the regions of Europe and because its search volume is neither too high (else the volume for another query would be drowned out by it and reported as 0) nor too low (else the volume for another query would overwhelm it and have it be reported as 0, in which case it would be useless as a reference point).

For each title, we measure the relative Google search volume for the search term equal to the original content title in each week from 2011 to 2014 (inclusive) over the whole world, in each European country, and in each country subdivision (states in Germany, cantons in Switzerland, autonomous communities in Spain, etc.). In each such region, after normalizing against the volume of our baseline query, the measurement can be interpreted as the fraction of Google searches for the title in a given week out of all searches in the region, measured on an arbitrary but (approximately) common scale between regions.

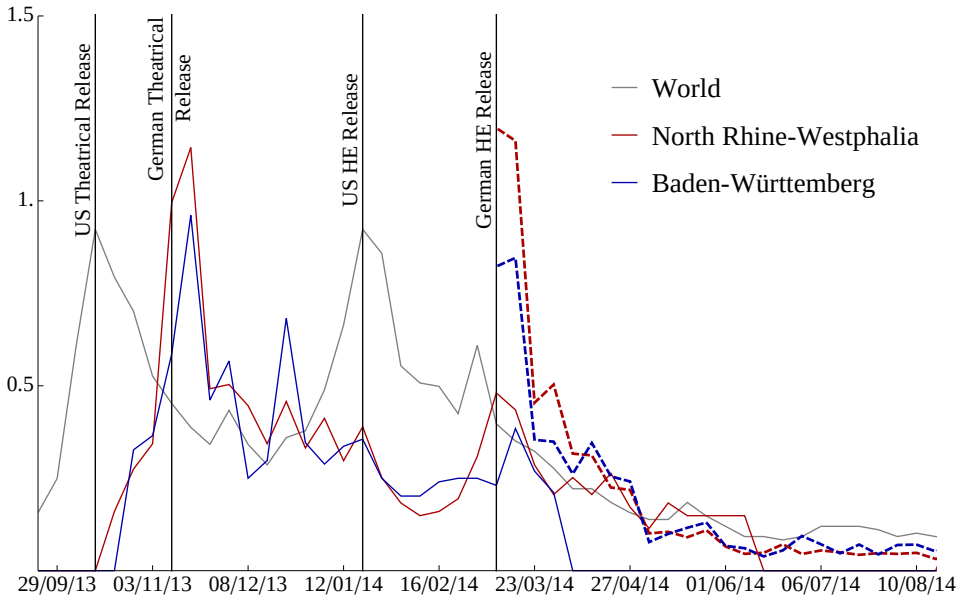
In Figure 2-10, we compare this search engine attention to sales figures in Germany for two unnamed films.[5] Comparing panel (a) and (b), we first notice that the overall scale of sales correlates with the overall scale of *local* search engine attention at the time of theatrical release, whereas the global search engine attention is less meaningful (note vertical axis scales, which are common between the two figures). Looking closer at differences between regions in panel (b), we see that, while showing in cinemas, unnamed film 2 garnered more search engine attention in North Rhine-Westphalia (NW) than in Baden-Württemberg (BW) and, correspondingly, HE sales in NW in the first weeks after HE release were greater than in BW. In panel (a), unnamed film 1 garnered similar search engine attention in both NW and BW and similar HE sales as well. In panel (b), we see that the search engine attention to unnamed film 2 in NW accelerated in advance of the HE release, which was particularly successful in NW. In panel (a), we see that a slight bump in search engine attention 3 months into

[5]These films must remain unnamed because a simple search can reveal their European distributor and hence the vendor who prefers their identity be kept confidential.

Figure 2-10: Search Engine Attention and Sales



(a) *Unnamed film 1*



(b) *Unnamed film 2*

Note: Solid lines show weekly search engine attention for the films in the world and in two populous German states. Dashed lines show weekly HE sales for the same films in the same states. Search engine attention and sales are both shown relative to corresponding overall totals in the respective region. The scales are arbitrary but common between regions and the two plots.

HE sales corresponded to a slight bump in sales. These observations suggest that local search engine attention both at the time of local theatrical release and in recent weeks may be indicative of future sales volumes.

2.5.4 Constructing Auxiliary Features and a Random Forest Prediction

For each instance (t, r) of problem (2.20) and for each item i we construct a vector of numeric predictive features x_{tri} that consist of backward cumulative sums of the sale volume of the item i at location r over the past 3 weeks (as available; e.g., none for new releases), backward cumulative sums of the total sale volume at location r over the past 3 weeks, the overall mean sale volume at location r over the past 1 year, the number of weeks since the original release data of the content (e.g., for a new release this is the length of time between the premier in theaters to release on DVD), an indicator vector for the country of the location r, an indicator vector for the identity of chain to which the location r belongs, the total search engine attention to the title i over the first two weeks of local theatrical release globally, in the country, and in the country-subdivision of the location r, backward cumulative sums of search engine attention to the title i over the past 3 weeks globally, in the country, and in the country-subdivision of the location r, and features capturing item information harvested from IMDb and RT as described below.

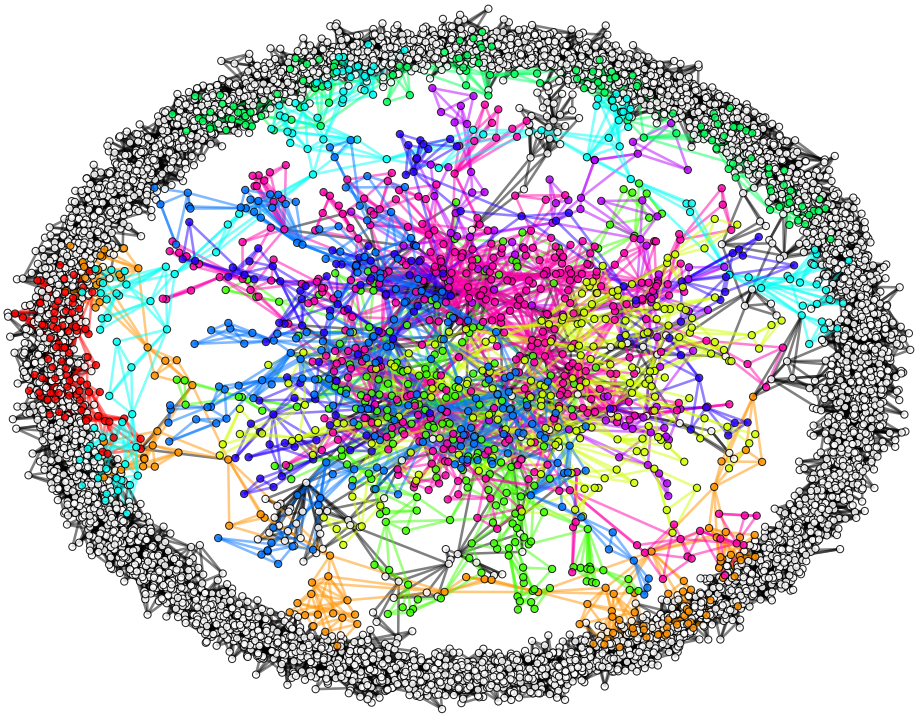
For some information harvested from IMDb and RT, the corresponding numeric feature is straightforward (e.g. number of awards). For other pieces of information, some distillation is necessary. For genre, we create an indicator vector. For MPAA rating, we create a single ordinal (from 1 for G to 5 for NC-17). For plot, we measure the cosine-similarity between plots,

$$\text{similarity}(P_1,\, P_2) = \frac{p_1^T p_2}{||p_1||\,||p_2||},$$

where p_{ki} denotes the number of times word i appears in plot text P_k and i indexes the collection of unique words appearing in plots P_1, P_2 ignoring common words like "the". We use this as a distance measure to hierarchically cluster the plots using Ward's method (cf. Ward (1963)). This captures common themes in titles. We construct 12 clusters based solely on historical data and, for new data, include a feature vector of median cosine similarity to each of the clusters. For actors, we create a graph with titles as nodes and with edges between titles that share actors, weighted by the number of actors shared. We use the method of Blondel et al. (2008) to find communities of titles and create an actor-counter vector for memberships in the 10 largest communities (see Figure 2-11). This approach is motivated by the existence of such actor groups as the "Rat Pack" (Humphery Bogart and friends), "Brat Pack" (Molly Ringwald and friends), and "Frat Pack" (Owen Wilson and friends) that often co-star in titles with a similar theme, style, and target audience.

We end up with $d_x = 91$ numeric predictive features. Having summarized these numerically, we train a RF of 500 trees to predict sales. In training the RF, we

Figure 2-11: The Graph of Actors



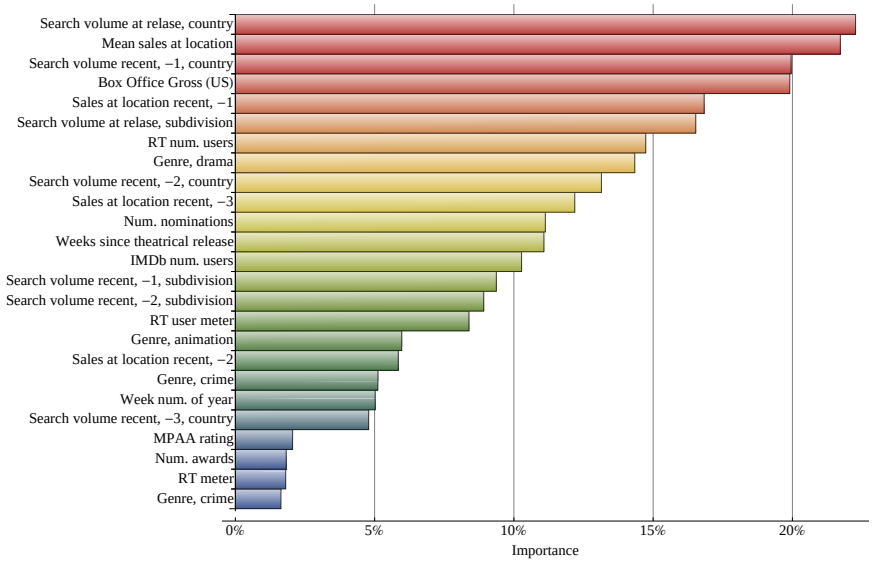
Note: Actors are connected via common movies where both are first-billed. Colored nodes correspond to the 10 largest communities of actors. Colored edges correspond to intra-community edges.

normalize each the sales in each instance by the training-set average sales in the corresponding location; we de-normalize after predicting. To capture the decay in demand from time of release in stores, we train a separate RFs for sale volume on the k^{th} week on the shelf for $k = 1, \ldots, 35$ and another RF for the "steady state" weekly sale volume after 35 weeks.

For $k = 1$, we are predicting the demand for a new release, the uncertainty of which, as discussed in Section 2.5.1, constitutes one of the greatest difficulties of the problem to the company. In terms of predictive quality, when measuring out-of-sample performance we obtain an $R^2 = 0.67$ for predicting sale volume for new releases. The 25 most important features in this prediction are given in Figure 2-12. In Figure 2-13, we show the R^2 obtained also for predictions at later times in the product life cycle, compared to the performance of a baseline heuristic that always predicts for next week the demand of last week.

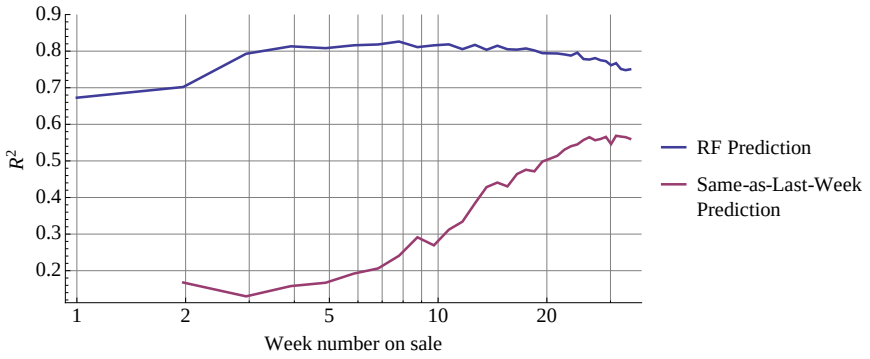
Considering the uncertainty associated with new releases, we feel that this is a positive result, but at the same time what truly matters is the performance of the prescription in the problem. We discuss this next.

Figure 2-12: The 25 Top Variables in Predictive Importance



Note: Predictive importance is measured as the average over forest trees of the change in mean-squared error of the tree (shown here as percentage of total variance) when the value of the variables is randomly permuted among the out-of-bag training data.

Figure 2-13: Out-of-Sample R^2 for Predicting Demand at Different Stages of Product Life

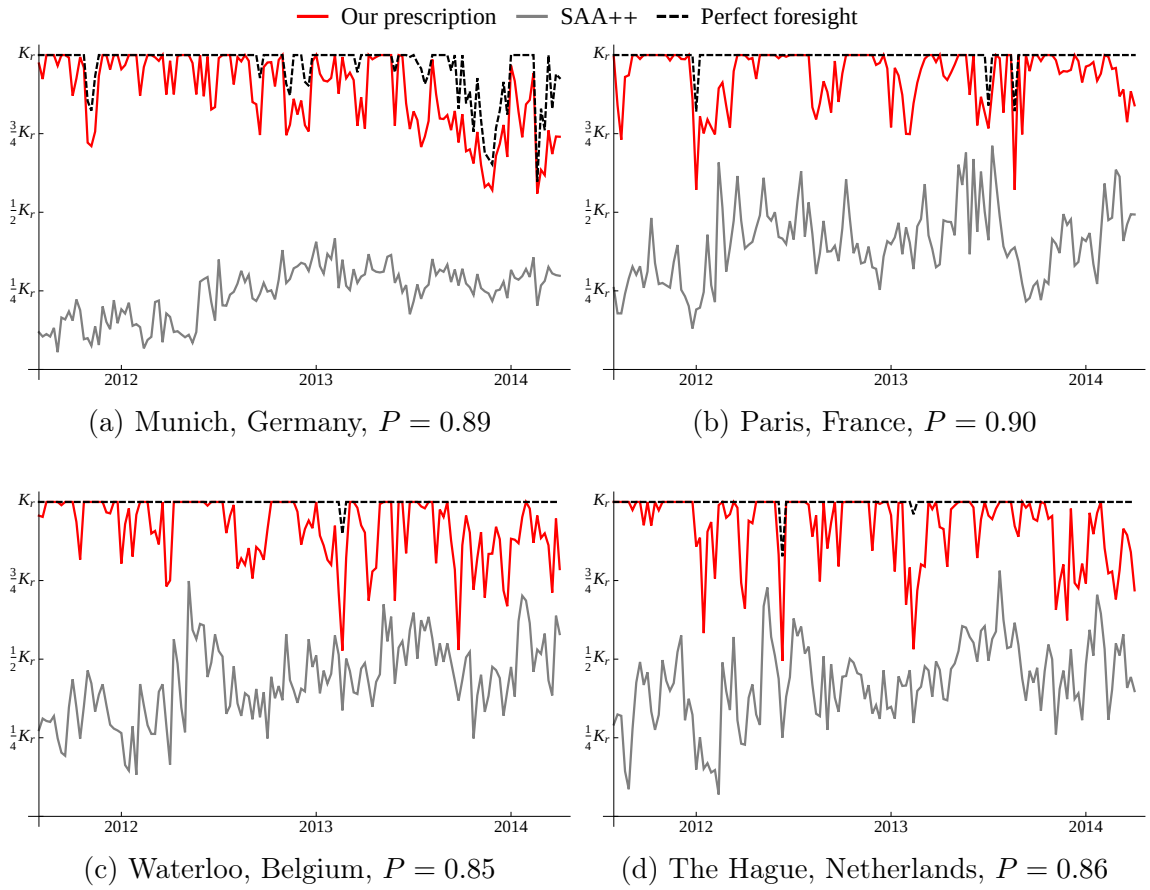


2.5.5 Applying Our Predictive Prescriptions to the Problem

In the last section we discussed how we construct RFs to predict sales, but our problem of interest is to prescribe order quantities. To solve our problem (2.20), we use the trees in the forests we trained to construct weights $w_{N,i}(x)$ exactly as in (2.18), then we transform these as in (2.21), and finally we prescribe data-driven order quantities $\hat{z}_N(x)$ as in (2.8). Thus, we use our data to go from an observation $X = x$ of our varied auxiliary data directly to a replenishment decision on order quantities.

We would like to test how well our prescription does out-of-sample and as an

Figure 2-14: The Performance of Our Prescription Over Time



(a) Munich, Germany, $P = 0.89$

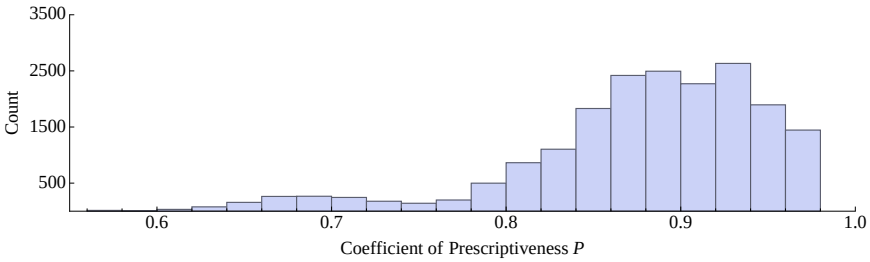
(b) Paris, France, $P = 0.90$

(c) Waterloo, Belgium, $P = 0.85$

(d) The Hague, Netherlands, $P = 0.86$

Note: The vertical axis is shown in terms of the location's capacity, K_r.

Figure 2-15: The Distribution of Coefficients of Prescriptiveness P over Retail Locations



actual live policy. To do this we consider what we would have done over the 150 weeks from December 19, 2011 to November 9, 2014 (inclusive). At each week, we consider only data from time prior to that week, train our RFs on this data, and apply our prescription to the current week. Then we observe what had actually materialized and score our performance.

There is one issue with this approach to scoring: our historical data only consists

of sales, not demand. While we corrected for the adverse effect of demand censorship on our prescriptions using the transformation (2.21), we are still left with censored demand when scoring performance as described above. In order to have a reasonable measure of how good our method is, we therefore consider the problem (2.20) with capacities K_r that are a *quarter* of their nominal values. In this way, demand censorship hardly ever becomes an issue in the scoring of performance. To be clear, this correction is necessary just for a counterfactual scoring of performance; not in practice. The transformation (2.21) already corrects for prescriptions trained on censored observations of the quantity Y that affects true costs.

We compare the performance of our method with two other quantities. One is the performance of the perfect-forecast policy, which knows future demand exactly (no distributions). Another is the performance of a data-driven policy without access to the auxiliary data (i.e., SAA). Because the decay of demand over the lifetime of a product is significant, to make it a fair comparison we let this policy depend on the distributions of product demand based on how long its been on the market. That is, it is based on T separate datasets where each consists of the demands for a product after t weeks on the market (again, considering only past data). Due to this handicap we term it SAA++ henceforth.

The ratio of the difference between our performance and that of the prescient policy and the difference between the performance of SAA++ and that of the prescient policy is the coefficient of prescriptiveness P. When measured out-of-sample over the 150-week period as these policies make live decisions, we get $P = 0.88$. Said another way, in terms of our objective (sales volumes), our data X and our prescription $\hat{z}_N(x)$ gets us 88% of the way from the best data-poor decision to the impossible perfect-foresight decision. This is averaged over just under 20,000 locations. In Figure 2-14, we plot the performance over time at four specific locations, the city of which is noted. In Figure 2-15, we plot the overall distribution of coefficients of prescriptiveness P over all retail locations in Europe.

2.6 Alternative Approaches

In the beginning of Section 2.2, we noted that the empirical distribution is insufficient for approximating the full-information problem (2.2). The solution was to consider local neighborhoods in approximating conditional expected costs; these were computed separately for each x. Another approach would be to develop an explicit decision rule and impose structure on it. In this section, we consider an approach to constructing a predictive prescription by selecting from a family of linear functions restricted in some norm,

$$\mathcal{F} = \left\{ z(x) = Wx \ : \ W \in \mathbb{R}^{d_z \times d_x}, \, ||W|| \leq R \right\}, \tag{2.22}$$

so to minimize the empirical marginal expected costs as in (2.4),

$$\hat{z}_N(\cdot) \in \arg\min_{z(\cdot) \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} c(z(x^i); y^i).$$

The linear decision rule can be generalized by transforming X to include nonlinear terms.

We consider two examples of a norm on the matrix of linear coefficients, W. One is the row-wise p, p'-norm:

$$||W|| = \left|\left|\left(\gamma_1 \left|\left|W_1\right|\right|_p, \ldots, \gamma_d \left|\left|W_d\right|\right|_p\right)\right|\right|_{p'}.$$

Another is the Schatten p-norm:

$$||W|| = \left|\left|\left(\tau_1, \ldots, \tau_{\min\{d_z, d_x\}}\right)\right|\right|_p \quad \text{where } \tau_i \text{ are } W\text{'s singular values.}$$

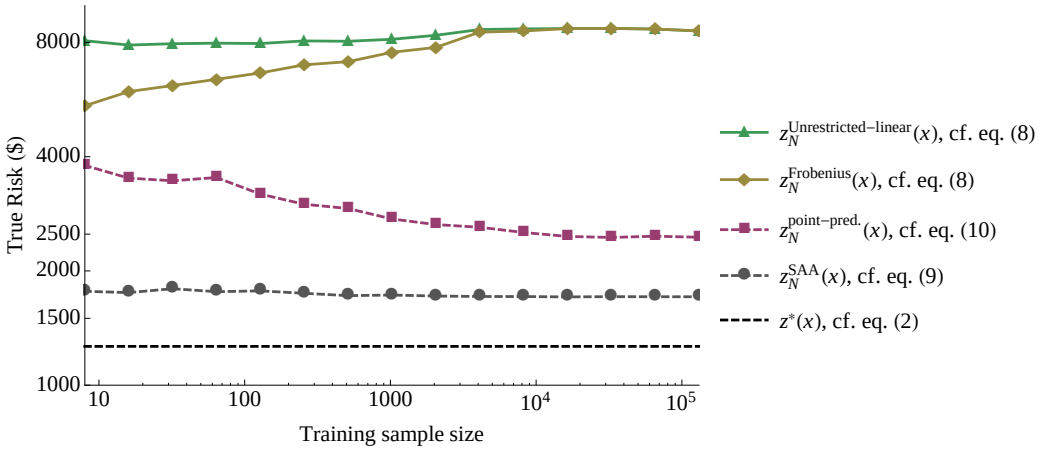
For example, the Schatten 1-norm is the matrix nuclear norm. In either case, the restriction on the norm is equivalent to an appropriately-weighted regularization term incorporated into the objectives of (2.4).

Problem (2.4) corresponds to the traditional framework of empirical risk minimization in statistical learning with a general loss function. There is no great novelty in this formulation except for the potential multivariateness of Y and z. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z; y) = (z - y)^2$, problem (2.4) corresponds to least-squares regression. For $d_z = d_y = 1$, $\mathcal{Z} = \mathbb{R}$, and $c(z; y) = (y - z)(\tau - \mathbb{I}[y - z < 0])$, problem (2.4) corresponds to quantile regression, which estimates the conditional τ-quantile as a function of x. Rearranging terms, $c(z; y) = (y - z)(\tau - \mathbb{I}[y - z < 0]) = \max\{(1 - \tau)(z - y), \tau(y - z)\}$ is the same as the newsvendor cost function where τ is the service level requirement. Rudin and Vahn (2014) consider this cost function and the selection of a linear decision rule with regularization on ℓ_1-norm in order to solve a newsvendor problem with auxiliary data. Quantile regression (cf. Koenker (2005)) and ℓ_1-regularized quantile regression (cf. Belloni and Chernozhukov (2011)) are standard techniques in regression analysis. Because most OR/MS problems involve multivariate uncertainty and decisions, in this section we generalize the approach and its associated theoretical guarantees to such multivariate problems ($d_y \geq 1$, $d_z \geq 1$).

Before continuing, we note a few limitations of any approach based on (2.4). For general problems, there is no reason to expect that optimal solutions will have a linear structure (whereas certain distributional assumptions lead to such conclusions in least-squares and quantile regression analyses). In particular, unlike the predictive prescriptions studied in Section 2.2, the approach based on (2.4) does not enjoy the same universal guarantees of asymptotic optimality. Instead, we will only have out-of-sample guarantees that depend on our class $\mathcal{F}$ of decision rules.

Another limitation is the difficulty in restricting the decisions to a constrained feasible set $\mathcal{Z} \neq \mathbb{R}^{d_z}$. Consider, for example, the portfolio allocation problem from Section 2.1.1, where we must have $\sum_{i=1}^{d_x} z_i = 1$. One approach to applying (2.4) to this problem might be to set $c(z; y) = \infty$ for $z \notin \mathcal{Z}$ (or, equivalently, constrain $z(x^i) \in \mathcal{Z}$ $\forall i$). However, not only will this not guarantee that $z(x) \in \mathcal{Z}$ for x outside the dataset, but we would also run into a problem of infeasibility as we would have N linear equality constraints on $d_z \times d_x$ linear coefficients (a constraint such as $\sum_{i=1}^{d_x} z_i \leq 1$ that does not reduce the affine dimension will still lead to an undesirably flat linear decision rule as N grows). Another approach may be to compose $\mathcal{F}$ with a

Figure 2-16: Performance of ERM Prescriptions in the Shipment Planning Example.



Note: The horizontal and vertical axes are on log scales.

projection onto $\mathcal{Z}$, but this will generally lead to a non-convex optimization problem that is intractable to solve. Therefore, the approach is limited in its applicability to OR/MS problems.

In a few limited cases, we may be able to sensibly extend the cost function synthetically outside the feasible region while maintaining convexity. For example, in the shipment planning example of Section 2.1.1, we may allow negative order quantities z and extend the first-stage costs to depend only on the positive part of z, i.e. $p_1 \sum_{i=1}^{d_z} \max\{z_i, 0\}$ (but leave the second-stage costs as they are for convexity). Now, if after training $\hat{z}_N(\cdot)$, we transform any resulting decision by only taking the positive part of each order quantity, we end up with a feasible decision rule whose costs are no worse than the synthetic costs of the original rule. If we follow this approach and apply (2.4), either without restrictions on norms or with a diminishing Frobenius norm penalty on coefficients, we end up with results as shown in Figure 2-16. The results suggest that, while we are able to apply the approach to the problem, restricting to linear decision rules is inefficient in this particular problem.

In the rest of this section we consider the application of the approach (2.4) to problems where y and z are multivariate and $c(z; y)$ is general, but only treat unconstrained decisions $\mathcal{Z} = \mathbb{R}^{d_z}$.

2.6.1 Tractability

We first develop sufficient conditions for the problem (2.4) to be optimized in polynomial time. The proof is in the E-companion.

Theorem 2.11. *Suppose that $c(z; y)$ is convex in z for every fixed y and let oracles be given for evaluation and subgradient in z. Then for any fixed x we can find an ϵ-optimal solution to* (2.4) *in time and oracle calls polynomial in n, d, $\log(1/\epsilon)$ for $\mathcal{F}$ as in* (2.22).

2.6.2 Out-of-Sample Guarantees

Next, we characterize the out-of-sample guarantees of a predictive prescription derived from (2.4). All proofs are in the E-companion. In the traditional framework of empirical risk minimization in statistical learning such guarantees are often derived using Rademacher complexity but these only apply to univariate problems (c.f. Bartlett and Mendelson (2003)). Because most OR/MS problems are multivariate, we generalize this theory appropriately. We begin by generalizing the definition of Rademacher complexity to multivariate-valued functions.

Definition 2.12. Given a sample $S_N = \{s_1, \ldots, s_N\}$, The *empirical multivariate Rademacher complexity* of a class of functions $\mathcal{F}$ taking values in $\mathbb{R}^d$ is defined as

$$\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N) = \mathbb{E}\left[\frac{2}{N}\sup_{g\in\mathcal{F}}\sum_{i=1}^{n}\sum_{k=1}^{d}\sigma_{ik}g_k(s_i)\Big|s_1, \ldots, s_n\right]$$

where σ_{ik} are independently equiprobably $+1, -1$. The *marginal multivariate Rademacher complexity* is defined as

$$\mathfrak{R}_N(\mathcal{F}) = \mathbb{E}\left[\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N)\right]$$

over the sampling distribution of S_N.

Note that given only data S_N, the quantity $\widehat{\mathfrak{R}}_N(\mathcal{F}; S_N)$ is observable. Note also that when $d = 1$ the above definition coincides with the common definition of Rademacher complexity.

The theorem below relates the multivariate Rademacher complexity of $\mathcal{F}$ to out-of-sample guarantees on the performance of the corresponding predictive prescription $\hat{z}_N(x)$ from (2.4). A generalization of the following to mixing processes is given in the supplemental Section A.2. We denote by $S_N^x = \left\{x^1, \ldots, x^N\right\}$ the restriction of our sample to data on X.

Theorem 2.13. *Suppose $c(z; y)$ is bounded and equi-Lipschitz in z:*

$$\sup_{z\in\mathcal{Z},\, y\in\mathcal{Y}} \quad c(z; y) \leq \overline{c},$$
$$\sup_{z\neq z'\in\mathcal{Z},\, y\in\mathcal{Y}} \quad \frac{c(z; y) - c(z'; y)}{||z_k - z'_k||_\infty} \leq L < \infty.$$

Then, for any $\delta > 0$, we have that with probability $1 - \delta$,

$$\mathbb{E}\left[c(z(X); Y)\right] \leq \frac{1}{N}\sum_{i=1}^{N} c(z(x^i); y^i) + \overline{c}\sqrt{\log(1/\delta')/2N} + L\mathfrak{R}_N(\mathcal{F}) \quad \forall z \in \mathcal{F}, \quad (2.23)$$

and that, again, with probability $1-\delta$*,*

$$\mathbb{E}\left[c(z(X);Y)\right] \leq \frac{1}{N}\sum_{i=1}^{N} c(z(x^i);y^i) + 3\overline{c}\sqrt{\log(2/\delta'')/2N} + L\widehat{\mathfrak{R}}_N(\mathcal{F};S_N^x) \quad \forall z \in \mathcal{F}. \tag{2.24}$$

In particular, these hold for $z = \hat{z}_N(\cdot) \in \mathcal{F}$*.*

Equations (2.23) and (2.24) provide a bound on the out-of-sample performance of any predictive prescription $z(\cdot) \in \mathcal{F}$. The bound is exactly what we minimize in problem (2.4) because the extra terms do not depend on $z(\cdot)$. That is, we minimize the empirical risk, which, with additional confidence terms, bounds the true out-of-sample costs of the resulting predictive prescription $\hat{z}_N(\cdot)$.

These confidence terms involve the multivariate Rademacher complexity of our class $\mathcal{F}$ of decision rules. In the next lemmas, we compute appropriate bounds on the complexity of our examples of classes $\mathcal{F}$. The theory, however, applies beyond linear rules.

Lemma 2.14. *Consider* $\mathcal{F}$ *as in* (2.22) *with row-wise* p, p' *norm for* $p \in [2,\infty)$ *and* $p' \in [1,\infty]$*. Let* q *be the conjugate exponent of* p *(*$1/p + 1/q = 1$*) and suppose that* $||x||_q \leq M$ *for all* $x \in \mathcal{X}$*. Then*

$$\mathfrak{R}_N(\mathcal{F}) \leq 2MR\sqrt{\frac{p-1}{N}}\sum_{k=1}^{d_z}\frac{1}{\gamma_k}.$$

Lemma 2.15. *Consider* $\mathcal{F}$ *as in* (2.22) *with Schatten p-norm. Then*

$$\widehat{\mathfrak{R}}_N(\mathcal{F};S_N^x) \leq 2Rd_z^r\sqrt{\frac{1}{N}}\sqrt{\widehat{\mathbb{E}}_{S_N^x}||X||_2^2}$$
$$\mathfrak{R}_N(\mathcal{F}) \leq 2Rd_z^r\sqrt{\frac{1}{N}}\sqrt{\mathbb{E}||X||_2^2},$$

where $r = \max\{1-1/p,\, 1/2\}$ *and* $\widehat{\mathbb{E}}_{S_N^x}$ *denotes empirical average over the sample* S_N^x*.*

The above results indicate that the confidence terms in equations (2.23) and (2.24) shrink to 0 as $N \to \infty$ even if we slowly relax norm restrictions. Hence, we can approach the optimal out-of-sample performance over the class $\mathcal{F}$ without restrictions on norms.

2.7 Conclusions

In this chapter, we combined ideas from ML and OR/MS in developing a framework, along with specific methods, for using data to prescribe optimal decisions in OR/MS problems that leverage auxiliary observations. We motivate our methods based on

existing predictive methodology from ML, but, in the OR/MS tradition, focus on the making of a decision and on the effect on costs, revenues, and risk. Our approach is

a) generally applicable,

b) tractable,

c) asymptotically optimal,

d) and leads to substantive and measurable improvements in a real-world context.

We feel that the above qualities, together with the growing availability of data and in particular auxiliary data in OR/MS applications, afford our proposed approach a potential for substantial impact in the practice of OR/MS.

Chapter 3

Prediction vs Prescription in Data-Driven Pricing

Pricing in revenue management is based on one's understanding of consumers' response to price changes. This, in turn, is often based on analytics of historical price-demand data. We discuss the distinction between predictive and prescriptive approaches to data-driven pricing. Through examples both synthetic and real, we show that a naive but common predictive approach can leave money on the table whereas a prescriptive approach is theoretically sound and performs well in practice. Extending recent work, we develop a statistical hypothesis test for revenue optimality of a particular pricing approach that works with observational data. Applying this test to data from an automotive loan provider, we demonstrate that predictive approaches clearly miss the mark in practical applications, looking only to actual revenues generated at the end of the day rather than model soundness. On the other hand, parametric approaches to pricing often suffice, but only when they take into full account the prescriptive nature of the problem.

3.1 Introduction

Pricing is one of the most fundamental instruments for revenue management. Hence, effective pricing hinges on the manager's understanding of consumers' response to price changes (Phillips 2005). In pricing applications, this response is estimated from observations of past sale attempts. This can be done through repeated experiments (as in Bertsimas and Perakis (2006), Besbes and Zeevi (2009), Harrison et al. (2012)), but in many real-world applications this is done based on analytics of a corpus of historical observational data (examples include Besbes et al. (2010), Cohen et al. (2014), Johnson et al. (2014)) – for lack of a better term, we call this *data-driven pricing*. Since prices are usually set at least somewhat strategically, an important concern, which we have found is not fully and consciously addressed in data-driven pricing theory and applications, is the distinction between *prediction* and *prescription*. In prediction, the analyst is an external observer that simply wants to predict as best as possible an outcome (such as demand) based on partial observations (such as price).

In prescription, on the other hand, the analyst is a manager that seeks herself to set a control (such as price) in order to optimize an objective (such as maximum revenue) that depends on a response to the control (such as demand). Ignoring the fact that such a response in the latter is distinct from a predicted outcome in the former can lead to suboptimal revenues if price is optimized on the basis predicted demand.

In this chapter, we explore this problem through the lens of a fundamental building block of data-driven pricing: the choice of a single price for a single product in a single sale attempt based on historical observations of the outcomes of past sale attempts. The implications of the work extend to more complicated multi-product and inventory constrained problem as well as more sophisticated customized schemes.

We consider pricing based on data that is observational, i.e. it is not the result of controlled experiments on consumers' response to prices but rather consists of observations of prices chosen historically and associated demand. This is by far the most common situation in every practical pricing application. Observed prices are usually set strategically, in consideration and anticipation of future demand, rather than randomly as in an experiment. The distinction between a predicted outcome and the causal effect of a control on an outcome is common (Spirtes 2010) even in econometric supply-demand-price analyses (Phillips et al. 2012, Berry et al. 1995, Bijmolt et al. 2005).

Here, however, we explore the ramifications of the dichotomy specifically for a *prescriptive* problem such as price optimization, where continuous controls are optimized for maximum effect. Therefore, we review specific examples of data-driven pricing and explore how this issue can substantively affect revenues negatively. Taking a step back, we study in generality the theoretical issue of identifiability of optimal prices from historical data – when is there no hope and under what conditions there is. We provide solutions to the pricing problem based on observational data by drawing on the literatures of nonparametric estimation and causal inference. Specifically, we provide a nonparametric method for price optimization with guarantees of asymptotic optimality, but, since large amounts of data may be necessary before these asymptotic kick in, we also provide a parametric method based on a generalization of the propensity score to continuous interventions. Recognizing that in a prescriptive problem fitting a complete price model is not the objective – optimizing revenue is – we develop a hypothesis test for asymptotic revenue optimality of a given pricing strategy based on observational data, extending recent work (Besbes et al. 2010). In fact, we apply this new test to study a real-world auto-loan dataset studied in Besbes et al. (2010) and show that both the parametric and nonparametric pricing strategies used therein lead to revenues that are statistically distinguishable as suboptimal, whereas our new parametric pricing strategy yields revenues that cannot be distinguished from optimal.

The purpose of this chapter is first and foremost to highlight a common issue with data-driven pricing applications that can have negative ramifications in real practice and to provide a framework in which to understand the limits and capacities of the data available. We provide pricing strategies apt for observational data and extend revenue-optimality testing methodologies tailored for prescriptive problems to the new setting of observational data. We provide both synthetic and real examples.

3.2 The Problem

Let us first describe precisely the basic pricing problem we consider without data – that is, in terms of hypothetical primitives that are in practice unknown. Consider choosing a unit price $p \in \mathcal{P} \subset \mathbb{R}_+$ at which to offer a product in one sale event (e.g. the price for a supermarket good for the week or the price offered at one point to one online customer). Let the per-unit revenue netted when selling at price p be $r(p) = p - c$ where c is the per-unit production or procurement cost. Let us denote by the random variable $D(p) \geq 0$ the potential stochastic demand that the product offered at price p would garner. This demand can be nonnegative continuous, nonnegative integral, or binary. A sale event is an instance of the stochastic process $\{D(p) : p \in \mathcal{P}\}$ since it encapsulates all the relevant (but unknown) information about a particular opportunity for sale. A sale event may or may not also have some other identifying or idiosyncratic information. We call the function $\mathbb{E}[D(p)]$ the *price response function* (PRF). The hypothetical price optimization problem we would then like to solve, had we full information on the unknown PRF, can be expressed as follows:

$$p^* \in \arg\max_{p \in \mathcal{P}} \{R(p) := r(p)\mathbb{E}[D(p)]\}. \tag{3.1}$$

Now we consider the corresponding data-driven pricing problem. Instead of having full knowledge of the PRF, we have observations from past n sale events. For each of the past sale events, $i = 1, \ldots, n$, we know the price offered $P_i \in \mathcal{P}$ and the demand seen $D_i = D_i(P_i)$. We may also have some additional information about each event. The standing assumptions is that past prices and sale events are independently and identically drawn (iid) from some stationary joint distribution, a generic draw from which we will denote without subscripts (e.g. P, D). Note that historical price P is a random variable and hence conditional expectations can be defined, which is in contrast to the full-information setting where p is a control variable. The problem of interest is to set a data-driven pricing strategy $\hat{p}_n$ so to achieve high revenue as measured by $R(p)$, the (unknown) revenue objective of (3.1).

In contrast, some work has focused on data-driven pricing strategies that attempt to solve a different problem than (3.1) that in our notation can be expressed as follows:

$$\tilde{p} \in \arg\max_{p \in \mathcal{P}} \left\{ \tilde{R}(p) := r(p)\mathbb{E}\left[D \middle| P = p\right] \right\}. \tag{3.2}$$

Such data-driven strategies, which we term *predictive* approaches, estimate the conditional expectation of demand given price $\mathbb{E}\left[D\middle|P = p\right]$, i.e. the best *prediction* for what demand is in a random instance of (P, D) where P is reveled, and plug in the estimate into (3.2).

The questions we wish to address are: what is the differences between (3.1) and (3.2), when can we even solve (3.1), how can we solve it when we can, and how we know if all of this actually matters in practice.

3.3 Prediction, Causation, and Prescription

Let us delineate three different problems that may be associated with price data. In the prediction problem, we draw some price-demand pair where price is known but demand is not and we want to have the best guess for this hidden value of demand.

> **The Prediction Problem in Pricing:** For a sale event and price drawn from a stationary distribution with demand hidden, predict the hidden value of demand with least error.

With perfect knowledge of distributions and if error is defined as squared difference, the optimal solution to this problem is the conditional expectation $\mathbb{E}\left[D\middle|P=p\right]$. With only data, we would solve this problem by fitting a linear, other parametric, or non-parametric regression. For example, the non-parametric Nadaraya-Watson kernel regression used in Besbes et al. (2010) as an estimator for the true price-response function is a universally consistent estimator of the conditional expectation under very mild conditions.

Prediction provides a best guess for a missing value given the value of a related observation, but it does not give us the effect of intervening and setting this value at a particular level. In causal estimation in the context of pricing, we are interested in the value of demand if we were to intervene and set the price at a particular level.

> **The Causal Estimation Problem in Pricing:** For a sale event drawn from a stationary distribution and for a fixed price, predict with least error the value of demand were the price to be set as given.

With perfect knowledge of distributions and if error is defined as squared difference, the optimal solution to this problem is the expectation $\mathbb{E}\left[D(p)\right]$. Describing the potential demand as $D(p)$ is in fact the Neyman-Rubin potential outcome notation (cf. Sekhon (2008)). The Neyman-Rubin framework is generally applied to binary interventions (control vs treatment), but here the "interventions" are prices that are potentially continuous. This distinction between the problems of prediction and causal estimation is common (cf. Spirtes (2010)).

In price optimization, however, we are not directly interested in either prediction or causal estimation. Instead, our primary interest is in netting as high revenues as possible. Thus, the problem of interest is to prescribe a price to achieve this objective.

> **The Prescription Problem in Pricing:** For a sale event drawn from a stationary distribution, set the price so to maximize the expected total revenues that would be netted at this price.

With perfect knowledge of distributions, the optimal solution to this problem is as in (3.1). Note that the solution p^* is determined by the solution $\mathbb{E}\left[D(p)\right]$ to the causal

estimation problem via optimization. Where the problems differ is in their objective and hence in how we evaluate a *data-driven* solution to either. In causal estimation, as in prediction, model accuracy is of primary interest. In prescription, however, it is revenue netted that is of primary interest. Thus, if a particular pricing strategy produces revenues that cannot be statistically distinguished from optimal revenues then we are content with it, even if a demand model that underlies this strategy can be distinguished as invalid. This is similar to the distinction made in Besbes et al. (2010) but here, in the context of observational data, we are concerned with prescribing prices that optimize the revenue we net when we *set* the price to p, and not with identifying the price at which the conditional expectation of revenues $\tilde{R}(p)$ is highest. However, if such a pricing strategy $\tilde{p}$ were to achieve revenues that are statistically indistinguishable from optimal then we would be content – we develop a statistical test for such a scenario.

Examples

To put the notions described above in context, let us consider some examples, both synthetic and real.

The following synthetic example shows that, in general, revenues generated by a predictive approach can be far lower than revenues generated by a prescriptive approach.

Example 3.1 (Artificial Continuous Example)**.** Suppose the unit procurement price is $c = 50$ and that we are interested in setting a price p in $\mathcal{P} = [50, 300]$. Suppose the demand process is such that

$$D(p) = 15\,(300 - p)_+ + 1500pXe^{-(p-50)X/10} + \epsilon(p),$$

where $X \sim \text{Exp}(1)$ is an exponentially distributed random disturbance and $\epsilon(p)$ is a Gaussian process with kernel $\text{Cov}\,(\epsilon(p), \epsilon(p')) = \sigma^2\mathbb{I}\,[p = p']$. Suppose, moreover, that historically prices have been set as $P = 50 + 10Y/X$ where $Y \sim \text{Exp}(1)$ is an exponentially distributed random disturbance.

Let us first consider the problem of demand modeling. The answer to the predictive problem is

$$\mathbb{E}\left[D\middle|P = p\right] = 15\,(300 - p)_+ + \frac{3750p}{p - 50},$$

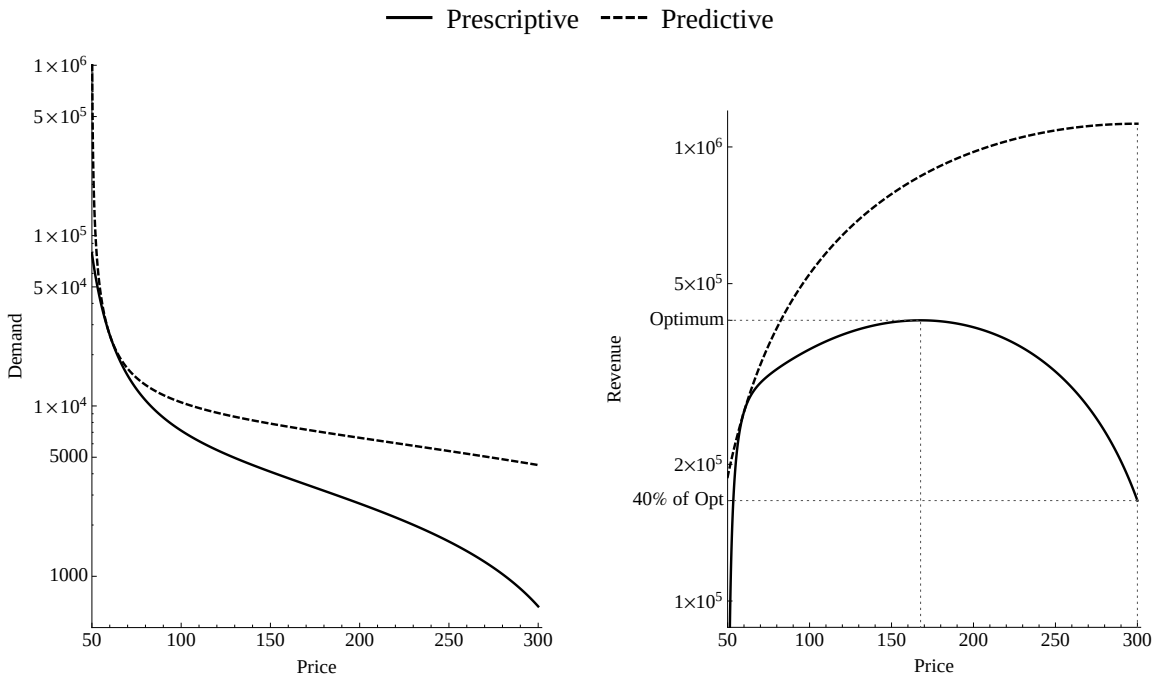
whereas the answer to the causal estimation problem is

$$\mathbb{E}\,[D(p)] = 15(300 - p)_+ + \frac{150000p}{(p - 40)^2}.$$

We plot these two in Figure 3-1(a).

Now consider the prescription problem for the optimal price. The true profit function, $R(p) = (p - c)\mathbb{E}\,[D(p)]$, is optimized at $p^* = 167.72$. On the other hand, a predictive approach would suggest that we optimize $\tilde{R}(p) = (p - c)\mathbb{E}\left[D\middle|P = p\right]$, leading to the price $\tilde{p} = 300$. This may seem very far from p^*, but the correct metric

Figure 3-1: Prediction vs Prescription in Example 3.1



(a) Demand (b) Revenue

Note: Solid lines show the true demand and revenue curves and dashed lines show the spurious ones that would arise from an invalid predictive analysis, which would lead to a 60% loss in revenues compared to the true, prescriptive optimum.

for evaluating a pricing strategy in the prescription problem is via the objective of revenues. It is that the revenue under $\tilde{p}$ is 60% less than the revenue under p^* that suggests that $\tilde{p}$ is not a good pricing strategy. We plot R, $\tilde{R}$, p^*, and $\tilde{p}$ in Figure 3-1(b).

Example 3.2 (Auto Loan Rate Optimization)**.** In Besbes et al. (2010), the authors study the problem of prescribing interest rates for automobile loans based on historical, observational data. The on-line auto lending data, provided by Columbia University Center for Pricing and Revenue Management (2012), consists of past sale events where a customer fills out a loan application, if approved an interest rate is quoted (price), and the customer either accepts or rejects the loan (binary demand). The authors apply a predictive approach and estimate $\mathbb{E}[D|P = p]$ using either Nadaraya-Watson kernel regression or logistic regression. The authors study the differences – and how to test for them – between pricing strategies generated by optimizing based on either regression since one is model-dependent (logistic regression) and the other is not (kernel regression). Both approaches, however, address the prediction problem and start by estimating $\mathbb{E}[D|P = p]$.

We will return to this example twice more to discuss further the relationship to

a prescription problem and then to conduct an empirical study of the distinction between prediction and prescription and whether it has any practical relevance to loss of revenue.

Example 3.3 (Other Examples from the Literature)**.** In Cohen et al. (2014), the authors study a data-driven multi-period dynamic pricing problem for supermarket goods based on historical, observational data. "To predict demand as a function of prices", the authors "estimate a log-log demand model" and report that "estimated demand models are accurate in the sense of having low forecast error," where recent prices and week index are included as explanatory variables besides the price to be set.

In Johnson et al. (2014), the authors study a data-driven pricing problem for online sales of apparel based on historical, observational data. Using tree regression, the authors "formulate a price optimization model to maximize revenue from first exposure styles, using demand predictions from the regression trees as inputs" and write the objective of this price optimization problem in terms of conditional expectations "$\mathbb{E}\left[D_{ijk}\middle|p_j, k\right]$," a shorthand they use for conditional expectation of demand given price, product and other sale event features, and the price of competing styles. They report that "a key requirement for our pricing decision support tool is the ability to accurately predict demand."

We will return to these examples to discuss further the relationship to a prescription problem and the relevance of the specific additional explanatory variables included.

3.4 Identifiability

In the last section we saw that the prescription problem (3.1) is distinct from (3.2). Next we ask the question of when can we even solve the problem (3.1) based on data. We show that without further assumptions the solution is not identifiable. Let us begin with an example.

Example 3.4 (Consulting for the MIT Coop)**.** Nathan and Dimitris are hired by the MIT Coop to help determine an optimal sale price for the classic MIT hoodie, which the MIT Coop procures at a unit price of \$19 (so $r(p) = p - 19$). The MIT Coop is debating between a retail price of \$20 and a retail price of \$28. In any given week in the past, the MIT Coop has offered the hoodie at either of the two prices and observed either no units sold, one thousand units sold, or two thousand units sold. The Coop has a great deal of historical data.

Nathan and Dimitris collate the data into a table that shows the frequency of each price-demand combination over history:

Joint Distribution of Historical Price and Demand

	Price	
Demand (thousands)	\$20	\$28
0	0	8/28
1	8/18	1/18
2	1/18	0

Due to the abundance of data, Nathan and Dimitris are confident that this is a faithful representation of the joint distribution of (P, D).

Dimitris regresses demand on price by computing a weighted average in each of the above columns and finds that

$$\mathbb{E}\left[D\middle|P=p\right] = \begin{cases} 10/9 & p = 20 \\ 1/9 & p = 28 \end{cases}.$$

Taking a predictive approach, Dimitris seeks a model wherein the PRF is equal to the conditional expectation. He arrives at this model: and concludes that $p^* = 20$ is the optimal price. Looking back, Dimitris comes up with a model to explain the data and his PRF:

Dimitris's Demand Model

	$P = 20$		$P = 28$	
	$D(28) = 0$	$D(28) = 1$	$D(28) = 0$	$D(28) = 1$
$D(20) = 1$	32/81	4/81	32/81	4/81
$D(20) = 2$	4/81	1/162	4/81	1/162

Dimitris confirms that his model agrees completely with the observed data and with $\mathbb{E}\left[D(p)\right] = \mathbb{E}\left[D\middle|P=p\right]$. Dimitris computes

$$R(p) = r(p)\mathbb{E}\left[D(p)\right] = \begin{cases} 10/9 & p = 20 \\ 1 & p = 28 \end{cases}, \tag{3.3}$$

and concludes that $p^* = 20$ is the optimal price.

Nathan, working from home that day and unaware of Dimitris's progress, has independently come up with another model in order to explain the data:

Nathan's Demand Model

	$P = 20$		$P = 28$	
	$D(28) = 0$	$D(28) = 1$	$D(28) = 0$	$D(28) = 1$
$D(20) = 1$	$40/99$	$4/99$	$4/9$	$2/45$
$D(20) = 2$	0	$1/18$	0	$1/90$

Nathan, too, is able to confirm that his model completely agrees with the observed data. Under this model, Nathan calculates

$$\mathbb{E}\left[D(p)\right] = \begin{cases} 16/15 & p = 20 \\ 5/33 & p = 28 \end{cases}, \quad R(p) = r(p)\mathbb{E}\left[D(p)\right] = \begin{cases} 16/15 & p = 20 \\ 15/11 & p = 28 \end{cases}, \tag{3.4}$$

and concludes, differently from Dimitris, that $p^* = 28$ is in fact the optimal price.

Nathan and Dimitris had both come up with demand models that fully concur with the observed data but recommended different prices as optimal. Both models support the data fully. Both models give rise to the same conditional expectation (regression) function,

$$\mathbb{E}\left[D \middle| P = p\right] = \begin{cases} 10/9 & p = 20 \\ 1/9 & p = 28 \end{cases},$$

which, in particular, agrees with the price response function $\mathbb{E}\left[D(p)\right]$ under Dimitris's model, but not under Nathan's model. Both models, as well as the data, fully agree with a homoskendastic linear model,

$$D = \frac{65}{18} - \frac{1}{8}P + \epsilon, \qquad \epsilon = \begin{cases} -1/9 & \text{with prob. } 8/9 \\ 8/9 & \text{with prob. } 1/9 \end{cases}, \qquad \epsilon \perp\!\!\!\perp P.$$

Therefore, there cannot be an issue of a demand-functional model misspecification. Moreover, the regressor P is independent of the error ϵ.

Nonetheless, the two models recommended different optimal prices.

3.4.1 Non-Identifiability

The issue we encountered above is one of identifiability.

Definition 3.5. Let $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ be a model for the distribution of the observed data. We say that $\phi : \Theta \to \Phi$ is *identifiable* if for any F_{θ_1}, $F_{\theta_2} \in \mathcal{F}$ such that $F_{\theta_1} = F_{\theta_2}$, we have $\phi(\theta_1) = \phi(\theta_2)$.

In the above, neither Θ nor Φ need have any topology or algebraic structure, that is, the model need not be parametric. Note that if any ϕ is not identifiable then any finer quantity, such as θ itself, is not identifiable.

Letting Θ denote the joint distribution of price P and demand process $D(p)$ and letting ϕ map this to the optimal price (or, set thereof if many), we have that Example

3.4 above provides a proof by example of the following result:

Corollary 3.6. *The optimal price p^* is not identifiable on the basis of observations of (P, D).*

In fact, we proved this as a corollary of the stronger result (i.e., smaller Θ):

Theorem 3.7. *The optimal price p^* is not identifiable on the basis of observations of (P, D) even under the Gauss-Markov assumptions:*

i. Linearity: there is a random variable ϵ such that $D = \beta_0 + \beta_1 P + \epsilon$.

ii. Exogeneity of independent variables: $\mathbb{E}\left[\epsilon \middle| P\right] = 0$.

iii. Homoskendasticity: $\mathrm{Var}\left(\epsilon \middle| P\right) = \mathrm{Var}\left(\epsilon\right)$ is constant.

iv. No collinearity: P is not constant.

In Example 3.4, exogeneity and homoskendasticity are a consequence of $\mathbb{E}\left[\epsilon\right] = 0$ and $\epsilon \perp\!\!\!\perp P$. Exogeneity implies $\mathrm{Cov}(\epsilon, P) = 0$. Note that whenever the optimal price is not identifiable, the price response function $\mathbb{E}\left[D(p)\right]$, a finer quantity, is not identifiable either.

3.4.2 Conditions for Identifiability

In Corollary 3.6 we saw that observations of (P, D) are not in general sufficient to identify an optimal price. Therefore, we need something more. Let us now consider also auxiliary observations X of some characteristics of the historical sale events, e.g. characteristics of the online customer, whether a product was featured in a promotional flyer, seasonality. Our data is now $\{(P_1, X_1, D_1), \ldots, (P_n, X_n, D_n)\}$. Even with this data, for the moment, we still restrict ourselves to the problem of choosing a single price for the whole population of sale events; we consider the customized extension in Section 3.7.1.

In Section 3.3 we saw that one of the important distinctions between prediction and prescription is that in the former price P is random variable whereas in the latter price p is a control. The covariates X could help go from prediction to prescription if they allow us to factor out the association of the random variable P with the particular sale event and its demand process $D(p)$, which are the only relevant things in the prescription problem. A sufficient condition is that X accounts for the association between $D(p)$ and P:

Assumption 3.8 (Weak Ignorability)**.** For all $p \in \mathcal{P}$, we have

$$D(p) \perp\!\!\!\perp P \;\middle|\; X$$

Under this condition, which we discuss below, we have identifiability.

Theorem 3.9. *Under weak ignorability, the optimal price p^* is identifiable on the basis of observations of (P, X, D).*

Proof. Under ignorability, using iterated expectations, we have

$$\begin{aligned}\mathbb{E}\left[D(p)\right] &= \mathbb{E}\left[\mathbb{E}\left[D(p)\middle|X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[D(p)\middle|P=p,X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[D(P)\middle|P=p,X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[D\middle|P=p,X\right]\right].\end{aligned}$$

The last expectation is expressed solely in terms of the joint distribution of (P, X, D), which gives the identifiability of the price response function $\mathbb{E}\left[D(p)\right]$ and hence the revenue function $R(p)$. The optimal price is given by optimizing $R(p)$, completing the proof. □

Note the last expectation is not an iterated expectation because it is not conditioned on $P = p$. In words, it says to take the conditional expectation of D given $X = x$, $P = p$ and to average it over all x using the marginal distribution of X (and *not* the conditional distribution given $P = p$).

Discussion

In words, the weak ignorability condition says that, historically, X accounts for all the sale-event-specific features that influenced managerial price-setting. Managers usually set prices strategically rather than at random. Thus, if prices are set in (partial) anticipation of demand, then prices and demand are confounded. If X accounts for the information based upon which the price was selected then weak ignorability is nonetheless satisfied. Because of this, this sort of condition is sometimes termed selection on observables (this is, however, imprecise because weak ignorability does not imply that price is a *function* of observables). Note that the conditional independence in weak ignorability is of historical price and *potential* demand at a price p, not historically observed demand. The term *weak* means that the independence holds separately for each price p.

If the manager chooses prices without regard to any specific sale event then weak ignorability holds with a null X variable (formally, $\sigma(X) = \{\Omega, \varnothing\}$). In particular, this is the case in dynamic demand learning and pricing as in Bertsimas and Perakis (2006), Besbes and Zeevi (2009), Harrison et al. (2012) because each sale event is assumed independent and nothing about a present sale event is considered when setting the price. This is the experimental setting. Unfortunately, it rarely holds in practice for historical data.

If there is not sufficient recorded information in X to merit the weak ignorability assumption, it is said that there is residual endogeneity. In this case, Theorem 3.9 fails, but there may be other conditions that enable identification such as the availability of instrumental variables (see Bijmolt et al. (2005) for a discussion of endogeneity in demand-price-response estimation). Here we focus on data that arises from historical pricing by managers whose behavior is well understood or even documented and hence what informed pricing decisions is known and observable to the same managing entity.

Let us consider weak ignorability and its ramifications in some specific examples.

Example 3.10 (Consulting for the MIT Coop)**.** Consider again the hypothetical case of Example 3.4. Recall, Dimitris and Nathan both came up with models for demand that completely agreed with the data but gave rise to different optimal prices. Thus, we concluded that the data observed could not possibly identify the right optimal price. What would allow us to identify the right, unique model is weak ignorability as per Theorem 3.9.

Suppose weak ignorability holds with X being a null variable, i.e. without any extra information. This condition then eliminates Nathan's model – it no longer agrees with both the data and this condition. On the other hand, Dimitris's model remains valid – in fact it turns out to be the unique model that agrees with both the data and this condition. Hence, under this condition, $p^* = 20$ is the correct optimal price.

But for weak ignorability to hold with X being a null variable we would have needed experimental data, where prices are set at random for the sake of experiment. This is almost never the case in real, historical datasets.

Suppose instead that we recorded additional information about each sale event: whether a man dressed as Tim the Beaver was outside the Coop to promote MIT apparel ($X = 1$) or not ($X = 0$). On average, Tim the Beaver was promoting the store 2 days out of the month ($\mathbb{P}(X = 1) = 2/30$). Suppose tallying the historic observations led to the following summary of the data.

Joint Distribution of Historical Price, Demand, and Tim-the-Beaver Campaigns

	$X = 0$		$X = 1$	
Demand (k)	$P = 20$	$P = 28$	$P = 20$	$P = 28$
0	0	4/9	0	0
1	4/9	2/45	0	1/90
2	0	0	1/18	0

If prices are chosen independently of whether Tim is campaigning, the previous scenario still holds and Dimitris's model is the uniquely correct one. If, however, to complement Tim the Beaver's promotion and to better capitalize on the opportunity to attract purchases, prices were more often cut to \$20 when Tim was campaigning outside, then we are no longer in the experimental setting. In fact, if we assume weak ignorability holds with X being whether Tim is campaigning then it turns out that Dimitris's model is ruled out and Nathan's model is the unique model that accommodates both this condition and the data observed, in which case $p^* = 28$ is the correct optimal price. In fact, Nathan's model can be written as follows. If Tim is not campaigning then $D(20) = 1$, $D(28) = 0$ with probability 10/11 and otherwise 0, and $P = 20$ with probability 10/21 and otherwise 28, each independently. If Tim is campaigning then $D(20) = 2$, $D(28) = 1$, and $P = 20$ with probability 5/6 and otherwise 28, each independently. In this hypothetical example, we are seeking a universal

price, to be set a priori without regard to Tim, but the price can also be customized according to Tim's presence (as was done historically); we consider customization in Section 3.7.1.

Example 3.11 (Auto Loan Rate Optimization)**.** Consider again the case of Example 3.2. There is a great amount of information about each loan applicant and the associated sale event, including the FICO credit score of the applicant, the length of the term over which the loan is to be repaid, the dollar amount of the loan, whether the car to be purchased is new, used, or refinanced, competitors' rate, prime rate, and who referred the applicant.

The dataset description (Columbia University Center for Pricing and Revenue Management 2012) says that approval and rate is based on "credit information and other criteria." Such criteria would almost certainly also be associated with the potential likelihood of the consumer to accept a loan offer at any one particular rate. Therefore, weak ignorability does not hold with null X – the data is not experimental. If the rate is chosen solely based on FICO score then weak ignorability would hold with X being the FICO score. It is said, however, that "other criteria" are used too. If, nonetheless, the data represents all the information that is provided by the applicant and hence all the information that the on-line lender could potentially based its price on then weak ignorability would hold with X consisting of this data too.

In Besbes et al. (2010), the authors consider setting a single price for eight customer segments each defined by a range of FICO scores, range of term lengths, and season when they applied. Paraphrased, their approach to pricing is to estimate $\mathbb{E}\left[D\middle|P=p,Y=i\right]$ $(=\mathbb{P}\left(D=1\middle|P=p,Y=i\right)$ because demand is binary) within each segment either parametrically or non-parametrically, where $Y=s(X)\in\{1,\ldots,8\}$ denotes membership in the population segments based on some components of X, and prescribing the segment-wide price that maximizes this estimated conditional expectation times per-unit revenue. How $\mathbb{E}\left[D\middle|P=p,Y=i\right]$ relates to $\mathbb{E}\left[D(p)\middle|Y=i\right]$, and hence how this estimated objective relates to the true objective of the pricing problem, depends upon weak ignorability. The authors assume that data points (D_i,P_i) are iid (Assumption 1 therein) but, while this , this does not mean any sort of independence within each data point such as weak ignorability.

Since $Y=s(X)$ is coarser than X, weak ignorability with respect to X does not generally imply the same with respect to the coarser Y. In particular, since, besides coarsening, Y also removes price-driving covariates such as loan amount, weak ignorability with respect to Y is not a reasonable assumption suggesting that even the non-parametric model the authors consider need not converge to the true PRF. Conditional expectations given X can be estimated based on partitioning into segments, but these must be data-driven and shrinking with sample size, not fixed a priori by $s(X)$ and must not remove whole dimensions.

It is important, nonetheless, to keep in mind that all of this is completely moot if, at the end of the day, revenues generated by any particular pricing scheme are indistinguishable from optimal. Using a statistical test we develop in Section 3.6, we test whether this is the case – or whether a finer analysis leads to greater revenue – when we next return to this example.

Example 3.12 (Other Examples from the Literature)**.** Recall the case of Cohen et al. (2014), where the authors study a data-driven multi-period dynamic pricing problem for supermarket goods. The authors use a regression of demand on present and recent prices based on historical, observational data in order to formulate a revenue objective. For this regression to inform prescription, we would need weak ignorability with X being past prices. It is unlikely that this holds. The pricing problem in Cohen et al. (2014) is billed as promotion optimization. Indeed, price promotions usually go hand-in-hand with promotions in other aspects of the product mix (known as the four P's: price, product, promotion, and place). For example, promoting a product in the weekly ad flyer is usually coincident with price-cutting promotions. The same for promoting an item by placing it the end of aisles and many other types of promotions. These most certainly have a strong effect on potential demand, were price set at any one particular price. Since these non-price promotions would historically also have a statistical association with price in observational data means that weak ignorability does not hold. The same questions arise in the case of Johnson et al. (2014) – does the regression have a causal meaning and, hence, does the optimization problem have a prescriptive meaning.[1]

Again, all this would be moot if revenues generated could not be statistically distinguished from optimal as it would not be clear that another approach could potentially be better. The tools we develop in Section 3.6 allow for such a scenario to be tested.

3.5 Solutions to the Prescriptive Problem

In the last section, we saw that weak ignorability enables identification, that is, the data-driven pricing problem is hypothetically solvable. Now we turn to solutions. In this section, assuming weak ignorability, we propose specific data-driven solutions to the prescriptive problem.

3.5.1 A Non-Parametric Solution

We begin with a non-parametric solution that is non-model-dependent in that it will converge to optimal pricing regardless of the true underlying model, given sufficient data and some mild assumptions. Henceforth, we assume that X is a vector of covariates taking values in $\mathbb{R}^k$.

[1] In Johnson et al. (2014), the authors write, "we calculated the correlation between our error term, $\hat{d}_i - d_i$, and all of the features in our regression as one test to identify if there were any systematic biases from potentially unobserved factors or endogeneity." This, however, cannot possibly test for endogeneity because errors in the "short" regression (that with only the observed covariates) would always be orthogonal to the regressors by the very definition of the error term. The endogeneity that concerns causality is when errors in the "long" regression (an assumed model that represents a hypothetical causal relationship and so-called because it includes additional, hidden covariates) are not orthogonal to regressors in the "short" regression. Generally, instruments are needed to test endogeneity.

The proof of Theorem 3.9 says that under weak ignorability, the PRF is $\mathbb{E}\left[D(p)\right] = \mathbb{E}\left[\mathbb{E}\left[D\middle|P=p,X\right]\right]$. Thus, to estimate the revenue function, one approach may be to estimate the regression function $\mathbb{E}\left[r(P)D\middle|P=p,X=x\right]$ and then average its values over an estimate for the marginal distribution of X. This yields an estimate of the revenue function $R(p)$, which can then be optimized. Nadaraya-Watson kernel regression (Nadaraya 1964, Watson 1964) can be used to estimate the regression function non-parametrically. The estimate, based on a kernel function $K : \mathbb{R}^{1+k} \to \mathbb{R}_+$ and bandwidth h_n, is

$$\overline{R}_n(p,x) = \frac{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right) r(P_i)D_i}{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right)}, \tag{3.5}$$

where $K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right) = K\left(\frac{p-P_i}{h_n}, \frac{x_1-X_{i1}}{h_n}, \ldots, \frac{x_k-X_{ik}}{h_n}\right)$. A kernel function mimics a continuous distribution centered at the data points, the width of which is determined by the bandwidth. Indeed, such a regression estimator arises as the conditional expectation with respect to the Parzen window density estimator for the joint distribution of $(P, X, r(P)D)$ and that of (P, X), where the Parzen window density estimator is in essence a smoothed histogram with continuous distributions instead of Dirac deltas (Parzen 1962). There are a variety of kernels used in practice (Hardle 1990). Our requirements for a kernel function and bandwidth are as follows:

Assumption 3.13 (Kernel Conditions)**.**

i. $0 < \int_{\mathbb{R}^{1+k}} K < \infty$.

ii. K is zero outside a bounded set.

iii. K is twice Lipschitz-continuously differentiable.

iv. K has order at least $s \in \mathbb{N}$, that is, $\int K(u)u^\alpha du = 0 \;\; \forall \alpha \in \mathbb{N}^{1+k} : |\alpha| < s$.

v. K is symmetric in its first argument.

vi. $h_n \to 0$ and $nh_n^{2s+3} \to 0$.

vii. $nh_n^{k+5}/\log(n) \to \infty$ and $nh_n^{2k+1}/\log(n)^2 \to \infty$.

A non-parametric estimate of the marginal distribution of X is the empirical distribution, which places unit mass at each of the observations X_i. Combining these estimates as detailed above, we arrive at the following estimate for the revenue function

$$\overline{R}_n(p) = \frac{1}{n}\sum_{i=1}^n \overline{R}_n(p, X_i) = \frac{1}{n}\sum_{i=1}^n \frac{\sum_{j=1}^n K\left(\frac{p-P_j}{h_n}, \frac{X_i-X_j}{h_n}\right) r(P_j)D_j}{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{X_i-X_j}{h_n}\right)}, \tag{3.6}$$

which leads to the following non-parametric data-driven price prescription

$$\overline{p}_n \in \arg\max_{p\in\mathcal{P}} \overline{R}_n(p). \tag{3.7}$$

The question that arises is how does $\overline{p}_n$ behave asymptotically. In particular, does this pricing strategy converge to optimality, both the price itself and its revenue performance. Since the estimates are non-parametric, the expectation is that this can occur under model-free assumptions. Next we show that this is indeed the case. First, we require additional assumptions.

Assumption 3.14 (Optimal Price Conditions)**.**

i. $\mathcal{P}$ is compact.

ii. p^* uniquely maximizes $R(p)$ on $\mathcal{P}$.

iii. p^* lies in the interior of $\mathcal{P}$.

iv. $R(p)$ is twice continuously differentiable and $R''(p^*) < 0$.

Assumption 3.15 (Distributional Conditions)**.**

i. X and P are continuously distributed with joint density $f_{P,X}(p,x)$ and X has marginal density $f_X(x)$ that is bounded and continuously differentiable.

ii. (X,P) have compact support, on which $f_{P,X}$ is bounded away from zero.

iii. $\mathbb{E}\left[D^4\right] < \infty$ and $\mathbb{E}\left[D^4 \middle| P = p, X = x\right]$ is bounded.

iv. $\mathbb{E}\left[D^2 \middle| P = p, X = x\right]$ is continuously differentiable.

v. $\mathbb{E}\left[D \middle| P = p, X = x\right]$ and $f_{P,X}(p,x)$ are $s+1$ times continuously differentiable with bounded derivatives.

vi. There exists $\epsilon > 0$ such that

$$\int \sup_{|\xi|\leq\epsilon} \left(1 + r(p^*+\xi)^4 \mathbb{E}\left[D^4 \middle| P = p^*+\xi, X = x\right]\right) f_{P,X}(p^*+\xi, x)dx < \infty.$$

Under these conditions, we can show the following asymptotic optimality and rates.

Theorem 3.16. *Under Assumptions 3.8, 3.13, 3.14, and 3.15, we have that*

$$\begin{aligned}
\sqrt{nh_n}(R(p) - \overline{R}_n(p)) &\xrightarrow{d} \mathcal{N}\left(0,\, \eta_p\right) \quad \forall p \in \mathcal{P}, \\
\sqrt{nh_n^3}(p^* - \overline{p}_n) &\xrightarrow{d} \mathcal{N}\left(0,\, \frac{\eta'}{R''(p^*)^2}\right) \\
\left(nh_n^3\right)\left(R(p^*) - R(\overline{p}_n)\right) &\xrightarrow{d} \frac{-\eta'}{2R''(p^*)}\chi_1^2,
\end{aligned}$$

and, if also $nh_n^{2s+1} \to 0$*, then*

$$\sqrt{nh_n}(R(p^*) - \overline{R}_n(\overline{p}_n)) \xrightarrow{d} \mathcal{N}(0,\, \eta)\,,$$

where $\mathcal{N}(0,\, \sigma^2)$ *denotes a centered normal distribution with variance* σ^2*,* χ^2_1 *denotes a chi-squared distribution with* 1 *degree of freedom, and* η*,* η'*,* η_p *are constants defined as follows*

$$\eta = r(p^*)^2\ \mathbb{E}\left[\frac{\mathrm{Var}\left(D\middle|P = p^*, X\right)}{f_{P|X}(p^*|X)}\right]\int \tilde{K}(p)^2 dp,$$

$$\eta' = r(p^*)^2\ \mathbb{E}\left[\frac{\mathrm{Var}\left(D\middle|P = p^*, X\right)}{f_{P|X}(p^*|X)}\right]\int \tilde{K}'(p)^2 dp,$$

$$\eta_p = r(p^*)^2\ \mathbb{E}\left[\frac{\mathrm{Var}\left(D\middle|P = p, X\right)}{f_{P|X}(p|X)}\right]\int \tilde{K}(p)^2 dp,$$

where $\tilde{K}(p) = \int K(p,x)dx$ *and* $f_{P|X}(p|x) = f_{P,X}(p,x)/f_X(x)$ *is the conditional distribution.*

Proof. See appendix. □

The main take away from this theorem is that under regularity conditions, but without model specification, the non-parametric pricing strategy has revenues that converge to optimal as $1/n$. Note that Assumption 3.13 implies that $s \geq k$ when $k \geq 3$ and $s \geq k+1$ when $k \leq 2$. This means that a so-called bias-reducing kernel (order greater than 2) is necessary when $k \geq 2$ in order to faithfully satisfy the assumptions of Theorem 3.16. Such kernels must have a negative part and are not probability densities.

3.5.2 A Parametric Solution

In the preceding section we developed a non-parametric pricing strategy that converged to optimal without requiring any model to be specified. Non-parametric approaches, however, can sometimes be unwieldy because they may be slow to converge and their shapelessness makes them uninterpretable. In fact, there is a growing body of work (Besbes et al. 2010, Besbes and Zeevi 2015) arguing that parametric models are often sufficient for prescriptive problems, where the model may need only fit well near the optimum. In particular, what matters is not model fit but objective performance. In this section we develop a particular parametric pricing strategy using a generalization of the propensity score.

The propensity score is a common matching metric used in the comparison of binary treatments in observational data (Rosenbaum and Rubin 1983). In particular, the conventional propensity score of a study subject is equal to the conditional probability of receiving the treatment of interest given the subject's covariates X. If treatments are continuous, the generalized propensity score (Robins et al. 2000,

Hirano and Imbens 2004, Imai and Van Dyk 2004) of a unit is defined as the conditional density of the unit receiving whatever treatment it did receive given the subject's covariates.

In our problem, the generalized propensity score is $Q = \phi(p, x)$ where $\phi(p, x) = f_{P|X}(p|x)$, that is, one takes the conditional density $f_{P|X}(p|x)$, which is non-random, and plugs in as values the random variables P and X. The key property of the generalized propensity score is that it is sufficient as a control for identifying the PRF. The following is an adaption of a common result.

Theorem 3.17. *Suppose weak ignorability holds. Then the PRF satisfies*

$$\mathbb{E}\left[D(p)\right] = \mathbb{E}\left[d(p, \phi(p, X))\right], \text{ where } d(p, q) = \mathbb{E}\left[D\middle|P = p, Q = q\right]$$

Proof. See appendix. □

The take away from this theorem is that the generalized propensity score allows for dimensionality reduction – it is sufficient to control just for this single univariate quantity rather than all of X. In the binary treatment case, the conventional propensity score is the coarsest such control (Rosenbaum and Rubin 1983). The limitation, of course, of using the generalized propensity score is that neither $\phi(p, x)$ nor the scores themselves are known. Before considering estimation, let us consider an artificial example to get a handle on generalized propensity scores and the uses of Theorem 3.17.

Example 3.18 (Artificial Continuous Example)**.** Recall the setup from Example 3.1: $c = 50$, $\mathcal{P} = [50, 300]$,

$$D(p) = 15\left(300 - p\right)_+ + 1500pXe^{-(p-50)X/10} + \epsilon(p),$$

where $X \sim \text{Exp}(1)$ is an exponentially distributed random disturbance and $\epsilon(p)$ is a Gaussian process with kernel $\text{Cov}\left(\epsilon(p), \epsilon(p')\right) = \sigma^2\mathbb{I}\left[p = p'\right]$, and $P = 50 + 10Y/X$ where $Y \sim \text{Exp}(1)$ is an exponentially distributed random disturbance. Recall that the PRF in this case is

$$\mathbb{E}\left[D(p)\right] = 15(300 - p)_+ + \frac{150000p}{(p - 40)^2}.$$

First, note that weak ignorability holds with respect to X. Since $P = 50 + 10Y/X$, we have that

$$\phi(p, x) = \frac{x}{10}e^{-(p-50)x/10}.$$

Hence the generalized propensity score is

$$Q = \frac{X}{10}e^{-(P-50)X/10} = \frac{X}{10}e^{-Y}.$$

This means that we can write $D = D(P)$ as

$$D = 15\left(300 - P\right)_+ + 15000PQ + \epsilon(P)$$

and hence

$$d(p,q) = \mathbb{E}\left[D|P=p,Q=q\right] = 15\left(300-p\right)_+ + 15000pq.$$

Since

$$\mathbb{E}\left[\phi(p,X)\right] = \mathbb{E}\left[\frac{X}{10}e^{-(p-50)X/10}\right] = \frac{10}{(p-40)^2},$$

we conclude that

$$\begin{aligned}\mathbb{E}\left[d(p,\phi(p,X))\right] &= 15\left(300-p\right)_+ + 15000p\mathbb{E}\left[\phi(p,X)\right] \\ &= 15(300-p)_+ + \frac{150000p}{(p-40)^2},\end{aligned}$$

which indeed agrees with the PRF $\mathbb{E}\left[D(p)\right]$.

Theorem 3.17 motivates the following procedure: estimate a conditional probability model to fit $\phi(p,x)$, impute generalized propensity scores $\hat{Q}_i = \phi(P_i, X_i)$, regress demand on price and imputed scores, average this regression over $\phi(p, X_i)$, and prescribe the price p that maximizes per-unit revenue times this estimate of the PRF. Specifically, the following parametric approach can be followed:

1. Regress P on X by fitting a generalized linear model (GLM) in order to estimate $\phi(p,x)$. For example, one can fit a simple linear regression

   $$\left(P\middle|X=x\right) \sim \mathcal{N}\left(\beta_0 + \beta^T x,\, \sigma^2\right),$$

   estimating $\hat{\beta}_n$ and $\hat{\sigma}_n$ by ordinary least squares, and then estimate $\phi(p,x)$ by

   $$\hat{\phi}_n(p,x) = \frac{1}{\sqrt{2R\hat{\sigma}_n^2}} e^{-\frac{(p-\hat{\beta}_{n0}-\hat{\beta}_n^T x)^2}{2\hat{\sigma}_n^2}}.$$

   Alternatively, we can fit a GLM with any overdispersed exponential family, i.e., choose $\hat{\beta}_n$, $\hat{\tau}_n$ by maximum likelihood given the model

   $$f_{P|X}(p|x;\beta,\tau) = h(p,\tau)\exp\left(\frac{b(\beta_0+\beta^T x)T(p) - A(\beta_0+\beta^T x)}{d(\tau)}\right).$$

   The above means that, given X, P is a member of an overdispersed exponential family with canonical parameter $\beta_0 + \beta^T x$ and dispersion parameter τ. See McCullagh et al. (1989) for more detail on GLMs. Fitting the GLM provides us with a more general estimate for $\phi(p,x)$:

   $$\hat{\phi}_n(p,x) = f_{P|X}(p|x;\hat{\beta}_n,\hat{\tau}_n).$$

2. Use $\hat{\phi}_n(p,x)$ to impute generalized propensity scores, setting $\hat{Q}_i = \hat{\phi}_n(P_i, X_i)$.

3. Regress D on P and $\hat{Q}$ using a flexible parametric regression to estimate $d(p,q)$,

e.g.

$$D = b(\alpha_0 + \alpha_1 p + \alpha_2 q + \alpha_3 q^2 + \epsilon)$$

via some link function b. Depending on the particular case, it may be appropriate to regress $\log(D)$ on $\log(P)$ and $\hat{Q}$ instead (log-log model). Call $\hat{d}_n(p, q)$ our estimate of $d(p, q)$.

4. Use $\hat{\phi}_n(p, x)$ and $\hat{d}_n(p, q)$ to estimate the PRF,

$$\hat{d}_n(p) = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_n(p, \hat{\phi}_n(p, X_i)),$$

and prescribe the price that optimizes estimated revenues,

$$\hat{p}_n \in \arg\max_{p \in \mathcal{P}} r(p) \hat{d}_n(p).$$

The above procedure provides a flexible parametric framework for data-driven pricing with observational data. When we apply it to examples both real and synthetic in Section 3.6.3 we find that it performs well and produces revenue that is statistically indistinguishable from optimal.

3.5.3 A Semi-Parametric Solution

We can also use the generalized propensity score to correct the naïve application of kernel regression as in Besbes et al. (2010) to observational data. If scores are estimated parametrically and the PRF is estimated non-parametrically as delineated below and then optimized, we arrive at a semi-parametric solution.

We begin by showing the following relationship.

Theorem 3.19. *Under weak ignorability,*

$$\mathbb{E}\left[D(p)\right] = \frac{\mathbb{E}\left[D/Q \middle| P = p\right]}{\mathbb{E}\left[1/Q \middle| P = p\right]}$$

Proof. Consider any random variable Y. Note that

$$\begin{aligned}
\mathbb{E}\left[Y \middle| P = p, X = x\right] &= \mathbb{E}\left[Y \mathbb{I}\left[P = p\right] \middle| X = x\right] / f_{P|X}(p|x) \\
&= \mathbb{E}\left[\frac{Y \mathbb{I}\left[P = p\right]}{\phi(p, x)} \middle| X = x\right] \\
&= \mathbb{E}\left[\frac{Y \mathbb{I}\left[P = p\right]}{Q} \middle| X = x\right].
\end{aligned}$$

Therefore, considering $Y = D$,

$$\begin{aligned}\mathbb{E}\left[D(p)\right] &= \mathbb{E}\left[\mathbb{E}\left[D(p)\middle|X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[D\middle|P=p,X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{D\mathbb{I}\left[P=p\right]}{Q}\middle|X\right]\right] = \mathbb{E}\left[\frac{D\mathbb{I}\left[P=p\right]}{Q}\right] = \mathbb{E}\left[D/Q\middle|P=p\right]f_P(p).\end{aligned}$$

Similarly, considering $Y = 1$, we have

$$\begin{aligned}1 &= \mathbb{E}\left[\mathbb{E}\left[1|X,P\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left[P=p\right]/Q\middle|X\right]\right] \\ &= \mathbb{E}\left[\mathbb{I}\left[P=p\right]/Q\right] = \mathbb{E}\left[1/Q\middle|P=p\right]f_P(p).\end{aligned}$$

Dividing and canceling $\mathbb{P}\left(P=p\right)$ yields the result. □

If we knew Q, the kernel estimator for $\mathbb{E}\left[D/Q\middle|P=p\right]$ would be

$$\frac{\sum_{i=1}^{n} {Q_i}^{-1} K\left(\frac{p-P_i}{h_n}\right) D}{\sum_{i=1}^{n} K\left(\frac{p-P_i}{h_n}\right)}$$

and the kernel estimator for $\mathbb{E}\left[1/Q\middle|P=p\right]$ would be

$$\frac{\sum_{i=1}^{n} {Q_i}^{-1} K\left(\frac{p-P_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{p-P_i}{h_n}\right)}.$$

Taking their ratios and imputing parametrically estimated scores $\hat{Q}_i$, Theorem 3.19 suggests the following semi-parametric estimator for the PRF:

$$\hat{d}_n(p) = \frac{\sum_{i=1}^{n} \hat{Q}_i^{-1} K\left(\frac{p-P_i}{h_n}\right) D}{\sum_{i=1}^{n} \hat{Q}_i^{-1} K\left(\frac{p-P_i}{h_n}\right)}.$$

Compared to the kernel estimator used in Besbes et al. (2010), here the kernel weights are corrected by an inverse-score weighting. This eliminates the bias due to the observational nature of the data (under weak ignorability). The similar structure makes clear some connections. In particular, in the experimental setting Q is constant and then in that case would cancel out in the above, leaving us with the simple kernel estimator.

We do not explore the semi-parametric approach further because, in agreement with previous findings, we find in Section 3.6.3 that a parametric approach is sufficient (where the fully non-parametric approach is used as a benchmark for testing). The above also suggests an alternative fully non-parametric approach where the scores $\hat{Q}_i$ are estimated non-parametrically also using kernel regression, but there is no clear benefit to such a two-step approach over the direct non-parametric approach explored in the earlier section.

3.6 A Test for Revenue Optimality

In the previous sections we considered various data-driven pricing strategies for observational data. All of these proceeded by estimating the PRF and optimizing resulting estimated revenue. Similarly, in their own context, each of Besbes et al. (2010), Cohen et al. (2014), Johnson et al. (2014) first estimated demand then optimized price. It may be argued that it is important that estimated revenues faithfully represent true revenues, but in fact this point is moot insofar as actual revenues generated by the resulting pricing strategy are satisfactory. This is the key point made by Besbes et al. (2010) where the authors develop a hypothesis test to inspect revenue optimality instead of predictive model fit. In the perspective presented herein, this test does not apply to observational data because it relies on a naïve kernel estimate of conditional expectation of revenues in order to estimate the true revenue function, but this may in general bear no relationship to the conditional expectation being estimated. The purpose of this section is to build on their work in developing an analogous test for the observational setting under the assumption of weak ignorability.

The standard hypothesis to be tested when verifying model fit is whether one's estimation of the PRF is consistent with the true PRF, that is, equal at all prices p. This is not necessarily of direct interest in a prescriptive problem such as data-driven pricing. Instead, the concern is whether the given pricing strategy is misguided in that it approaches suboptimal revenues. The null hypothesis is that it does not.

Consider some data-driven pricing strategy $\hat{p}_n$. Let $\hat{p}$ be the price that this strategy, given infinite data, will eventually arrive at. We leave this somewhat vague to make the test flexible, only requiring the following condition in defining what $\hat{p}$ means.

Assumption 3.20 (Convergent Pricing Strategy)**.** $\hat{p}_n - \hat{p} = O_p(1/\sqrt{n})$ for some fixed $\hat{p} \in \mathcal{P}$.[2]

For example, consider the strategy that fits a naïve kernel regression to estimate conditional expectation of revenues and optimizes it, without regard to causality. Such a strategy, given infinite data, will eventually arrive at the price $\hat{p}$ that optimizes the true conditional expectation of revenues. In particular, it is true that $\hat{p}_n - \hat{p} = O_p(1/\sqrt{n})$ (Ziegler 2002). A similar condition is true of the strategy that uses a parametric maximum-likelihood regression (Besbes et al. 2010).

The hypothesis we would like to test is

$$H_0 : R(p^*) = R(\hat{p})$$

against the alternative

$$H_1 : R(p^*) > R(\hat{p}).$$

That is, we would like to test whether the nominal price that our pricing strategy would be prescribing is generating optimal revenues. The difference between the hypothesis we consider and the one considered by Besbes et al. (2010) is in the definition of $R(p)$ (i.e. $\mathbb{E}\left[r(p)D(p)\right]$ vs. $r(p)\mathbb{E}\left[D\middle|P = p\right]$).

[2]The notation $Y_n = O_p(a_n)$ means that for any $\epsilon > 0$ there is $M > 0$ such that $\mathbb{P}\left(|Y_n/a_n| > M\right) < \epsilon$ eventually. In particular, if Y_n/a_n converges in distribution then $Y_n = O_p(a_n)$.

A hypothesis test is a procedure that either rejects H_0 in favor of H_1 or claims that there is insufficient evidence to reject H_0. The test should only falsely reject H_0 at a bounded rate, known as *significance*. A consistent hypothesis test will eventually reject H_0 whenever it is false (i.e., given sufficient data). Thus, a consistent test must balance caution in rejecting H_0 when it might actually be true and audacity in rejecting it when the evidence supports it or risk letting the false hypothesis slide. Since a hypothesis test is based on data and should not reject H_0 without significant evidence in the data, a test for our hypothesis can be interpreted as only rejecting a pricing strategy if it *generates revenues that are distinguishable from optimal to a statistically significant degree*. If revenues are not statistically distinguishable from optimal, a pricing strategy should not be rejected (but this does not mean H_0 is true).

3.6.1 Test Statistic and Large Sample Theory

The impediment to verifying our hypothesis is that $R(p)$, p^*, and $\hat{p}$ are all unknown; were they known, we would compute $\rho = R(p^*) - R(\hat{p})$ and compare it to 0. Therefore, we must come up with an observable test statistic as a proxy to ρ. We do this by replacing the unknowns by our consistent estimates for them. We replace $R(p)$ and p^* by our non-parametric estimates $\overline{R}(p)$ as in (3.6) and $\overline{p}$ as in (3.7) and we replace $\hat{p}$ by $\hat{p}_n$. The resulting test statistic is

$$\rho_n = \overline{R}_n(\overline{p}_n) - \overline{R}_n(\hat{p}_n).$$

If ρ_n is small, we have reason to believe that $\rho = 0$, whereas if ρ_n is large, we would believe that $\rho > 0$. The question is where to draw the line.

Theorem 3.21. *Suppose Assumptions 3.8, 3.13, 3.14, 3.15, and 3.20 hold. Let* $\Gamma = \frac{-\eta'}{2R''(p^*)}$ *with* η' *defined as in Theorem 3.16. Then,*

i. *under* H_0, $(nh_n^3)\,\rho_n \xrightarrow{d} \Gamma\chi_1^2$, *and*

ii. *under* H_1, $(nh_n^3)\,\rho_n \xrightarrow{d} \infty$.

Proof. See appendix. □

Theorem 3.21 says that if we only reject H_0 when $\rho_n > n^{-1}h_n^{-3}\Gamma F_{\chi_1^2}^{-1}(1-\alpha)$ (where $F_{\chi_1^2}^{-1}$ is the chi-squared quantile function), then when H_0 is true we would only falsely reject H_0 a $1-\alpha$ fraction of the time (asymptotically). On the other hand, if H_0 is false, then we would eventually reject it using such a procedure. The problem is that Γ is unknown meaning that this exact procedure cannot be implemented in practice.

3.6.2 A Hypothesis Test

One way to implement a hypothesis is to estimate Γ and replace the estimate into the results of Theorem 3.21. In particular, given any estimate $\hat{\Gamma}_n$ that converges in probability to Γ, we would have as an immediate consequence of Theorem 3.21 that

$(nh_n^3)\,\hat{\Gamma}_n^{-1}\rho_n$ converges in distribution to χ_1^2 under H_0 and to ∞ under H_1. This would give an implementable test. Non-parametric estimators for Γ, however, would tend to be convoluted and unwieldy, involving partial means of estimators of conditional variance and density as well as fragile estimates of second derivatives of partial means.

Instead, we consider an alternative approach that estimates Γ via the bootstrap (Efron and Tibshirani 1993). The proof of Theorem 3.21 shows that the term that dominates the behavior of ρ_n under H_0 is $A_n = \overline{R}_n(\overline{p}_n) - \overline{R}_n(p^*)$ and that, in particular, $(nh_n^3)\,A_n \xrightarrow{d} \Gamma\chi_1^2$. To estimate Γ we can estimate the mean of A_n. Following Besbes et al. (2010), we use the bootstrap to do this. The fact that ρ_n is asymptotically pivotal (the asymptotic distribution is independent of $\hat{p}$) suggests that a bootstrap procedure could be particularly powerful (Horowitz 2001).

The bootstrap procedure proceeds as follows. Compute $\overline{p}_n$ as in (3.7) based on the data $\mathcal{S}_n$. Fix B large. For $b = 1, \ldots, B$, do:

1. Draw n samples with replacement from $\mathcal{S}_n$ to form the resampled dataset $\mathcal{S}_n^{(b)}$.
2. Compute $\overline{R}_n^{(b)}$ and $\overline{p}_n^{(b)}$ as in (3.6)-(3.7) based on the data $\mathcal{S}_n^{(b)}$.
3. Set $A_n^{(b)} = \overline{R}_n^{(b)}(\overline{p}_n^{(b)}) - \overline{R}_n^{(b)}(\overline{p}_n)$.

Let $\hat{\Gamma}_n = \frac{nh_n^3}{B}\sum_{i=1}^n A_n^{(b)}$. Reject H_0 if $\rho_n > n^{-1}h_n^{-3}\hat{\Gamma}_n F_{\chi_1^2}^{-1}(1-\alpha)$.

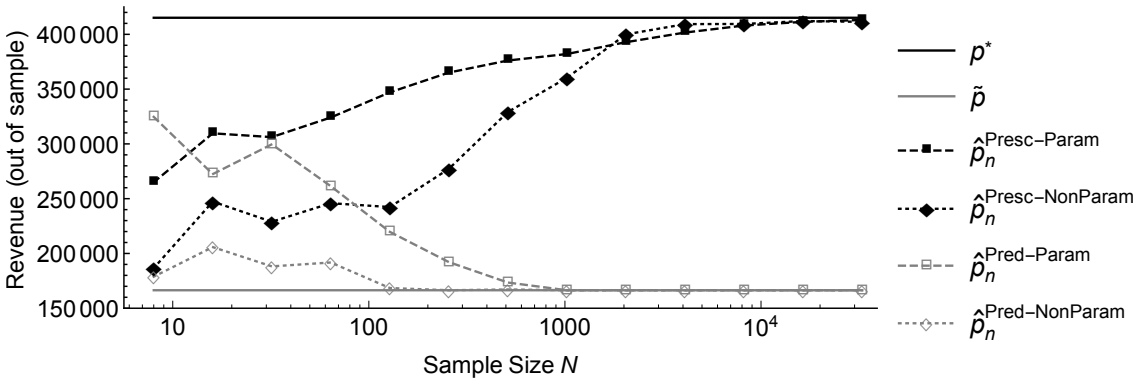
This bootstrap procedure is more attractive than convoluted kernel estimates of Γ because it is less dependent on parameters and it deals more directly with the finite-sample distribution of ρ_n. We use this bootstrap test in our numerical experiments in the next section.

3.6.3 Examples

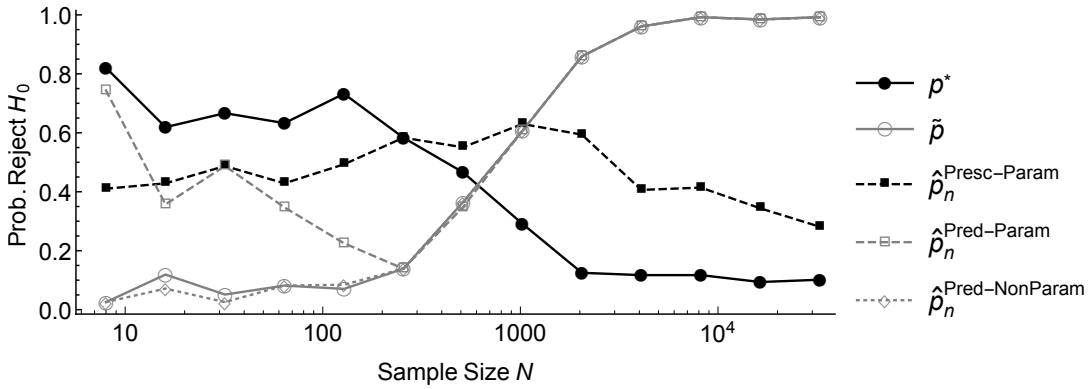
In this section we use our test to study the distinction between prediction and prescription in both synthetic and real examples and whether it has any operational consequence. We consistently find that ignoring the distinction in cases with observational data can significantly hurt revenues. On the other hand, we find that a parametric approach is sufficient for good performance as long as it takes into account this distinction.

Example 3.22 (Artificial Continuous Example)**.** Consider again the setup from Example 3.1 and consider recording n observations from (P, X, D). We compare four different pricing strategies: prescriptive non-parametric, prescriptive parametric, predictive non-parametric, and predictive parametric. The prescriptive non-parametric strategy is $\overline{p}_n$ as in (3.7) using a second order Gaussian kernel $K(u) = e^{-\frac{||u||_2^2}{2h_n^2}}$ and $h_n = (n\log(n))^{-1/7}$, which satisfy Assumption 3.13 with $s = 2$, $k = 1$. Note that Assumptions 3.14 and 3.15 are also satisfied by construction. For the prescriptive parametric strategy, we follow our procedure from Section 3.5.2 using a conditional gamma model for the GLM in step 1 and a quadratic regression in step 3. For the predictive non-parametric strategy, we apply a naïve kernel regression of D on P

Figure 3-2: Comparing Predictive and Prescriptive Data-Driven Pricing Strategies



(a) Revenues



(b) Test results

(using the same kernel and bandwidth) and optimize the estimated conditional expectation of revenues. Finally, for the predictive parametric strategy, we perform a linear regression of D on P and optimize the estimated conditional expectation of revenues.

First, we consider the revenue performance of each of these strategies. We plot the corresponding out-of-sample revenues, $R(\hat{p}_n)$, along with optimal revenue $R(p^*)$, in Figure 3-2(a). The example shows that a predictive approach, whether parametric or not, can potentially leave much on the table in terms of revenues. In contrast to the predictive approach, the prescriptive non-parametric approach converges to optimum, in agreement with Theorem 3.16. On the other hand, the prescriptive parametric approach offers better out-of-sample performance for small samples.

Next, we apply our hypothesis test for revenue optimality. We plot the frequencies of rejecting a pricing strategy as significantly suboptimal at a significance of 0.1 in Figure 3-2(b). We see that with sufficient data, the test can distinguish those pricing strategies that generate suboptimal revenues (i.e. solely predictive strategies) from those that cannot be distinguished from optimal for all prescriptive intents and purposes. In particular, it takes about a thousand data points before the test has the desired significance of 0.1 (i.e. p^* is rejected no more than 10% of the time).

Example 3.23 (Auto Loan Rate Optimization)**.** Consider again the case of Example 3.2. In Besbes et al. (2010), the authors consider whether a parametric model suffices for the problem of fixed pricing within various customer segments of loan applicants. The segments are defined in terms of three factors:

1. FICO score: (690, 715] (range 1) or (715, 740] (range 2),
2. Loan term in months: ≤ 36 (class 1), (36, 48] (class 2), (48, 60] (class 3), or > 60 (class 4).
3. Season: first half of data (half 1) or second half (half 2).

Customers with FICO scores outside of (690, 740] are not considered (see ibid. for reasoning). Term classes 2 and 4 are not considered either, but we consider these here. The authors use a per-unit revenue function $r(p) = r - 2\%$. As in Example 3.11, let us use $Y \in \{1, \ldots, 16\}$ to denote membership in a segment. Within each segment, the authors approach is to estimate (either parametrically or non-parametrically) the conditional expectation of demand given price (same as conditional probability since demand is binary) and to optimize per-unit revenue times this conditional expectation. Using a test that compares the parametric and non-parametric approaches, they conclude that a parametric model suffices.

We consider the very same problem again here, paying closer attention to the observational nature of the data. In Example 3.11 we argued that even within each segment, the data cannot be treated as experimental (i.e. satisfying weak ignorability with respect to segment alone) and therefore that purely predictive approaches will not produce the true PRF. We noted, however, that this is moot if, in the end of the day, revenues generated by such approaches cannot be distinguished from optimal. We now use our hypothesis test to determine whether this is the case. We also apply the test to our parametric prescriptive approach from Section 3.5.2 to determine whether a parametric approach suffices to achieve revenues that cannot be distinguished from optimal. For the predictive approaches, we use the same methods as used in Besbes et al. (2010): kernel regression with the Gaussian kernel (non-parametric) and logistic regression (parametric). For our parametric prescriptive approach we fit a log-normal model for price via linear regression on X as our GLM for price, i.e.,

$$\big(\log(P)\big|X = x\big) \sim \mathcal{N}\left(\beta_0 + \beta_1^T x,\, \sigma\right),$$

and we fit a logistic regression for demand that is linear in price and quadratic in generalized propensity score, i.e.,

$$\mathbb{P}\left(D = 1\middle|P = p,\, Q = q\right) = \frac{1}{1 + e^{-\alpha_0 - \alpha_1 p - \alpha_2 q - \alpha_3 q^2}}.$$

We let X consist of FICO score, the loan amount, the loan term, whether the car is new or used, whether the loan refinancing, and if so what was the previous rate (otherwise 0). In our experience, each of these covariates has direct impact on the interest rate quoted to applicants – and each can arguably impact the decision of the

Table 3.1: Testing Revenue Optimality in the Auto Loan Rate Optimization Example

		FICO range 1 (690, 715]		FICO range 2 (715, 740]		
		Half 1	Half 2	Half 1	Half 2	
n		1359	732	1386	781	Term class 1
p-values	Prescriptive, Parametric	0.15	0.16	0.50	0.012 (*)	
	Predictive, Parametric	0.030 (*)	0.049 (*)	0.54	1.1 e −5 (***)	
	Predictive, Non-parametric	0.0012 (**)	4.3 e −4 (***)	0.082	7.6 e −10 (***)	
n		1394	832	1327	690	Term class 2
p-values	Prescriptive, Parametric	0.11	0.19	0.054	0.89	
	Predictive, Parametric	0.0020 (**)	0.11	0.035 (*)	0.047 (*)	
	Predictive, Non-parametric	3.3 e −5 (***)	0.0040 (**)	9.2 e −5 (***)	0.047 (*)	
n		4495	3147	3803	2865	Term class 3
p-values	Prescriptive, Parametric	9.8 e −8 (***)	0.39	0.070	0.082	
	Predictive, Parametric	1.9 e −9 (***)	0.43	0.0021 (**)	0.011 (*)	
	Predictive, Non-parametric	2.8 e −5 (***)	0.08	5.3 e −8 (***)	0.0034 (**)	
n		2347	1506	1834	1206	Term class 4
p-values	Prescriptive, Parametric	0.0070 (**)	0.63	0.28	0.42	
	Predictive, Parametric	0.026 (*)	0.18	0.24	0.23	
	Predictive, Non-parametric	3.0 e −7 (***)	8.0 e −15 (***)	0.17	0.0054 (**)	

Note: (*) denotes reject H_0 at significance $p < 0.05$, (**) at $p < 0.01$, and (***) at $p < 0.001$, while gray values denote $p \geq 0.05$.

The data clearly distinguishes from optimal the revenues generated by predictive approaches, rejecting the null hypothesis in all but 3 segments (non-parametric) or 6 segments (parametric). The prescriptive approach, albeit parametric, on the other hand, generates revenues that cannot be distinguished from optimal in all but 3 segments (in each of which, the other approaches also failed the test).

applicant to accept any one rate. At the same time, this summarizes all relevant data provided and thus encapsulates all customer-specific information that could have gone into a rate quote decision. Therefore, we conclude that weak ignorability reasonably holds with respect to X, while it is likely to fail with respect to any subset of X.

We run the test within each of the 16 customer segments and report the resulting p-values in Table 3.1. The results overwhelmingly support the case that a predictive approach is insufficient and leaves revenue on the table – the data clearly distinguishes the revenues generated by these approaches from optimal in all but 3 segments (non-parametric) or 6 segments (parametric). Our prescriptive parametric approach, on the other hand, passes the test in all but 3 segments (in each of which, the other approaches also failed the test). We note that in the same segments in which the analysis in Besbes et al. (2010) suggested that logistic regression on P is sufficient to estimate the PRF for pricing purposes, our findings show that it is in fact insufficient if one is concerned with optimizing revenues in a prescriptive setting. On the other hand, a parametric approach that address the observational nature of the data and the prescriptive nature of the problem seems to suffice for pricing in most cases, generating revenues that the data cannot be outright distinguish from optimal. In practice, it is known that parametric models, even if misspecified, can be helpful in extracting useful models from smaller datasets. Our findings confirm this and agree with the conclusions of Besbes et al. (2010) but not entirely with the methods.

3.7 Extensions

We explored the predictive-prescriptive dichotomy through a particular, simple pricing problem and presented one treatment of the issue. The ideas, however, extend and relate to a wider range of operational problems and statistical techniques. We discuss these extensions in this section.

3.7.1 Customized Pricing

Up to now, we have considered the problem of assigning a single price based on data, but in this data prices were potentially set based partially on observed covariates. The sole customization considered was of the form of discrete binning and splitting of the data. The single-pricing problem offered the clearest parallel to existing work and, as a building block of price revenue optimization, provided a framework in which to study the distinction between prediction and prescription.

In this section we briefly expand our scope to the problem of customized pricing, where each price can be dependent on customer characteristics. In the full-information case, the problem is the same as the single-price problem and simply involves adjusting one's definition of the relevant population. In the data-driven case, however, the difference is that data from heterogeneous customers must be used to estimate the PRF for a particular customer either because customer characteristics are defined using continuous quantities or because there are many segments. The standing assumption will still be, as before, weak ignorability with respect to X.

Let us consider the fully customized pricing problem where price should be customized on the basis of the full set of covariates X. That is, we are interested in the problem of choosing a unit price $p(x) \in \mathcal{P} \subset \mathbb{R}_+$ for each customer characteristic x. The hypothetical price optimization problem we would then like to solve can be expressed as follows:

$$p^*(x) \in \arg\max_{p \in \mathcal{P}} \left\{ R(p, y) := r(p)\mathbb{E}\left[D(p)\middle|X = x\right] \right\}. \tag{3.8}$$

Assuming customers arrive from some stationary distribution, our expected revenue generated from a measurable pricing strategy $p(x)$ is

$$R(p(\cdot)) = \mathbb{E}\left[r(p(X))D(p(X))\right].$$

Note that we have

$$R(p^*(\cdot)) = \mathbb{E}\left[\max_{p \in \mathcal{P}} r(p)\mathbb{E}\left[D(p)\middle|X\right]\right].$$

One approach to customized pricing is to estimate the customized PRF, i.e. $\mathbb{E}\left[D(p)\middle|X = x\right]$, and optimize customized pricing with respect to it. To estimate the customized PRF, we can rely on weak ignorability. The proof of Theorem 3.9 argued that under weak ignorability, $\mathbb{E}\left[D(p)\middle|X = x\right] = \mathbb{E}\left[D\middle|P = p,\, X = x\right]$ so that the customized PRF is given by regressing D on P and X. Since we customize the price based on X we do not average over it as we have before. (If customization were done on the basis of a subset $Y = s(X)$ of the features, we would need to average over the conditional distribution of X given Y, which would require additional estimation.)

Non-parametric approach.

As before, we can use Nadaraya-Watson kernel regression to come up with a consistent non-parametric estimate for $\mathbb{E}\left[r(P)D\middle|P = p, X = x\right]$. This is exactly what we did in deriving $\overline{R}_n(p, x)$ in eq. (3.5) in Section 3.5.1, which leads to the following non-parametric data-driven customized price prescription

$$\overline{p}_n(x) \in \arg\max_{p \in \mathcal{P}} \overline{R}_n(p, x). \tag{3.9}$$

Estimating the marginal distribution of X by the empirical distribution, a corresponding non-parametric estimate of the expected revenue generated from a pricing strategy $p(\cdot)$ is

$$\overline{R}_n(p(\cdot)) = \frac{1}{n}\sum_{i=1}^{n} \overline{R}_n(p(X_i), X_i). \tag{3.10}$$

A hypothesis test.

As before, it can be argued that for pricing purposes, the fit of a customized demand model is irrelevant insofar as the model leads to revenues that cannot be discerned from optimal. We can develop a hypothesis test akin to that of Section 3.6, which

asses whether this is the case in the customized pricing case.

Let us consider any customized pricing scheme $\hat{p}_n(\cdot)$ and let $\hat{p}(\cdot)$ be its large-sample equivalent. That is, let us assume the following.

Assumption 3.24 (Convergent Customized Pricing Strategy)**.** For some fixed $\hat{p}(\cdot)$,

$$\sup_x |\hat{p}_n(x) - \hat{p}(x)| = O_p(1/\sqrt{n}).$$

We are interested in testing the hypotheses

$$H_0 : R(p^*(\cdot)) = R(\hat{p}(\cdot)),$$
$$H_1 : R(p^*(\cdot)) > R(\hat{p}(\cdot)).$$

Again, the impediment to testing this is that $R(p(\cdot))$, $p^*(\cdot)$, and $\hat{p}(\cdot)$ are all unknown; were they known, we could compute $\kappa = R(p^*(\cdot)) - R(\hat{p}(\cdot))$ and compare it to 0. A test statistic that proxies κ that uses our non-parametric estimates from the last section is

$$\kappa_n = \overline{R}_n(\overline{p}_n(\cdot)) - \overline{R}_n(\hat{p}_n(\cdot)) = \frac{1}{n}\sum_{i=1}^{n} \left(\overline{R}_n(\overline{p}_n(X_i), X_i) - \overline{R}_n(\hat{p}_n(X_i), X_i)\right).$$

As before, it can be shown that under appropriate conditions, our statistic diverges under H_1 and converges in distribution under H_0, with

$$\tilde{A}_n = \frac{1}{n}\sum_{i=1}^{n} \left(\overline{R}_n(\overline{p}_n(X_i), X_i) - \overline{R}_n(p^*(X_i), X_i)\right)$$

being the dominating term. Therefore, an approximate rejection threshold for κ_n can be gleaned from the bootstrap estimates

$$\tilde{A}_n^{(b)} = \frac{1}{n}\sum_{i=1}^{n} \left(\overline{R}_n^{(b)}(\overline{p}_n^{(b)}(X_i), X_i) - \overline{R}_n^{(b)}(\overline{p}_n(X_i), X_i)\right).$$

The details are beyond the scope of this chapter.

3.7.2 Related Problems

The distinction between prediction and prescription extends to other data-driven prescriptive problems that leverage observational data. Within pricing, this includes data-driven multi-product or inventory-constrained pricing (Oren et al. 1984, Gallego and Van Ryzin 1994, Elmaghraby and Keskinocak 2003). More generally, the distinction is relevant whenever the effect of the decision being optimized on the objective is unknown and needs to be estimated from experimental data, including, e.g., newsvendor models where on-hand inventory affect demand (Lee et al. 2012). The same concepts, such as the central role of weak ignorability, extend to these problems.

Problems where the effect of the decision on the objective is known a priori are unaffected by this distinction. Consider, for example, the classic stochastic optimization problem,

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right],$$

with decision z and random disturbance Y. Here, knowledge of the cost structure $c(z; y)$ encapsulates the decision's effect on the objective and, in a data-driven case, all that needs to be estimated from the data is the distribution of Y – e.g. via the sample average approximation (Kleywegt et al. 2002b) or robust sample average approximation (Bertsimas et al. 2014b). The same is true of the more intricate conditional stochastic optimization problem studied in (Bertsimas and Kallus 2015b),

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y) \middle| X = x\right],$$

where X represents predictive observations. The cost function gives the prescriptive effect of the decision z and is assumed known. The effect of the predictive features X are to be estimated, but because they are not being optimized but only observed, their causal effect is not of interest. This breaks down if the assumption of a known cost function breaks down – then the prescriptive effect of, instead of prediction based on, a decision needs to be estimated.

3.7.3 Related Statistical Methods

The Neyman-Rubin potential outcome framework is not the only framework used to describe causal relationships, although it is largely the most popular in statistics. We find that potential outcome notation fits well with the problem we explore here and also matches with familiar notation already used in other work in operations research, such as Lee et al. (2012), where the notation $G(q, \cdot)$ is used for the distribution of demand when the initial on-hand inventory is set to q.

Other notable frameworks for causality include structural equation models (SEM; cf. Goldberger (1972)), popular in econometrics, and Pearl's framework of causal Bayesian networks and do-calculus (cf. Pearl (2000)), popular in epidemiology. The SEM framework may well be applied to the problem but we choose not to use it because of its need for a priori models, the common restriction to linear relationships, incompatible notation, and the less clear question of model-free identifiability, which here drives our pricing solution and the nonparametric test for revenue optimality.

Pearl's framework in some senses encompasses both potential outcomes and SEM. Its dependence on directed acyclic graph (DAG) models to describe a priori causal relationships, however, makes it potentially too unwieldy for application to the problem herein and its notation and extensive nomenclature too complex for a succinct presentation. In effect, a causal DAG, correctly specified, can specify the correct subset of the covariates X that should be included in order to achieve weak ignorability. The standard practice in applications of the Neyman-Rubin framework is generally to condition on all observed covariates X that are potentially relevant (cf. Rubin (2009)), but one can come up with contrived scenarios where the inclusion of a

covariate in such conditioning can (asymptotically) bias causal estimates (cf. Shrier (2009), Pearl (2009)). Because these scenarios are usually restricted to self-selection via hidden factors, rather than selection by a manager based on available data, the relevance of such concerns to the problems explored herein is limited.

3.8 Conclusions

We studied the distinction between prediction and prescription in the context of data-driven pricing and showed that a naive but common predictive approach leaves money on the table whereas a theoretically-sound prescriptive approaches performs well in practice. We demonstrated this using a novel statistical test applied to data from an automotive loan provider. Our results indicated that in many circumstances parametric approaches suffice, but only when they take into account the prescriptive nature of the problem. We highlight the predictive-prescriptive dichotomy using the lens of a simple pricing problem, but the ideas extend to many other operational decision-making problems where the effect of a control is unknown and our decision-making process is informed by observational data.

Part II

The Interface Between Hypothesis Testing and Optimization Under Uncertainty

Chapter 4

Robust SAA

Sample average approximation (SAA) is a widely popular approach to data-driven decision-making under uncertainty. Under mild assumptions, SAA is both tractable and enjoys strong asymptotic performance guarantees. Similar guarantees, however, do not typically hold in finite samples. In this chapter, we propose a modification of SAA, which we term Robust SAA, which retains SAA's tractability and asymptotic properties and, additionally, enjoys strong finite-sample performance guarantees. The key to our method is linking SAA, distributionally robust optimization, and hypothesis testing of goodness-of-fit. Beyond Robust SAA, this connection provides a unified perspective enabling us to characterize the finite sample and asymptotic guarantees of various other data-driven procedures that are based upon distributionally robust optimization. We present examples from inventory management and portfolio allocation, and demonstrate numerically that our approach outperforms other data-driven approaches in these applications.

4.1 Introduction

In this chapter, we treat the stochastic optimization problem

$$z_{\text{stoch}} = \min_{x \in X} \mathbb{E}_F[c(x;\xi)], \tag{4.1}$$

where $c(x,\xi)$ is a given cost function depending on a random vector ξ following distribution F and a decision variable $x \in X \subseteq \mathbb{R}^n$. This is a widely used modeling paradigm in operations research, encompassing a number of applications Shapiro and Andrzej (2003), Birge and Louveaux (2011).

In real-world applications, however, the distribution F is unknown. Rather, we are given data $\xi^1, \ldots, \xi^N$, which are typically assumed to be drawn IID from F. The most common approach in these settings is the sample average approximation (SAA). SAA approximates the true, unknown distribution F by the empirical distribution $\hat{F}_N$, which places $1/N$ mass at each of the data points. In particular, the SAA approach

approximates (4.1) by the problem

$$z_{\text{SAA}} = \min_{x \in X} \frac{1}{N} \sum_{j=1}^{N} c(x, \xi^j). \tag{4.2}$$

Variants of the SAA approach in this and other contexts are ubiquitous throughout operations research, often used tacitly without necessarily being referred to by this name.

Under mild conditions on the cost function $c(x; \xi)$ and the sampling process, SAA enjoys two important properties:

Asymptotic Convergence: As the number of data points $N \to \infty$, both the optimal value z_{SAA} of (4.2) and an optimal solution x_{SAA} converge to the optimal value z_{stoch} of (4.1) and an optimal solution x_{stoch} almost surely (e.g. Kleywegt et al. (2002a), King and Wets (1991)).

Tractability: Finding the optimal value of and an optimal solution to (4.2) is computationally tractable (e.g. Birge and Louveaux (2011)).

In our opinion, these two features – asymptotic convergence and tractability – underly SAA's practical success in data-driven settings. Similar performance guarantees, however, do not hold for SAA for finite N, except in certain special cases (e.g. Kleywegt et al. (2002a), Levi et al. (2012)).

In this chapter, we propose a novel approach to (4.1) in data-driven settings which we term *Robust SAA*. Robust SAA inherits SAA's favorable asymptotic convergence and tractability. Unlike SAA, however, Robust SAA enjoys a strong *finite sample* performance guarantee for a wide class of optimization problems. The key idea of Robust SAA is to approximate (4.1) by a particular data-driven, distributionally robust optimization problem using ideas from statistical hypothesis testing.

More specifically, a distributionally robust optimization (DRO) problem is

$$\overline{z} = \min_{x \in X} \mathcal{C}(x, \mathcal{F}), \tag{4.3}$$

$$\text{where } \mathcal{C}(x, \mathcal{F}) = \sup_{F_0 \in \mathcal{F}} \mathbb{E}_{F_0}[c(x; \xi)], \tag{4.4}$$

where $\mathcal{F}$ is a set of potential distributions for ξ. We call such a set a *distributional uncertainty set* or DUS in what follows. Initial research (see literature review below) focused on DUSs $\mathcal{F}$ specified by fixing the first few moments of a distribution or other structural features, but did not explicitly consider the data-driven setting. Recently, the authors of Calafiore and El Ghaoui (2006), Delage and Ye (2010) took an important step forward proposing data-driven DRO formulations in which the DUS $\mathcal{F}$ is a function of the data, i.e., $\mathcal{F} = \mathcal{F}(\xi^1, \ldots, \xi^N)$, and showing that (4.3) remains tractable. Loosely speaking, their DUSs consist of distributions whose first few moments are close to the sample moments of the data. The authors show how to tailor these DUS so that for any $0 \leq \alpha \leq 1$, the probability (with respect to data sample) that the true (unknown) distribution $F \in \mathcal{F}(\xi^1, \ldots, \xi^N)$ is at least $1 - \alpha$.

Consequently, solutions to (4.3) based on these DUSs enjoy a distinct, finite-sample guarantee:

Finite-Sample Performance Guarantee: With probability at least $1 - \alpha$ with respect to the data sampling process, for any optimal solution $\overline{x}$ to (4.3), $\overline{z} \geq \mathbb{E}_F[c(\overline{x}, \xi)]$, where the expectation is taken with respect to the true, unknown distribution F.

In contrast to SAA, however, the methods of Calafiore and El Ghaoui (2006), Delage and Ye (2010) do not generally enjoy asymptotic convergence. (We make this claim precise Section 4.4.3).

Our approach, Robust SAA, is a particular type of data-driven DRO. Unlike existing approaches, however, our DUSs are not defined in terms of the sample moments of the data, but rather are specified as the confidence region of a goodness-of-fit (GoF) hypothesis test. Intuitively, our DUSs consist of all distributions which are "small" perturbations of the empirical distribution – hence motivating the name Robust SAA – where the precise notion of "small" is determined by the choice of GoF test. Different GoF tests yields different DUSs with different computational and statistical properties.

We prove that like other data-driven DRO proposals, Robust SAA also satisfies a finite-sample performance guarantee. Moreover, we prove that for a wide-range of cost functions $c(x; \xi)$, Robust SAA can be reformulated as a single-level convex optimization problem suitable for off-the-shelf solvers and is tractable theoretically and practically. Unlike other data-driven DRO proposals, however – and this is key – we prove that Robust SAA also satisfies an asymptotic convergence property similar to SAA. In other words, Robust SAA combines the strengths of both the classical SAA and data-driven DRO. Computational experiments in inventory management and portfolio allocation confirm that these properties translate into higher quality solutions for these applications in both small and large sample contexts.

In addition to proposing Robust SAA as an approach to addressing (4.1) in data-driven settings, we highlight a connection between GoF hypothesis testing and data-driven DRO more generally. Specifically, we show that any DUS that enjoys a finite-sample performance guarantee, including the methods of Calafiore and El Ghaoui (2006), Delage and Ye (2010), can be recast as the confidence region of *some* statistical a hypothesis test. Thus, hypothesis testing provides a unified viewpoint. Adopting this viewpoint, we characterize the finite-sample and asymptotic performance of DROs in terms of certain statistical properties of the underlying hypothesis test, namely significance and consistency. This characterization highlights an important, new connection between statistics and data-driven DRO. From a practical perspective, our results allow us to describe which DUSs are best suited to certain applications, providing important modeling guidance to practitioners. Moreover, this connection motivates the use of well-established statistical procedures like bootstrapping in the DRO context. Numerical experimentation confirms that these procedures can significantly improve upon existing algorithms and techniques.

To summarize our contributions:

1. We propose a new approach to optimization in data-driven settings, termed Robust SAA, which enjoys both finite sample and asymptotic performance guarantees for a wide-class of problems.

2. We develop new connections between SAA, DRO and statistical hypothesis testing. In particular, we characterize the finite-sample and asymptotic performance of data-driven DROs in terms of certain statistical properties of a corresponding hypothesis test, namely its significance and consistency.

3. Leveraging the above characterization, we shed new light on the finite sample and asymptotic performance of existing DRO methods and Robust SAA. In particular, we provide practical guidelines on designing appropriate DRO formulations for specific applications.

4. We prove that Robust SAA yields tractable optimization problems that are solvable in polynomial time for a wide class of cost functions. Moreover, for many cases of interest, including two-stage convex optimization with linear recourse, Robust SAA leads a single-level convex optimization formulations that can be solved using off-the-shelf software for linear or second-order optimization.

5. Through numerical experiments in inventory management and portfolio allocation, we illustrate that Robust SAA leads to better performance guarantees than existing data-driven DRO approaches and has performance similar to classical SAA in the large-sample regime.

6. Finally, we show how Robust SAA can be used to obtain approximations to the "price of data" – the price one would be willing to pay in a data-driven setting for additional data.

The remainder of this chapter is structured as follows. We next provide a brief literature review and describe the model setup. In Section 4.2, we illustrate the fundamental connection between DRO and the confidence regions of GoF tests and explicitly describe Robust SAA. Section 4.3 connects the significance of the hypothesis test to the finite-sample performance of a DRO. Section 4.4 connects the consistency of the hypothesis test to the asymptotic performance of the DRO. Section 4.5 proves that for the tests we consider, Robust SAA leads to a tractable optimization problem for many choices of cost function. Finally, Section 4.7 presents an empirical study and Section 4.8 concludes. All proofs except that for Theorem 4.3 are in the appendix.

4.1.1 Literature Review

DRO was first proposed by the author in Scarf (1958), where $\mathcal{F}$ is taken to be the set of distributions with a given mean and covariance in a specific inventory context. DRO has since received much attention in the literature, with many authors focusing on DUSs $\mathcal{F}$ defined by fixing the first few moments of the distribution Birge and Wets (1986), Prékopa (1995), Popescu (2007), Bertsimas and Popescu (2005), although some also consider other structural information such as unimodality Dupačová (1987).

In Wiesemann et al. (2013), the authors characterized the computational tractability of (4.3) for a wide range of DUSs $\mathcal{F}$ by connecting tractability to the geometry of $\mathcal{F}$.

As mentioned, in Delage and Ye (2010), Calafiore and El Ghaoui (2006), the authors extended DRO to the data-driven setting. In Calafiore and El Ghaoui (2006), the authors studied chance constraints, but their results can easily be cast in the DRO setting. Both papers focus on tractability and the finite-sample guarantee of the resulting formulation. Neither considers asymptotic performance. In Jiang and Guan (2013), the authors also propose a data-driven approach to chance constraints, but do not discuss either finite sample guarantees or asymptotic convergence. Using our hypothesis testing viewpoint, we are able to complement these existing works and establish a unified set of conditions under which the above methods will enjoy a finite-sample guarantee and/or be asymptotically convergent.

Recently, several other authors have considered hypothesis testing in certain, specific optimization contexts. In Bertsimas et al. (2013), the authors show how hypothesis tests can be used to construct uncertainty sets for robust, linear optimization problems, and establish a finite-sample guarantee that is similar in spirit to our own. They do not, however, consider asymptotic performance. In Ben-Tal et al. (2013), the authors consider robust optimization problems described by phi-divergences over uncertain, discrete probability distributions with finite support and provides tractable reformulations of these constraints. The authors mention that these divergences are related to GoF tests for discrete distributions, but do not explicitly explore asymptotic convergence of their approach to the full-information optimum or the case of continuous distributions. Similarly, in Klabjan et al. (2013), the authors study a stochastic lot-sizing problem under discrete distributional uncertainty described by Pearson's χ^2 GoF test and develop a dynamic programming approach to this particular problem. The authors establish conditions for asymptotic convergence for this problem but do not discuss finite sample guarantees.

By contrast, we provide a systematic study of GoF testing and data-driven DRO. By connecting these problems with the existing statistics literature, we provide a unified treatment of both discrete and continuous distributions, finite-sample guarantees, and asymptotic convergence. Moreover, our results apply in a general optimization context for a large variety of cost functions. We consider this viewpoint to both unify and extend these previous results.

4.1.2 Setup

In the remainder, we denote the support of ξ by Ξ. We assume $\Xi \subseteq \mathbb{R}^d$ is closed, and denote by $\mathcal{P}(\Xi)$ the set of (Borel) probability distributions over Ξ. For any probability distribution $F_0 \in \mathcal{P}(\Xi)$, $F_0(A)$ denotes the probability of the event $\xi \in A$. To streamline the notation when $d = 1$, we let $F_0(t) = F_0((-\infty, t])$. When $d > 1$ we also denote by $F_{0,i}$ the univariate marginal distribution of the i^{th} component, i.e., $F_{0,i}(A) = F_0(\{\xi : \xi_i \in A\})$. We assume that $X \subseteq \mathbb{R}^{d_x}$ is closed and that for any $x \in X$, $\mathbb{E}_F[c(x;\xi)] < \infty$ with respect to the true distribution, i.e., the objective function of the full-information stochastic problem (4.1) is well-defined.

When Ξ is unbounded, (4.3) may not admit an optimal solution. (We will see non-

pathological examples of this behavior in Section 4.3.2.) To be completely formal in what follows, we first establish sufficient conditions for the existence of an optimal solution. First, recall the definition of equicontinuity:

Definition 4.1. A set of functions $\mathcal{H} = \{h : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}\}$ is *equicontinuous* if for any given $x \in \mathbb{R}^{m_1}$ and $\epsilon > 0$ there exists $\delta > 0$ such that for all $h \in \mathcal{H}$, $\|h(x) - h(x')\| < \epsilon$ for any x' with $\|x - x'\| < \delta$.

In words, equicontinuity generalizes the usual definition of continuity of a function to continuity of a set of functions.

Our sufficient conditions constitute an analogue of the classical Weierstrass Theorem for deterministic optimization (see, e.g., Bertsekas (1999), pg. 669):

Theorem 4.2. *Suppose there exists $x_0 \in X$ such that $\mathcal{C}(x_0; \mathcal{F}) < \infty$ and that $c(x; \xi)$ is equicontinuous in x over all $\xi \in \Xi$. If either X is compact or $\lim_{||x|| \to \infty} c(x; \xi) = \infty$ for any ξ, then the optimal value $\overline{z}$ of (4.3) is finite and is achieved at some $\overline{x} \in X$.*

4.2 Goodness-of-Fit Testing and Robust SAA

In this section, we provide a brief review of GoF testing as it relates to Robust SAA. For a more complete treatment, including the wider range of testing cases possible, we refer the reader to D'Agostino and Stephens (1986), Thas (2009).

Given IID data $\xi^1, \ldots, \xi^N$ and a distribution F_0, a GoF test considers the hypothesis

$$H_0 : \text{ The data } \xi^1, \ldots, \xi^N \text{ were drawn from } F_0 \tag{4.5}$$

and rejects it if there is sufficient evidence against it, otherwise making no particular conclusion. A test is said to be of significance level α if the probability of incorrectly rejecting H_0 is at most α.

A typical test specifies a statistic

$$S_N = S_N(F_0, \xi^1, \ldots, \xi^N)$$

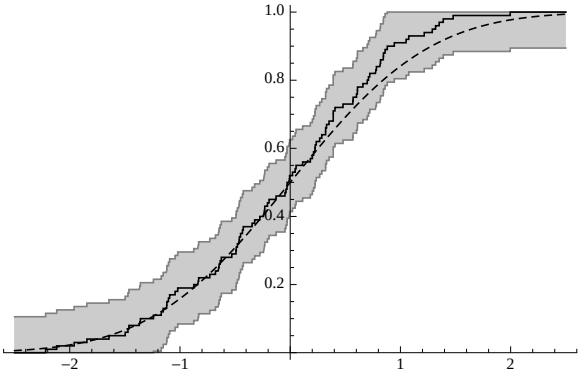
that depends on the data $\xi^1, \ldots, \xi^N$ and hypothesis F_0 and also specifies a threshold $Q_{S_N}(\alpha)$ that depends only on α. The test rejects H_0 if $S_N > Q_{S_N}(\alpha)$.

The threshold $Q_{S_N}(\alpha)$ is usually the $(1-\alpha)^{\text{th}}$ quantile of the distribution of S_N under the assumption that the data have the distribution F_0. For some tests, $Q_{S_N}(\alpha)$ can be computed (or bounded) in closed-form. More generally, $Q_{S_N}(\alpha)$ can be approximated numerically using techniques like the bootstrap, in particular when it may depend on F_0 (see Efron and Tibshirani (1993)). Implementations of bootstrap procedures for computing thresholds $Q_{S_N}(\alpha)$ are available in many popular software packages, e.g., the function *one.boot* in the [R] package *simpleboot*.

One example is the Kolmogorov-Smirnov (KS) test for univariate distributions. The KS test uses the statistic

$$D_N = \max_{i=1,\ldots,N} \left\{ \max \left\{ \frac{i}{N} - F_0(\xi^{(i)}), F_0(\xi^{(i)}) - \frac{i-1}{N} \right\} \right\}.$$

Figure 4-1: The Confidence Region of the Kolmogorov-Smirnov Test



Note: The significance in this example is 20%. The dashed curve is the true cumulative distribution function, that of a standard normal. The solid curve is the empirical cumulative distribution function having observed 100 draws from the true distribution. The confidence region contains all distributions with cumulative distribution functions that take values inside the grey region.

Tables for $Q_{D_N}(\alpha)$ are widely available (see e.g. D'Agostino and Stephens (1986), Stephens (1970)).

The set of all distributions F_0 that pass a test is called the *confidence region* of the test and is denoted by

$$\mathcal{F}^{\alpha}_{S_N}(\xi^1,\ldots,\xi^N) = \left\{F_0 \in \mathcal{P}(\Xi) : S_N(F_0,\xi^1,\ldots,\xi^N) \leq Q_{S_N}(\alpha)\right\}. \tag{4.6}$$

As an example, Figure 4-1 illustrates the confidence region of the KS test. Observe that by construction, the confidence region of a test with significance level α is a DUS which contains the true, unknown distribution F with probability at least $1-\alpha$.

4.2.1 The Robust SAA Approach

Given data $\xi^1,\ldots,\xi^N$, the Robust SAA approach involves the following steps:

1. Choose a significance level $0 < \alpha < 1$ and goodness-of-fit test at level α independently of the data.
2. Let $\mathcal{F} = \mathcal{F}_N(\xi^1,\ldots,\xi^N)$ be the confidence region of the test.
3. Solve
$$\overline{z} = \arg\min_{x\in X} \sup_{F_0\in\mathcal{F}_N(\xi^1,\ldots,\xi^N)} \mathbb{E}_{F_0}[c(x;\xi)]$$
and let $\overline{x}$ be an optimal solution.

Section 4.5 illustrates how to solve the optimization problem in the last step for various choices of goodness-of-fit test and classes of cost functions.

4.2.2 Connections to Existing Methods

As observed in Section 4.2, we can use a GoF test at significance level α to construct a DUS that contains the true distribution F with probability at least $1-\alpha$ via its confidence region. It is possible to do the reverse as well. Given a data-driven DUS $\mathcal{F}_N(\xi^1,\ldots,\xi^N)$ that contains the true distribution with probability at least $1-\alpha$ with respect to the sampling distribution, we can construct a GoF test with significance level α that rejects the hypothesis (4.5) whenever $F_0 \notin \mathcal{F}_N(\xi^1,\ldots,\xi^N)$. This is often termed the duality between hypothesis tests and confidence regions (see for example §9.3 of Rice (2007)).

This reverse construction can be applied to existing data-driven DUSs in the literature such as Delage and Ye (2010), Calafiore and El Ghaoui (2006) to construct their corresponding hypothesis tests. In this way, hypothesis testing provides a common ground on which to understand and compare the methods.

In particular, the hypothesis tests corresponding to the DUSs of Delage and Ye (2010), Calafiore and El Ghaoui (2006) test only the first moments of the true distribution (cf. Section 4.4.3). By contrast, we will for the most part focus on tests (and corresponding confidence regions) that test the entire distribution, not just the first two moments. This feature is key to achieving both finite-sample and asymptotic guarantees.

4.3 Finite-Sample Performance Guarantees

We first study the implication of a test's significance on the finite-sample performance of Robust SAA. Let us define the following random variables expressible as functions of the data $\xi^1,\ldots,\xi^N$:

The DRO solution:	$\overline{x} \in \arg\min_{x\in X} \sup_{F_0\in\mathcal{F}_N(\xi^1,\ldots,\xi^N)} \mathbb{E}_{F_0}\left[c(x;\xi)\right].$
The DRO value:	$\overline{z} = \min_{x\in X} \sup_{F_0\in\mathcal{F}_N(\xi^1,\ldots,\xi^N)} \mathbb{E}_{F_0}\left[c(x;\xi)\right].$
The true cost of the DRO solution:	$z = \mathbb{E}_F\left[c(\overline{x};\xi)\middle\vert\xi^1,\ldots,\xi^N\right].$

The following is an immediate consequence of significance.

Theorem 4.3. *If $\mathcal{F}_N(\xi^1,\ldots,\xi^N)$ is the confidence region of a valid GoF test at significance α, then, with respect to the data sampling process,*

$$\mathbb{P}\left(\overline{z} \geq z\right) \geq 1-\alpha.$$

Proof. Suppose $F \in \mathcal{F}_N$. Then $\sup_{F_0\in\mathcal{F}_N} \mathbb{E}_{F_0}[c(x;\xi)] \geq \mathbb{E}_F[c(x;\xi)]$ for any $x \in X$.

Therefore, we have $\overline{z} \geq z$. In terms of probabilities, this implication yields,

$$\mathbb{P}\left(\overline{z} \geq z\right) \geq \mathbb{P}\left(F \in \mathcal{F}_N\right) \geq 1 - \alpha.ineq$$

□

This makes explicit the connection between the statistical property of significance of a test with the objective performance of the corresponding Robust SAA decision in the full-information stochastic optimization problem.

Next we review the particular GoF tests we will employ.

4.3.1 Tests for Distributions with Known Discrete Support

When ξ has known finite support $\Xi = \{\hat{\xi}^1, \ldots, \hat{\xi}^n\}$ there are two popular tests of GoF: Pearson's χ^2 test and the G-test (see D'Agostino and Stephens (1986)). Let $p(j) = F(\{\hat{\xi}^j\})$, $p_0(j) = F_0(\{\hat{\xi}^j\})$, and $\hat{p}_N(j) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\xi^i = \hat{\xi}^j\right]$ be the true, hypothetical, and empirical probabilities of observing $\hat{\xi}^j$, respectively.

Pearson's χ^2 test uses the statistic

$$X_N = \left(\sum_{j=1}^{n} \frac{\left(p_0(j) - \hat{p}_N(j)\right)^2}{p_0(j)}\right)^{1/2},$$

whereas the G-test uses the statistic

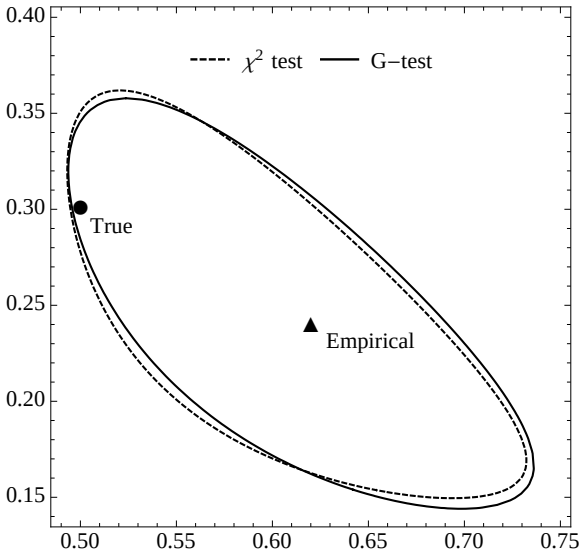
$$G_N = \left(2\sum_{j=1}^{N} \hat{p}_N(j) \log\left(\frac{\hat{p}_N(j)}{p_0(j)}\right)\right)^{1/2}.$$

The confidence regions of these take the form of (4.6) for S_N being either X_N or G_N. An illustration of these ($n = 3$, $N = 50$) is given in Figure 4-2. Intuitively, these can be seen as being generalized balls around the empirical distribution $\hat{p}_N$. The metric is given by the statistic S_N and the radius diminishes as $Q_{S_N}(\alpha) = O(N^{-1/2})$ (see D'Agostino and Stephens (1986)).

4.3.2 Tests for Univariate Distributions

Suppose ξ is a univariate continuous random variable that is known to have lower support greater than $\underline{\xi}$ and upper support less than $\overline{\xi}$. These bounds could possibly be infinite. The most commonly used GoF tests in this setting are the Kolmogorov-Smirnov (KS) test, the Kuiper test, the Cramér-von Mises (CvM) test, the Watson test, and the Anderson-Darling (AD) test. The KS (D_N), the Kuiper (V_N), the CvM (W_N), the Watson (U_N), and the AD (A_N) tests use the statistics (see D'Agostino

Figure 4-2: Distributional Uncertainty Sets for the Discrete Case



Note: The distributional uncertainty sets are visualized projected onto the first two components. The example has with $n = 3$, $\alpha = 0.8$, $N = 50$, $p = (0.5, 0.3, 0.2)$. The dot denotes the true frequencies p and the triangle the observed fractions $\hat{p}_{50}$.

and Stephens (1986))

$$
\begin{aligned}
D_N &= \max_{i=1,\ldots,N} \left\{ \max \left\{ \frac{i}{N} - F_0(\xi^{(i)}), F_0(\xi^{(i)}) - \frac{i-1}{N} \right\} \right\}, \\
V_N &= \max_{1 \le i \le N} \left(F_0(\xi^{(i)}) - \frac{i-1}{N} \right) + \max_{1 \le i \le N} \left(\frac{i}{N} - F_0(\xi^{(i)}) \right), \\
W_N &= \left(\frac{1}{12N^2} + \frac{1}{N} \sum_{i=1}^{N} \left(\frac{2i-1}{2N} - F_0(\xi^{(i)}) \right)^2 \right)^{1/2}, \\
U_N &= \left(W_N^2 - \left(\frac{1}{N} \sum_{i=1}^{N} F_0(\xi^{(i)}) - \frac{1}{2} \right)^2 \right)^{1/2}, \\
A_N &= \left(-1 - \sum_{i=1}^{N} \frac{2i-1}{N^2} \left(\log F_0(\xi^{(i)}) + \log(1 - F_0(\xi^{(N+1-i)})) \right) \right)^{1/2}.
\end{aligned}
\tag{4.7}
$$

We let $S_N \in \{D_N, W_N, A_N, V_N, U_N\}$ be any one of the above statistics and $Q_{S_N}(\alpha)$ the corresponding threshold. Tables for $Q_{S_N}(\alpha)$ are widely available (see D'Agostino and Stephens (1986), Stephens (1970)). Moreover, $Q_{S_N}(\alpha)$ can be computed by simulation as the $(1-\alpha)^{\text{th}}$ percentile of the distribution of S_N when $F_0(\xi^i)$ in (4.7) are replaced by IID uniform random variables on $[0, 1]$.

The confidence regions of these tests take the form of (4.6). Recall Figure 4-1 illustrated $\mathcal{F}^{\alpha}_{D_N}$. As in the discrete case, $\mathcal{F}^{\alpha}_{S_N}$ can also be seen as a generalized ball around the empirical distribution $\hat{F}_N$. Again, the radius diminish as $Q_{S_N}(\alpha) = O(N^{-1/2})$ (see D'Agostino and Stephens (1986)).

When $\underline{\xi}$ and $\overline{\xi}$ are finite, we take $\mathcal{F}^{\alpha}_{S_N}$ to be our DUS corresponding to these tests. When either $\underline{\xi}$ or $\overline{\xi}$ is infinite, however, $\overline{z}$ in (4.3) may also be infinite as seen in the following proposition.

Proposition 4.4. *Fix x, α, and $S_N \in \{D_N, W_N, A_N, V_N, U_N\}$. If $c(x;\xi)$ is continuous but unbounded on Ξ then $\mathcal{C}(x;\mathcal{F}^{\alpha}_{S_N}) = \infty$ almost surely.*

The conditions of Proposition 4.4 are typical in many applications. For example, in Wang et al. (2009), the authors briefly propose a data-driven DRO formulation of the newsvendor problem that is equivalent to our Robust SAA formulation using the KS test. Using Proposition 4.4, however, one can show that if the uncertain demand is supported on the positive real-line, the optimal value of this formulation is infinite. We will return to the data-driven newsvendor in Example 4.28 below.

Consequently, when either $\underline{\xi}$ or $\overline{\xi}$ is infinite, we will employ an alternative, non-standard, GoF test in Robust SAA. The confidence region of our proposed test will satisfy the conditions of Theorem 4.2, and, therefore, (4.3) will attain a finite, optimal solution.

Our proposed test combines one of the above GoF tests with a second test for a generalized moment of the distribution. Specifically, fix any function $\phi : \Xi \to \mathbb{R}_+$ such that $\mathbb{E}_F[\phi(\xi)] < \infty$ and $|c(x_0,\xi)| = O(\phi(\xi))$ for some $x_0 \in X$. For a fixed μ_0, consider the null hypothesis

$$H_0' : \mathbb{E}_F[\phi(\xi)] = \mu_0. \tag{4.8}$$

There are many possible hypothesis tests for (4.8). Any of these tests can be used as the second test in our proposal. For concreteness, we focus on a test with rejects (4.8) if

$$M_N = \left| \mu_0 - \frac{1}{N}\sum_{i=1}^{N} \phi(\xi^i) \right| > Q_{M_N}(\alpha). \tag{4.9}$$

As mentioned in Section 4.2, the threshold $Q_{M_N}(\alpha)$ can be computed via the bootstrap. In our numerical experiments in Section 4.7.1 we approximate $Q_{M_N}(\alpha)$ as $\hat{\sigma}_N Q_{T_{N-1}}(\alpha/2)/\sqrt{N}$ where $Q_{T_{N-1}}(\alpha/2)$ is the $(1-\alpha/2)^{\text{th}}$ quantile of the Student-T distribution with $N-1$ degrees of freedom and $\hat{\sigma}^2_N$ is the sample variance of $\phi(\xi)$. This is a widely used approximation in statistics which is known to perform well in similar applications Rice (2007).

Given $0 < \alpha_1, \alpha_2 < 1$, combining S_N and (4.9), we propose the following GoF test:

$$\text{Reject } F_0 \text{ if either } S_N > Q_{S_N}(\alpha_1) \text{ or } \left| \mathbb{E}_{F_0}[\phi(\xi)] - \frac{1}{N}\sum_{i=1}^{N} \phi(\xi^i) \right| > Q_{M_N}(\alpha_2).$$

By the union bound, the probability of incorrectly rejecting F_0 is at most

$$\mathbb{P}(S_N > Q_{S_N}(\alpha_1)) + \mathbb{P}\left(\left|\mathbb{E}_{F_0}[\phi(\xi)] - \frac{1}{N}\sum_{i=1}^{N}\phi(\xi^i)\right| > Q_{M_N}(\alpha_2)\right) \leq \alpha_1 + \alpha_2.$$

Thus, our proposed test has significance level $\alpha_1 + \alpha_2$.

The confidence region of the above test is given by the intersection of the confidence region of our original goodness-of-fit test and the confidence region of our test for (4.8):

$$\begin{aligned}\mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2} &= \mathcal{F}_{S_N}^{\alpha_1} \cap \mathcal{F}_{M_N}^{\alpha_2} \\ &= \left\{F_0 \in \mathcal{P}(\Xi) : S_N \leq Q_{S_N}(\alpha_1),\ \left|\mathbb{E}_{F_0}[\phi(\xi)] - \frac{1}{N}\sum_{i=1}^{N}\phi(\xi^i)\right| \leq Q_{M_N}(\alpha_2)\right\}.\end{aligned} \tag{4.10}$$

Observe that since $|c(x_0;\xi)| = O(\phi(\xi))$, i.e., $\exists \nu, \eta$ such that $|c(x_0;\xi)| \leq \nu + \eta\phi(\xi)$, we have

$$\begin{aligned}\mathcal{C}(x_0; \mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2}) &= \sup_{F_0 \in \mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2}} \mathbb{E}_{F_0}[c(x_0;\xi)] \leq \nu + \eta \sup_{F_0 \in \mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2}} \mathbb{E}_{F_0}[\phi(\xi)] \\ &\leq \nu + \frac{\eta}{N}\sum_{i=1}^{N}\phi(\xi^i) + \eta Q_{M_N}(\alpha_2) < \infty,\end{aligned}$$

so that unlike $\mathcal{F}_{S_N}^{\alpha}$, our new confidence region $\mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2}$ does indeed satisfy the conditions of Theorem 4.2, even if $\underline{\xi}$ or $\overline{\xi}$ are infinite.

4.3.3 Tests for Multivariate Distributions

In this section, we propose two different tests for the case $d \geq 2$. The first is a standard test based on testing marginal distributions. The second is a new test we propose that tests the full joint distribution.

Testing Marginal Distributions

Let $\alpha_1, \ldots, \alpha_d > 0$ be given such that $\alpha = \alpha_1 + \cdots + \alpha_d < 1$. Consider the test for the hypothesis $F = F_0$ that proceeds by testing the hypotheses $F_i = F_{0,i}$ for each $i = 1, \ldots, d$ by applying a test from the previous two sections at significance level α_i to the sample $\xi_i^1, \ldots, \xi_i^N$ and rejecting F_0 if any of these fail. The corresponding confidence region is

$$\mathcal{F}_{\text{marginals}}^{\alpha} = \left\{F_0 \in \mathcal{P}(\Xi) \,:\, F_{0,i} \in \mathcal{F}_i^{\alpha_i}\left(\xi_i^1, \ldots, \xi_i^N\right)\ \forall i = 1, \ldots, d\right\},$$

where $\mathcal{F}_i^{\alpha_i}$ denotes the confidence region corresponding to the test applied on the i^{th} component. By the union bound we have

$$\mathbb{P}\left(F \notin \mathcal{F}^{\alpha}_{\text{marginals}}\right) \leq \sum_{i=1}^{d} \mathbb{P}\left(F_i \notin \mathcal{F}_i^{\alpha_i}\right) \leq \sum_{i=1}^{d} \alpha_i = \alpha,$$

so the test has significance α.

Testing Linear-Convex Ordering

In this section, we first provide some background on the linear-convex ordering (LCX) of random vectors first proposed in Scarsini (1998), and then use LCX to motivate a new GoF test for multivariate distributions. To the best of our knowledge, we are the first to propose GoF tests based on LCX.

Given two multivariate distributions G and G', we write

$$G \preceq_{\text{LCX}} G' \iff \mathbb{E}_G[\phi(a^T\xi)] \leq \mathbb{E}_{G'}[\phi(a^T\xi)] \ \ \forall\ a \in \mathbb{R}^d \text{ and convex functions } \phi \tag{4.11}$$

$$\iff \mathbb{E}_G[\max\{a^T\xi - b, 0\}] \leq \mathbb{E}_{G'}[\max\{a^T\xi - b, 0\}] \ \ \forall\ |a_1| + \ldots + |a_d| + |b| \leq 1, \tag{4.12}$$

where the second equivalence follows from Theorem. 3.A.1 of Shaked and Shanthikumar (2007).

Our interest in LCX stems from the following result from Scarsini (1998). Assuming $\mathbb{E}_G[\|\xi\|_2^2] < \infty$,

$$\mathbb{E}_G[\|\xi\|_2^2] \geq \mathbb{E}_{G'}[\|\xi\|_2^2] \text{ and } G \preceq_{\text{LCX}} G' \implies G = G'. \tag{4.13}$$

Equation (4.13) motivates our GoF test. Intuitively, the key idea of our test is that if $F \neq F_0$, i.e., we should reject F_0, then by (4.13) either $\mathbb{E}_{F_0}[\|\xi\|_2^2] < \mathbb{E}_F[\|\xi\|_2^2]$ or $F_0 \not\preceq_{LCX} F$. Thus, we can create a GoF test by testing for each of these cases separately.

More precisely, for a fixed μ_0, first consider the hypothesis

$$H_0' : \mathbb{E}_F[\|\xi\|_2^2] = \mu_0. \tag{4.14}$$

As in Section 4.3.2, there are many possible tests for (4.14). For concreteness, we focus on a one-tailed test which rejects (4.14) if $R_N = \frac{1}{N}\sum_{i=1}^{N} \|\xi^i\|_2^2 - \mu_0 > Q_{R_N}(\alpha)$. where $Q_{R_N}(\alpha)$ is a threshold which can be computed by bootstrapping.

Next, define the statistic

$$C_N(F_0) = \sup_{|a_1|+\cdots+|a_d|+|b|\leq 1} \left(\mathbb{E}_{F_0}[\max\{a^T\xi - b, 0\}] - \frac{1}{N}\sum_{i=1}^{N}[\max\{a^T\xi^i - b, 0\}] \right).$$

From (4.12), $C_N(F_0) \leq 0 \iff F_0 \preceq_{\text{LCX}} \hat{F}_N$. (Recall that $\hat{F}_N$ denotes the empirical

distribution.)

Finally, combining these pieces and given $0 < \alpha_1, \alpha_2 < 1$, our LCX-based GoF test is

$$\text{Reject } F_0 \text{ if either } C_N(F_0) > Q_{C_N}(\alpha_1) \text{ or } \mathbb{E}_{F_0}[\|\xi\|_2^2] < \frac{1}{N}\sum_{i=1}^{N}\|\xi\|_2^2 - Q_{R_N}(\alpha_2). \tag{4.15}$$

The threshold $Q_{C_N}(\alpha_1)$ can be computed by bootstrapping or exactly bounded explicitly. See Section C.1 in the appendix for further discussion. In our numerical experiments in Section 4.7.3, we use bootstrapped thresholds.

From a union bound we have that the LCX-based GoF test (4.15) has significance level $\alpha_1 + \alpha_2$. The confidence region of the LCX-based GoF test is

$$\mathcal{F}_{C_N,R_N}^{\alpha_1,\alpha_2} = \left\{ F_0 \in \mathcal{P}(\Xi) : C_N(F_0) \le Q_{C_N}(\alpha_1),\ \mathbb{E}_{F_0}[\|\xi\|_2^2] \ge \frac{1}{N}\sum_{i=1}^{N}\|\xi^2\|_2^2 - Q_{R_N}(\alpha_2) \right\}. \tag{4.16}$$

4.4 Convergence

Had we known the true distribution F we would solve problem (4.1). As we gather more data, we know more and more about F. Therefore, it is clearly desirable that our decisions converge to the optimal solutions of (4.1).

In this section, we study the relationship between the GoF test underlying an application of Robust SAA and convergence properties of the Robust SAA optimal values $\overline{z}$ and solutions $\overline{x}$. Recall from Section 4.2.2 that since many existing DRO formulations can be recast as confidence regions of hypothesis tests, our analysis will simultaneously also allow us to study the convergence properties of these methods as well.

The convergence conditions we seek are

i. Convergence of objective function:

$$\begin{aligned} &\mathcal{C}(x;\mathcal{F}_N) \to \mathbb{E}_F[c(x;\xi)] \\ &\text{uniformly over any compact subset of } X, \end{aligned} \tag{4.17}$$

ii. Convergence of optimal values:

$$\min_{x\in X} \mathcal{C}(x;\mathcal{F}_N) \to \min_{x\in X} \mathbb{E}_F[c(x;\xi)], \tag{4.18}$$

iii. Convergence of optimal solutions:

$$\begin{aligned} &\text{Every sequence } x_N \in \arg\min_{x\in X} \mathcal{C}(x;\mathcal{F}_N) \text{ has at least one limit} \\ &\text{point, and all of its limit points are in } \arg\min_{x\in X} \mathbb{E}_F[c(x;\xi)], \end{aligned} \tag{4.19}$$

all holding almost surely (a.s.). The key to these will be a restricted form of statistical consistency that we term uniform consistency.

4.4.1 Uniform Consistency and Convergence of Optimal Solutions

In statistics, *consistency of a test* (see Def. 4.5 below) is a well-studied property that a GoF test may exhibit. In this section, we define a new property of GoF tests that we call *uniform consistency*. Uniform consistency is a strictly stronger property than consistency, in the sense that every uniformly consistent test is consistent, but some consistent tests are not uniformly consistent. More importantly, we will prove that uniform consistency of the underlying GoF test tightly characterizes when conditions (4.17)-(4.19) hold. In particular, we show that when X and Ξ are bounded, uniform consistency of the underlying test implies conditions (4.17)-(4.19) for any cost function $c(x, \xi)$ which is equicontinuous in x, and if the test is not uniformly consistent, then there exist cost functions (equicontinuous in x) for which conditions (4.17)-(4.19) do not hold. When X or Ξ are unbounded, the same conclusions hold for all cost functions which are equicontinuous in x and satisfy an additional, mild, regularity condition. (See Theorem 4.12 for a precise statement.) In other words, we can characterize the convergence of Robust SAA and other data-driven, DRO formulations by studying if their underlying GoF test is uniformly consistent. In our opinion, these results highlight a new, fundamental connection between statistics and data-driven optimization. We will use this result to assess the strength of various DRO formulations for certain applications in what follows.

First, we recall the definition of consistency of a GoF test (cf. entry for *consistent test* in Dodge (2006)):

Definition 4.5. A GoF test is *consistent* if, for every $F_0 \neq F$, the probability of rejecting F_0 approaches 1 as $N \to \infty$.

Observe

Proposition 4.6. *If a test is consistent, then any $F_0 \neq F$ is a.s. rejected infinitely often (i.o.) as $N \to \infty$.*

Proof. $$\mathbb{P}(F_0 \text{ rejected i.o.}) = \mathbb{P}\left(\limsup_{N\to\infty} \{F_0 \notin \mathcal{F}_N\}\right) \geq \limsup_{N\to\infty} \mathbb{P}\left(F_0 \notin \mathcal{F}_N\right) = 1,$$

where the first inequality follows from Fatou's Lemma, and the second since the test is consistent. □

Consistency describes the test's behavior with respect to a single, fixed distribution F_0. In particular, the conclusion of Proposition 4.6 holds only when we consider the same, fixed distribution F_0 for each N. We would like to extend consistency to describe the test's behavior with respect to many alternatives F_0 simultaneously.

Motivated by an alternate definition of local uniform convergence,[1] we define uniform consistency by requiring that a condition similar to the conclusion of Proposition 4.6 hold for almost every sequence of distributions:

Definition 4.7. A GoF test is *uniformly consistent* if, a.s., every sequence F_N that does not converge weakly to F is rejected i.o.

The requirement that F_N does not converge weakly to F parallels the requirement that $F_0 \neq F$.

Uniform consistency is a strictly stronger requirement than consistency.

Proposition 4.8. *If a test is uniformly consistent, then it is consistent. Moreover, there exist tests which are consistent, but not uniformly consistent.*

Uniform consistency is the key property for the convergence of Robust SAA. Besides uniform consistency, convergence will be contingent on three assumptions.

Assumption 4.9. $c(x;\xi)$ is equicontinuous in x over all $\xi \in \Xi$.

Assumption 4.10. X is closed and either

a. X is bounded or

b. $\lim_{||x||\to\infty} c(x;\xi) = \infty$ uniformly over ξ in some $D \subseteq \Xi$ with $F(D) > 0$ and $\liminf_{||x||\to\infty} \inf_{\xi\notin D} c(x;\xi) > -\infty$.

Assumption 4.11. Either

a. Ξ is bounded or

b. $\exists \phi : \Xi \to \mathbb{R}_+$ such that $\sup_{F_0\in\mathcal{F}_N} \left| E_{F_0}\phi(\xi) - \frac{1}{N}\sum_{i=1}^{N}\phi(\xi^i) \right| \to 0$ almost surely and $c(x;\xi) = O(\phi(\xi))$ for each $x \in X$.

Assumptions 4.9 and 4.10 are only slightly stronger than those required for the existence of an optimal solution in Theorem 4.2. The second portion of Assumption 4.10b is trivially satisfied by cost functions which are bounded from below. Finally, observe that in the case that Ξ is unbounded, our proposed DUS in (4.10) satisfies Assumption 4.11b by construction.

Under these assumptions, the following theorem provides a tight characterization of convergence.

Theorem 4.12. *Assumptions 4.9, 4.10, and 4.11 imply conditions* (4.17)-(4.19) *hold a.s. if and only if $\mathcal{F}_N$ is the confidence region of a uniformly consistent test.*

[1]Recall: A sequence of functions $g_n : \mathbb{R}^{m_1} \mapsto \mathbb{R}^{m_2}$ converges locally uniformly to a continuous function $g : \mathbb{R}^{m_1} \mapsto \mathbb{R}^{m_2}$ if and only if for any convergent sequence $x_n \to x$ we have that $g_n(x_n) \to g(x)$.

Table 4.1: Summary of Convergence Results

GoF test	Support	Uniformly consistent	(4.17)-(4.19) hold a.s. for any $c(x;\xi)$ that is equicontinuous in x	separable (4.20)	as in (4.25)
χ^2 and G-test	Finite	Yes	Yes	Yes	Yes
KS, Kuiper, CvM, Watson, and AD tests	Univariate	Yes	Yes	Yes	Yes
Test of marginals using the above tests	Multivariate	No	No*	Yes	Yes
LCX-based test	Multivariate	Yes	Yes	Yes	Yes
Tests implied by DUSs of Delage and Ye (2010), Calafiore and El Ghaoui (2006)	Multivariate	No	No*	No*	Yes

Note: * denotes the result is tight in the sense that there are examples in this class that do not converge.

Thus, in one direction, we can guarantee convergence (i.e., conditions (4.17)-(4.19) hold a.s.) if Assumptions 4.9, 4.10, and 4.11 are satisfied and we use a uniformly consistent test in applying Robust SAA. In the other direction, if the test is not uniformly consistent, there will exist instances satisfying Assumptions 4.9, 4.10, and 4.11 for which convergence fails.

Some of the GoF tests in Section 4.3 are not consistent, and therefore, cannot be uniformly consistent. By Theorem 4.12, DROs built from these tests cannot exhibit asymptotic convergence for all cost functions. One might argue, then, that that these DRO formulations should be avoided in modeling and applications in favor of DROs based on uniformly consistent tests.

In most applications, however, we are not concerned with asymptotic convergence *for all* cost functions, but rather only for the *given* cost function $c(x, \xi)$. It may happen a DRO may exhibit asymptotic convergence for this particular cost function, even when its DUS is given by the confidence region of an inconsistent test. (We will see an example of this behavior with the multi-item newsvendor problem in Section 4.7.2.)

To better understand when this convergence may occur despite the fact that the test is not consistent, we introduce a more relaxed form of uniform consistency.

Definition 4.13. Given $c(x;\xi)$, we say that F_N c-converges to F if $\mathbb{E}_{F_N}[c(x;\xi)] \to \mathbb{E}_F[c(x;\xi)]$ for all $x \in X$.

Definition 4.14. A test is *c-consistent* if, a.s., every sequence F_N that does not c-converge to F is rejected i.o.

This notion may potentially be weaker than consistency, but is sufficient for convergence for a given instance as shown below.

Theorem 4.15. *Suppose Assumptions 4.9 and 4.11 hold and that $\mathcal{F}_N$ always contains the empirical distribution. If $\mathcal{F}_N$ is the confidence region of a c-consistent test, then conditions* (4.17)-(4.19) *hold a.s.*

In the next sections we will explore the consistency of the various tests introduced in Section 4.3. We summarize our results in Table 4.1.

4.4.2 Tests for Distributions with Discrete or Univariate Support

All of the classical tests we considered in Section 4.3 are uniformly consistent.

Theorem 4.16. *The χ^2 and G-tests are uniformly consistent.*

Theorem 4.17. *The KS, Kuiper, CvM, Watson, and AD tests are uniformly consistent.*

4.4.3 Tests for Multivariate Distributions

Testing Marginal Distributions

We first claim that the test of marginals is not consistent. Indeed, consider a multivariate distribution $F_0 \neq F$ which has the same marginal distributions, but a different joint distribution. By construction, the probability of rejecting F_0 is at most α for all N, and hence does not converge to 1. Since the test of marginals is not consistent, it cannot be uniformly consistent.

We next show that the test is, however, c-consistent whenever the cost is separable over the components of ξ.

Proposition 4.18. *Suppose $c(x;\xi)$ is separable over the components of ξ, that is, can be written as*

$$c(x;\xi) = \sum_{i=1}^{d} c_i(x;\xi_i), \tag{4.20}$$

and Assumptions 4.9, 4.10, and 4.11 hold for each $c_i(x;\xi_i)$. Then, the test of marginals is c-consistent if each univariate test is uniformly consistent.

That is to say, if the cost can be separated as in (4.20), applying the tests from Section 4.3.2 to the marginals is sufficient to guarantee convergence.

It is important to note that some cost functions may only be separable after a transformation of the data, potentially into a space of different dimension. If that is the case, we may transform ξ and apply the tests to the transformed components in order to achieve convergence.

Tests Implied by DUSs of Delage and Ye (2010), Calafiore and El Ghaoui (2006)

The DUS of Delage and Ye (2010) has the form

$$\mathcal{F}^{\alpha}_{\mathrm{DY},N} = \left\{ F_0 \in \mathcal{P}(\Xi) \ : \ \begin{array}{c} \left(\mathbb{E}_{F_0}[\xi] - \hat{\mu}_N\right)^T \hat{\Sigma}_N^{-1} \left(\mathbb{E}_{F_0}[\xi] - \hat{\mu}_N\right) \leq \gamma_{1,N}(\alpha), \\ \gamma_{3,N}(\alpha)\hat{\Sigma}_N \preceq \mathbb{E}_{F_0}[(\xi - \hat{\mu}_N)(\xi - \hat{\mu}_N)^T] \preceq \gamma_{2,N}(\alpha)\hat{\Sigma}_N \end{array} \right\} \tag{4.21}$$

$$\text{where } \hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \xi^i, \ \hat{\Sigma}_N = \frac{1}{N}\sum_{i=1}^{N} \left(\xi^i - \hat{\mu}_N\right)\left(\xi^i - \hat{\mu}_N\right)^T. \tag{4.22}$$

The thresholds $\gamma_{1,N}(\alpha)$, $\gamma_{2,N}(\alpha)\,\gamma_{3,N}(\alpha)$ are developed therein (for Ξ bounded) so as to guarantee a significance of α (in our GoF interpretation) and, in particular, have the property that

$$0 \leq \gamma_{1,N}(\alpha) \to 0, \ 1 \leq \gamma_{2,N}(\alpha) \to 1, \ 1 \geq \gamma_{3,N}(\alpha) \to 1. \tag{4.23}$$

The DUS of Calafiore and El Ghaoui (2006) has the form

$$\mathcal{F}^{\alpha}_{\text{CEG},N} = \left\{ F_0 \in \mathcal{P}(\Xi) \ : \ \begin{array}{c} ||\mathbb{E}_{F_0}[\xi] - \hat{\mu}_N||_2 \leq \gamma_{1,N}(\alpha), \\ \left|\left|\mathbb{E}_{F_0}\left[(\xi - \mathbb{E}_{F_0}[\xi])\,(\xi - \mathbb{E}_{F_0}[\xi])^T\right] - \hat{\Sigma}_N\right|\right|_{\text{F}} \leq \gamma_{2,N}(\alpha) \end{array} \right\}.$$

The thresholds $\gamma_{1,N}(\alpha)$, $\gamma_{2,N}(\alpha)$ are developed in Shawe-Taylor and Cristianini (2003) (for Ξ bounded) so as to guarantee a significance of α and with the property that

$$0 \leq \gamma_{1,N}(\alpha) \to 0,\ 0 \leq \gamma_{2,N}(\alpha) \to 0. \tag{4.24}$$

The GoF tests implied by these DUSs consider only the first two moments of a distribution (mean and covariance). Therefore, the probability of rejecting a multivariate distribution different from the true one but with the same mean and covariance is by construction never more than α, instead of converging to 1. That is, these tests are not consistent and therefore they are not uniformly consistent. We next provide conditions on the cost function that guarantee that the tests are nonetheless c-consistent.

Proposition 4.19. *Suppose $c(x;\xi)$ can be written as*

$$c(x;\xi) = c_0(x) + \sum_{i=1}^{d} c_i(x)\xi_i + \sum_{i=1}^{d}\sum_{j=1}^{i} c_{ij}(x)\xi_i\xi_j \tag{4.25}$$

and that $\mathbb{E}_F[\xi_i\xi_j]$ exists. Then, the tests with confidence regions given by $\mathcal{F}^{\alpha}_{DY,N}$ or $\mathcal{F}^{\alpha}_{CEG,N}$ are c-consistent.

Note that because we may transform the data to include components for each pairwise multiplication, the conditions on the cost function in Proposition 4.19 are stronger than those in Proposition 4.18. In particular, in one dimension, separability is trivially always true whereas the decomposition (4.25) is clearly not.

Testing Linear-Convex Ordering

The previous two multivariate GoF tests were neither consistent, nor uniformly consistent. By contrast,

Proposition 4.20. *The LCX-based test is consistent.*

Proposition 4.21. *Suppose Ξ is bounded. Then the LCX-based test is uniformly consistent.*

It is an open question whether the LCX-based test is uniformly consistent – in addition to being consistent – for unbounded Ξ. We conjecture that it is. Moreover, in our numerical experiments involving the LCX test, we have observed convergence of the Robust SAA solutions to the full-information optimum even when Ξ is unbounded. (See Section 4.7.3 for an example.)

4.5 Tractability

In this section, we characterize conditions under which problem (4.3) is theoretically tractable, i.e., can be solved with a polynomial-time algorithm. Additionally, we are interested in cases where (4.3) is practically tractable, i.e., can be solved using off-the-shelf linear or second-order cone optimization solvers. In the case of one problem – the newsvendor problem – we show that Robust SAA using the KS test admits a closed-form solution.

4.5.1 Tests for Distributions with Known Discrete Support

We begin this section with a reformulation of (4.3) as a single-level optimization problem for $\mathcal{F}^{\alpha}_{X_N}$ and $\mathcal{F}^{\alpha}_{G_N}$, from which tractability results will follow. The confidence regions of the discrete GoF tests we consider are a special case of those considered in Ben-Tal et al. (2013). As direct corollaries of the results therein we have the following:

Theorem 4.22. *Under the assumptions of Theorem 4.2, we have*

$$\begin{aligned}
\mathcal{C}\left(x; \mathcal{F}^{\alpha}_{X_N}\right) = \min_{r,s,t,c} \quad & r + \left(\left(Q_{X_N}(\alpha)\right)^2 + 2\right) s - 2\sum_{j=1}^{n} \hat{p}_N(j) t_j \\
s.t. \quad & r \in \mathbb{R},\, s \in \mathbb{R}_+,\, t \in \mathbb{R}^n,\, c \in \mathbb{R}^n \\
& s + r \geq c_j && \forall j = 1, \ldots, n \\
& (2s - c_j - r,\, 2t_j,\, c_j - r) \in C^3_{SOC} && \forall j = 1, \ldots, n \\
& c_j \geq c\left(x; \hat{\xi}^j\right) && \forall j = 1, \ldots, n
\end{aligned}$$

$$\begin{aligned}
\mathcal{C}\left(x; \mathcal{F}^{\alpha}_{G_N}\right) = \min_{r,s,t,c} \quad & r + \left(\frac{1}{2}\left(Q_{G_N}(\alpha)\right)^2 - 1\right) s - \sum_{j=1}^{n} \hat{p}_N(j) t_j \\
s.t. \quad & r \in \mathbb{R},\, s \in \mathbb{R}_+,\, t \in \mathbb{R}^n,\, c \in \mathbb{R}^n \\
& (t_j,\, s,\, r - c_j) \in C_{\mathrm{XC}} && \forall j = 1, \ldots, n \quad (4.26) \\
& c_j \geq c\left(x; \hat{\xi}^j\right) && \forall j = 1, \ldots, n
\end{aligned}$$

where $C^3_{SOC} = \left\{(x, y, z) \in \mathbb{R}^3 \;:\; x \geq \sqrt{y^2 + z^2}\right\}$ *is the three-dimensional second-order cone and* $C_{\mathrm{XC}} = \overline{\{(x, y, z) : ye^{x/y} \leq z,\, y > 0\}}$ *is the exponential cone.*

The DRO problem (4.3) is $\min_{x \in X} \mathcal{C}\left(x; \mathcal{F}\right)$. Therefore, for $\mathcal{F}^{\alpha}_{X_N}$ and $\mathcal{F}^{\alpha}_{G_N}$, (4.3) can formulated as a single-level optimization problem by augmenting the corresponding minimization problem above with the control variable $x \in X$. Note that apart from the constraints $x \in X$ and

$$c_j \geq c\left(x; \hat{\xi}^j\right), \tag{4.27}$$

the rest of the constraints, as seen in the problems in Theorem 4.22, are convex. The following result characterizes in general when solving these problems is tractable in a

theoretical sense.

Theorem 4.23. *Suppose that $X \subseteq \mathbb{R}^{d_x}$ is a closed convex set for which a weak separation oracle is given and that*

$$c\left(x; \hat{\xi}^j\right) = \max_{k=1,\ldots,K_j} c_{jk}(x)$$

where each $c_{jk}(x)$ is a convex function in x for which evaluation and subgradient oracles are given. Then, under the assumptions of Theorem 4.2, we can find an ϵ-optimal solution to (4.3) in the discrete case for $S_N = X_N$, G_N in time and oracle calls polynomial in $n, d_x, K_1, \ldots, K_n, \log(1/\epsilon)$.

For some problems the constraints $x \in X$ and (4.27) can also be conically formulated as the Example 4.24 below shows. In such a case, the DRO can be solved directly as a conic optimization problem. Optimization over the exponential cone – a non-symmetric cone — although theoretically tractable, is numerically challenging. Fortunately, the particular exponential cone constraints (4.26) can be recast as second-order cone constraints, albeit with constraint complexity growing in both n *and* N (see Lobo et al. (1998)).

Example 4.24. *Two-stage problem with linear recourse and a non-increasing, piecewise-linear convex disutility.* Consider the following problem

$$\begin{aligned} c(x; \hat{\xi}^j) &= \max_{k=1,\ldots,K} \left(\gamma_k R_j(x) + \beta_k\right), \quad \gamma_k \leq 0 \qquad (4.28)\\ \text{where} \quad R_j(x) &= \min_{y \in \mathbb{R}_+^{d_y}} f_j^T y \\ &\qquad \text{s.t. } A_j x + B_j y = b_j \\ X &= \{x \geq 0 : Hx = h\}. \end{aligned}$$

This problem was studied in a non-data-driven DRO settings in Žáčková (1966), Dupačová (1987), Bertsimas et al. (2010). To formulate (4.3), we may introduce variables $y \in \mathbb{R}_+^{n \times d_y}$ and replace (4.27) with

$$\begin{aligned} &c_j \geq \gamma_k \left(c^T x + f_j^T y_j\right) + \beta_k && \forall j = 1, \ldots, n, \ \forall k = 1, \ldots, K, \\ &A_j x + B_j y_j = b_j && \forall j = 1, \ldots, n. \end{aligned}$$

The resulting problem is then a second-order cone optimization problem for $\mathcal{F}^{\alpha}_{X_N}$ and $\mathcal{F}^{\alpha}_{G_N}$.

4.5.2 Tests for Univariate Distributions

We now consider the case where ξ is a general univariate random variable. We proceed by reformulating (4.3) as a single-level optimization problem by leveraging semi-infinite conic duality. This leads to corresponding tractability results. In the following we will use the notation $\xi^{(0)} = \underline{\xi}$ and $\xi^{(N+1)} = \overline{\xi}$.

The first observation is that the constraint $S_N(\zeta_1, \ldots, \zeta_N) \leq Q_{S_N}(\alpha)$ is convex in $\zeta_i = F_0(\xi^{(i)})$ and representable using *canonical cones.* By a canonical cone, we mean any cartesian product of the cones $\mathbb{R}^k$, $\{0\}$, $\mathbb{R}^k_+$ (positive orthant), C^k_{SOC} (second-order cone), and semidefinite cone. Optimization over canonical cones is tractable both theoretically and practically using state-of-the-art interior point algorithms Ben-Tal and Nemirovski (2001).

Theorem 4.25. *For each of* $S_N \in \{D_N, V_N, W_N, U_N, A_N\}$

$$S_N(\zeta_1, \ldots, \zeta_N) \leq Q_{S_N}(\alpha) \iff A_{S_N}\zeta - b_{S_N,\alpha} \in K_{S_N}$$

for convex cones K_{S_N}*, matrices* A_{S_N}*, and vectors* $b_{S_N,\alpha}$ *as follows:*

$$K_{D_N} = \mathbb{R}^{2N}_+, \quad b_{D_N,\alpha} = \begin{pmatrix} \frac{1}{N} - Q_{D_N}(\alpha) \\ \vdots \\ \frac{N}{N} - Q_{D_N}(\alpha) \\ -\frac{0}{N} - Q_{D_N}(\alpha) \\ \vdots \\ -\frac{N-1}{N} - Q_{D_N}(\alpha) \end{pmatrix}, \quad A_{D_N} = \begin{pmatrix} [I_N] \\ [-I_N] \end{pmatrix},$$

$$K_{V_N} = \left\{(x,y) \in \mathbb{R}^{2N} \ : \ \min_i x_i + \min_i y_i \geq 0\right\}, \quad b_{V_N,\alpha} = \begin{pmatrix} \frac{1}{N} - Q_{V_N}(\alpha)/2 \\ \vdots \\ \frac{N}{N} - Q_{V_N}(\alpha)/2 \\ -\frac{0}{N} - Q_{V_N}(\alpha)/2 \\ \vdots \\ -\frac{N-1}{N} - Q_{V_N}(\alpha)/2 \end{pmatrix},$$

$$A_{V_N} = \begin{pmatrix} [I_N] \\ [-I_N] \end{pmatrix},$$

$$K_{W_N} = C^{N+1}_{\text{SOC}}, \quad b_{W_N,\alpha} = \begin{pmatrix} \sqrt{N\left(Q_{W_N}(\alpha)\right)^2 - \frac{1}{2N}} \\ \frac{1}{2N} \\ \frac{3}{2N} \\ \vdots \\ \frac{2N-1}{2N} \end{pmatrix}, \quad A_{W_N} = \begin{pmatrix} 0 \cdots 0 \\ [I_N] \end{pmatrix},$$

$$K_{U_N} = C^{N+2}_{\text{SOC}}, \quad b_{U_N,\alpha} = \begin{pmatrix} \frac{-1}{2} + \left(\frac{N}{24} - \frac{N}{2}\left(Q_{U_N}(\alpha)\right)^2\right) \\ \frac{-1}{2} - \left(\frac{N}{24} - \frac{N}{2}\left(Q_{U_N}(\alpha)\right)^2\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$A_{U_N} = \begin{pmatrix} \frac{1-N}{2N} & \frac{3-N}{2N} & \cdots & \frac{N-1}{2N} \\ \frac{N-1}{2N} & \frac{N-3}{2N} & \cdots & \frac{1-N}{2N} \\ & [I_N - \frac{1}{N}E_N] & & \end{pmatrix},$$

$$K_{A_N} = \left\{ (z,x,y) \in \mathbb{R} \times \mathbb{R}^{2N}_+ \ : \ |z| \leq \prod_{i=1}^{N} (x_i y_i)^{\frac{2i-1}{2N^2}} \right\}, \quad b_{A_N,\alpha} = \begin{pmatrix} e^{-\left(Q_{A_N}(\alpha)\right)^2 - 1} \\ 0 \\ \vdots \\ 0 \\ -1 \\ \vdots \\ -1 \end{pmatrix},$$

$$A_{A_N} = \begin{pmatrix} 0 \cdots 0 \\ [I_N] \\ [-\tilde{I}_N] \end{pmatrix},$$

where I_N is the $N \times N$ identity matrix, $\tilde{I}_N$ is the skew identity matrix ($[\tilde{I}_N]_{ij} = \mathbb{I}[i = N - j]$), and E_N is the $N \times N$ matrix of all ones.

Note that the cones $K_{D_N}, K_{W_N}, K_{U_N}$ are canonical cones. The other cones can be expressed using canonical cones. The cone K_{V_N} is an orthogonal projection of an affine slice of $\mathbb{R}^{2n+2} \times \mathbb{R}^3_+$. The cone K_{A_N} is an orthogonal projection of an affine slice of the product of $2^{\lceil \log_2(2N^2) \rceil + 1} - 2 = O(N^2)$ three-dimensional second-order cones (see Lobo et al. (1998)). Therefore, the constraint $A_{S_N}\zeta - b_{S_N,\alpha} \in K_{S_N}$ can be expressed using canonical cones in each case.

Problem (4.3) is a two-level optimization problem. To formulate it as a single-level problem, we dualize the inner problem, $\mathcal{C}(x;\mathcal{F})$. For a cone $K \subseteq \mathbb{R}^k$, we use the notation K^* to denote the dual cone $K^* = \{y \in \mathbb{R}^k : y^T z \geq 0 \, \forall z \in K\}$. The following is a direct consequence of Proposition 3.4 of Shapiro (2001).[2]

Theorem 4.26. *Let $S_N \in \{D_N, W_N, A_N, V_N, U_N\}$. Under the assumptions of Theo-*

[2]The only nuance is that Proposition 3.4 of Shapiro (2001) requires a generalized Slater point. We use the empirical distribution function, $\hat{F}_N$, as the generalized Slater point in the space of distributions.

rem 4.2,

$$\begin{aligned}
\mathcal{C}\left(x;\mathcal{F}^{\alpha}_{S_N}\right) = \min_{r,c} \quad & b^T_{S_N,\alpha} r + c_{N+1} \\
\text{s.t} \quad & -r \in K^*_{S_N},\, c \in \mathbb{R}^{N+1} \\
& \left(A^T_{S_N} r\right)_i = c_i - c_{i+1} && \forall i = 1,\ldots,N \\
& c_i \geq \sup_{\xi \in (\xi^{(i-1)}, \xi^{(i)}]} c(x;\xi) && \forall i = 1,\ldots,N+1
\end{aligned} \tag{4.29}$$

$$\begin{aligned}
\mathcal{C}\left(x;\mathcal{F}^{\alpha_1,\alpha_2}_{S_N,M_N}\right) = \min_{r,t,s,c} \quad & b^T_{S_N,\alpha_1} r + c_{N+1} + \left(\hat{\mu} + Q^{\alpha_2}_{M_N}\right) t - \left(\hat{\mu} - Q^{\alpha_2}_{M_N}\right) s \\
\text{s.t} \quad & -r \in K^*_{S_N},\, t \geq 0,\, s \geq 0,\, c \in \mathbb{R}^{N+1} \\
& \left(A^T_{S_N} r\right)_i = c_i - c_{i+1} && \forall i = 1,\ldots,N \\
& c_i \geq \sup_{\xi \in (\xi^{(i-1)}, \xi^{(i)}]} (c(x;\xi) - (t-s)\phi(\xi)) && \forall i = 1,\ldots,N+1.
\end{aligned} \tag{4.30}$$

Note that the cones K_{D_N}, K_{W_N}, K_{U_N} are self-dual ($K^* = K$) and therefore the dual cones remain canonical cones. For K_{V_N} and K_{A_N}, the dual cones are

$$K^*_{V_N} = \left\{ (x,y) \in \mathbb{R}^{2N}_+ \;:\; \sum_{i=1}^N x_i = \sum_{i=1}^N y_i \right\}$$

$$K^*_{A_N} = \{(z,x,y) \;:\; (z/\gamma, x, y) \in K_{A_N}\} \text{ where } \gamma = \prod_{i=1}^d \left(\frac{2i-1}{2N^2}\right)^{\frac{2i-1}{N^2}},$$

and therefore they remain expressible using canonical cones.

Note that in the case of $\mathcal{F}^{\alpha}_{S_N}$, the worst-case distribution has discrete support on no more than $N+1$ points. This is because shifting probability mass inside the interval $(\xi^{(i-1)}, \xi^{(i)}]$ does not change any $F_0(\xi^{(i)})$. In the worst-case, all mass in the interval (if any) will be placed on the point in the interval with the largest cost (including the left endpoint in the limit).

The DRO problem (4.3) is $\min_{x \in X} \mathcal{C}\left(x;\mathcal{F}\right)$. Therefore, for $\mathcal{F}^{\alpha}_{S_N}$ and $\mathcal{F}^{\alpha_1,\alpha_2}_{S_N,M_N}$, (4.3) can formulated as a single-level optimization problem by augmenting the corresponding minimization problem above with the control variable $x \in X$. We next give general conditions that ensure the theoretical tractability of the problem.

Theorem 4.27. *Suppose that $X \subseteq \mathbb{R}^{d_x}$ is a closed convex set for which a weak separation oracle is given and that*

$$c\left(x;\xi\right) = \max_{k=1,\ldots,K} c_k(x,\xi) \tag{4.31}$$

where each $c_k(x;\xi)$ is convex in x for each ξ and continuous in ξ for each x and for which an oracle is given for the subgradient in x. If $\mathcal{F} = \mathcal{F}^{\alpha}_{S_N}$, suppose also that an

oracle is given for maximizing $c_k(x;\xi)$ over ξ in any closed (possibly infinite) interval for fixed x. If $\mathcal{F} = \mathcal{F}^{\alpha_1,\alpha_2}_{S_N,M_N}$, suppose also that an oracle is given for maximizing $c_k(x;\xi) + \eta\phi(\xi)$ over ξ in a closed interval for fixed x and $\eta \in \mathbb{R}$. Then, under the assumptions of Theorem 4.2, we can find an ϵ-optimal solution to (4.3) in time and oracle calls polynomial in $N, d_x, K, \log(1/\epsilon)$ for $\mathcal{F} = \mathcal{F}^{\alpha}_{S_N}$ or $\mathcal{F} = \mathcal{F}^{\alpha_1,\alpha_2}_{S_N,M_N}$.

As in the discrete case, when the constraints $x \in X$ and (4.29) (or, (4.30)) can be conically formulated, Theorem 4.26 provides an explicit single-level conic optimization formulation of the problem (4.3). In Examples 4.28, 4.30, and 4.31 below, we consider specific problems for which this is the case and study this formulation.

Example 4.28. *The newsvendor problem.* In the newsvendor problem, one orders in advance $x \geq 0$ units of a product to satisfy an unknown future demand for $\xi \geq 0$ units. Unmet demand is penalized by $b > 0$, representing either backlogging costs or lost profit. Left over units are penalized by $h > 0$, representing either holding costs or recycling costs. The cost function is therefore $c(x;\xi) = \max\{b(\xi - x), h(x - \xi)\}$, the lower support of ξ is $\underline{\xi} \geq 0$, and the space of controls is $X = \mathbb{R}_+$. In this case the constraints (4.29) for bounded-support case become

$$c_i \geq b(\xi^{(i)} - x), \quad c_i \geq h(x - \xi^{(i-1)}) \qquad \forall i = 1, \ldots, N+1$$

and $x \in X$ becomes $x \in \mathbb{R}_+$. In the unbounded case, we may use $\phi(\xi) = |\xi|$ in the construction of (4.10). Because $\underline{\xi} \geq 0$, we have $\phi(\xi) = \xi$. The constraints (4.30) then become

$$c_i \geq b(\xi^{(i)} - x) - (t-s)\xi^{(i)}, \quad c_i \geq h(x - \xi^{(i-1)}) - (t-s)\xi^{(i-1)} \quad \forall i = 1, \ldots, N+1$$

where the $(N+1)^{\text{th}}$ left constraint is equivalent to $b \leq t - s$ because $\xi^{(N+1)} = \infty$. Substituting these constraints in this way the DRO (4.3) becomes a conic optimization problem.

In the specific case of bounded support and $\mathcal{F} = \mathcal{F}^{\alpha}_{D_N}$ this reformulation yields a linear optimization problem, which admits a closed-form solution given next.

Proposition 4.29. *Suppose that $\Xi = [\underline{\xi}, \overline{\xi}]$ is compact, and N is large enough so that $Q_{D_N}(\alpha) < \frac{\min\{b,h\}}{b+h}$. Then, the the DRO (4.3) for the newsvendor problem with $\mathcal{F} = \mathcal{F}^{\alpha}_{D_N}$ admits the closed-form solution:*

$$\begin{aligned}
\overline{x} &= (1-\theta)\xi^{(i_{lo})} + \theta\xi^{(i_{hi})} \\
\overline{z} &= \frac{1}{N} \sum_{1 \leq i \leq i_{lo} \vee i_{hi} \leq i \leq N} c\left(\overline{x}; \xi^{(i)}\right) + Q_{D_N}(\alpha) c\left(\overline{x}; \underline{\xi}\right) + Q_{D_N}(\alpha) c\left(\overline{x}; \overline{\xi}\right) \\
&\quad - \left(\frac{\lceil N\left(\theta - Q_{D_N}(\alpha)\right) \rceil}{N} - \left(\theta - Q_{D_N}(\alpha)\right) \right) c\left(\overline{x}; \xi^{(i_{lo})}\right) \\
&\quad - \left(\left(\theta + Q_{D_N}(\alpha)\right) - \frac{\lfloor N\left(\theta + Q_{D_N}(\alpha)\right) \rfloor}{N} \right) c\left(\overline{x}; \xi^{(i_{hi})}\right)
\end{aligned}$$

where $\theta = b/(b+h)$, $i_{lo} = \lceil N(\theta - Q_{D_N}(\alpha)) \rceil$, and $i_{hi} = \lfloor N(\theta + Q_{D_N}(\alpha)) + 1 \rfloor$.

Importantly, this means that solving the Robust SAA newsvendor problem is no more difficult than solving the SAA newsvendor problem.

Example 4.30. *Max of bilinear functions.* More generally, we may consider cost functions of the form (4.31) with bilinear parts $c_k(x;\xi) = p_{k0} + p_{k1}^T x + p_{k2}\xi + \xi p_{k3}^T x$. In this case, (4.29) is equivalent to

$$c_i \geq p_{k0} + p_{k1}^T x + p_{k2}\xi^{(i-1)} + \xi^{(i-1)} p_{k3}^T x, \quad \forall i = 1,\ldots,N, \quad \forall k = 1,\ldots,K \tag{4.32}$$

$$c_i \geq p_{k0} + p_{k1}^T x + p_{k2}\xi^{(i)} + \xi^{(i)} p_{k3}^T x, \quad \forall i = 1,\ldots,N, \quad \forall k = 1,\ldots,K. \tag{4.33}$$

If the cost is fully linear, $p_{3k} = 0$ (as in the case of the newsvendor example), then (4.29) can be written in one linear inequality:

$$c_i \geq p_{k0} + p_{k1}^T x + \max\left\{p_{k2}\xi^{(i-1)}, p_{k2}\xi^{(i)}\right\} \quad \forall i = 1,\ldots,N, \quad \forall k = 1,\ldots,K. \tag{4.34}$$

For $\mathcal{F} = \mathcal{F}_{S_N,M_N}^{\alpha_1,\alpha_2}$ we may use $\phi(\xi) = |\xi|$ and simply add $\left|\xi^{(i-1)}\right|$ and $\left|\xi^{(i)}\right|$ to the left-hand sides of (4.32) and (4.33), respectively, or to the corresponding branches of the max in (4.34).

Example 4.31. *Two-stage problem.* Consider a two-stage problem similar to the one studied in Example 4.24:

$$c(x;\xi) = \max_{k=1,\ldots,K} \left(\gamma_k R(x;\xi) + \beta_k\right), \quad \gamma_k \leq 0 \tag{4.35}$$

$$\begin{aligned} \text{where} \quad R(x;\xi) = \min_{y \in \mathbb{R}_+^{d_y}} \; & (f + g\xi)^T y \\ \text{s.t. } & Ax + By = b + p\xi \\ X = \{x \geq 0 : & Hx = h\}. \end{aligned}$$

When only the right-hand-side vector is uncertain ($g = 0$), the recourse $R(x;\xi)$ is convex in ξ so that the supremum in (4.29) is taken at one of the endpoints and we may use a similar construction as in Example 4.30.

When only the cost vector is uncertain ($p = 0$), the recourse $R(x;\xi)$ is concave in ξ. By linear optimization duality we may reformulate (4.29) by introducing variables $R \in \mathbb{R}^{N+1}$, $y \in \mathbb{R}_+^{d_y \times (N+1)}$, $\eta \in \mathbb{R}_+^{N+1}$, $\theta \in \mathbb{R}_+^{N+1}$ and constraints

$$\begin{aligned} & c_i \geq \gamma_k c^T x + \gamma_k R_i + \beta_k && \forall i = 1,\ldots,N+1, \quad \forall i = k,\ldots,K \\ & \eta_i - \theta_i = f^T y_i, \; Ax + By_i \leq b && \forall i = 1,\ldots,N+1 \\ & R_i \geq g^T y_i + \xi^{(i)}\eta_i - \xi^{(i-1)}\theta_i && \forall i = 1,\ldots,N+1. \end{aligned}$$

4.5.3 Tests for Multivariate Distributions

Testing Marginal Distributions

Recall that when $c(x;\xi)$ is separable over the components of ξ, i.e.,

$$c(x;\xi) = \sum_{i=1}^{d} c_i(x;\xi_i),$$

Robust SAA converges for the test of marginals (cf. Section 4.4.3). We next show that Robust SAA is also tractable in this case. When $\mathcal{F} = \mathcal{F}^{\alpha}_{\text{marginals}}$ and costs are separable, (4.3) can be written as

$$\min_{x\in X} \sup_{F_0\in\mathcal{F}} \mathbb{E}_{F_0}[c(x;\xi)] = \min_{x\in X} \sum_{i=1}^{d} \sup_{F_{0,i}\in\mathcal{F}_i^{\alpha_i}} \mathbb{E}_{F_{0,i}}[c_i(x;\xi_i)].$$

Applying Theorems 4.22 and 4.26 separately to these d subproblems yields a single-level optimization problem. This problem is theoretically tractable when each subproblem satisfies the corresponding conditions in Theorems 4.23 and 4.27. Similarly, when each subproblem is of one of the forms treated in Examples 4.24, 4.28, 4.30, and 4.31, (4.3) can be formulated as a linear or second-order cone optimization problem.

Testing Linear-Convex Ordering

Next, we consider the case of the test based on LCX. For this section we restrict our attention to cost functions of the form

$$c(x;\xi) = \max_{k=1,\ldots,K} \left\{ p_{k0} + p_{k1}^T x + p_{k2}^T \xi + x^T P_k \xi \right\}. \tag{4.36}$$

The following result provides a semi-infinite linear optimization reformulation of (4.3) and a polynomial-time separation algorithm.

Theorem 4.32. *Suppose that we can express $c(x;\xi)$ as in (4.36). Suppose moreover that $X = \left\{x \in \mathbb{R}^{d_x} : x \geq 0,\, Hx = h\right\}$ with $h \in \mathbb{R}^{d'}$ and that $\Xi = \mathbb{R}^d$. Under the assumptions of Theorem 4.2, the optimal value of (4.3) for $\mathcal{F} = \mathcal{F}^{\alpha_1,\alpha_2}_{C_N,R_N}$ is given by*

the semi-infinite linear optimization problem

$$\begin{aligned}
\max_{r,s,t} \quad & \sum_{k=1}^{K} \left(p_{k0} r_k + p_{k2} s_k\right) + h^T t \\
s.t. \quad & r \in \mathbb{R}^k_+,\ s \in \mathbb{R}^{k \times d},\ t \in \mathbb{R}^{d'} \\
& \sum_{k=1}^{K} \max\{a^T s_k - b r_k, 0\} \leq Q_{C_N}(\alpha_1) + \frac{1}{N} \sum_{i=1}^{N} \max\{a^T \xi^i - b, 0\} \quad \forall\ ||a||_1 + |b| \leq 1 \qquad (4.37) \\
& \sum_{k=1}^{K} r_k = 1 \\
& H^T t - \sum_{k=1}^{K} \left(r_k p_{k1} - P_k z_k\right) \leq 0, \qquad (4.38)
\end{aligned}$$

and the optimal solution $\overline{x}$ is given by the dual variable associated with constraint (4.38)*.*

To separate over the constraint (4.37) *at a given s', r' we may solve the linear optimization problems*

$$v_\gamma = \max_{||a||_1 + |b| \leq 1} \left(\sum_{k=1}^{K} \gamma_k \left(a^T s'_k - b r'_k\right) - \frac{1}{N} \sum_{i=1}^{N} \max\{a^T \xi^i - b, 0\} \right)$$

for each $\gamma \in \{0,1\}^k \backslash \{(0,\ldots,0)\}$. If $v_\gamma \leq 0$ for every γ, the constraint is satisfied. Otherwise, for γ such that $v_\gamma > 0$ and a, b being the corresponding optimizers, the following is a separating hyperplane,

$$\sum_{k=1}^{K} \gamma_k \left(a^T s_k - b r_k\right) \leq Q_{C_N}(\alpha_1) + \frac{1}{N} \sum_{i=1}^{N} \max\{a^T \xi^i - b, 0\}.$$

Notice that the solution above does not explicitly involve α_2 – the significance of the test for $\mathbb{E}\left[||\xi||_2^2\right]$. This a consequence of the structure of the cost function (4.36) and the unbounded support $\Xi = \mathbb{R}^d$. The implication is that we may let $\alpha_2 \to 0$, increasing the probability of the finite-sample guarantee without affecting the solution $\overline{x}$ or the bound $\overline{z}$. This is the approach we take in the empirical study in Section 4.7.3.

Example 4.33. *Portfolio allocation.* There are d securities with unknown future returns ξ_i and we must divide our budget into fractions x_i invested in security i with $\sum_i x_i = 1$. The return on a unit budget is $x^T \xi$. There are two common cost functions used in this problem.

One is the negative utility of unit-budget returns for a piecewise-linear concave

nondecreasing utility function. That is, given parameters β_k, γ_k such that $\gamma_k \leq 0$,

$$c(x;\xi) = -u(x^T\xi) = \max_{k=1,\ldots,K} \left(\gamma_k x^T \xi + \beta_k\right), \tag{4.39}$$

which fits into the framework of (4.36) using $p_{k0} = \beta_k$, $p_{k1} = 0$, $p_{k2} = 0$, $P_k = \gamma_k I_d$.

A more popular choice in practice is the conditional value at risk of negative returns. The CVaR at level ϵ of a random loss L with quantile function F_L^{-1} is the expectation above the $(1-\epsilon)$-quantile:

$$\mathrm{CVaR}_\epsilon(L) = \mathbb{E}\left[L \;\middle|\; L \geq F_L^{-1}(1-\epsilon)\right] = \inf_{\beta\in\mathbb{R}} \mathbb{E}\left[\beta + \frac{1}{\epsilon}\max\{L-\beta, 0\}\right],$$

where the latter equivalent definition is due to Rockafellar and Uryasev (2000). We can formulate the min CVaR problem using (4.36) by augmenting the decision vector as (β_+, β_-, x) and setting $H = (0, 0, 1, \ldots, 1)$, $h = 1$, $K = 2$, and

$$p_{1,0} = 0,\ p_{1,2} = 0,\ P_1 = 0,\ p_{2,0} = 0,\ p_{2,2} = 0,$$

$$p_{1,1} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ p_{2,1} = \begin{pmatrix} 1 - 1/\epsilon \\ -1 + 1/\epsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ P_2 = \begin{pmatrix} 0 \cdots 0 \\ 0 \cdots 0 \\ \\ [-I_d/\epsilon] \\ \\ \end{pmatrix}.$$

Notice that since $K = 2$, separating over the constraint (4.37) requires solving only three linear optimization problems.

4.6 Estimating the Price of Data

Our framework allows one to compute the price one would be willing to pay for further data gathering. Given the present dataset, we define the price of data (PoD) as follows:

$$\mathrm{PoD} = \overline{z}\left(\xi^1, \ldots, \xi^N\right) - \mathbb{E}\left[\overline{z}\left(\xi^1, \ldots, \xi^N, \xi^{N+1}\right) \middle| \xi^1, \ldots, \xi^N\right]. \tag{4.40}$$

PoD is equal to the expected marginal benefit of one additional data point in reducing our bound on costs.

One way to estimate the above quantity is via resampling:

$$\mathrm{PoD} \approx \overline{z}\left(\xi^1, \ldots, \xi^N\right) - \frac{1}{N}\sum_{i=1}^{N} \overline{z}\left(\xi^1, \ldots, \xi^N, \xi^i\right). \tag{4.41}$$

The resampled average can also be, in turn, estimated by an average over a smaller random subsample from the data. This approach is illustrated numerically in Section 4.7.3.

In the case of the newsvendor problem using the KS test, the closed form solution yields a simpler approximation. Observe that in Proposition 4.29, small changes to the data change $\overline{x}$ very little and the costs for ξ near $\overline{x}$ (in particular, between i_{lo} and i_{hi}) are small compared to costs far away from $\overline{x}$. Thus, we suggest the approximation

$$\text{PoD} \approx \left(Q_{D_N}(\alpha) - Q_{D_{N+1}}(\alpha)\right)\left(c\left(\overline{x};\underline{\xi}\right) + c\left(\overline{x};\overline{\xi}\right)\right). \tag{4.42}$$

This approximation is illustrated numerically in section 4.7.1.

We can write a more explicit approximation using the asymptotic approximation of $Q_{D_N}(\alpha)$ (see Thas (2009)) and $1/\sqrt{N} - 1/\sqrt{N+1} \approx 1/(2N^{3/2})$ for large N:

$$\text{PoD} \approx \frac{q_\alpha}{2N^{3/2}}\left(c\left(\overline{x};\underline{\xi}\right) + c\left(\overline{x};\overline{\xi}\right)\right) \quad \text{where} \quad q_\alpha = \begin{cases} 1.36, & \alpha = 0.05, \\ 1.22, & \alpha = 0.1, \\ 1.07, & \alpha = 0.2. \end{cases}$$

4.7 Empirical Study

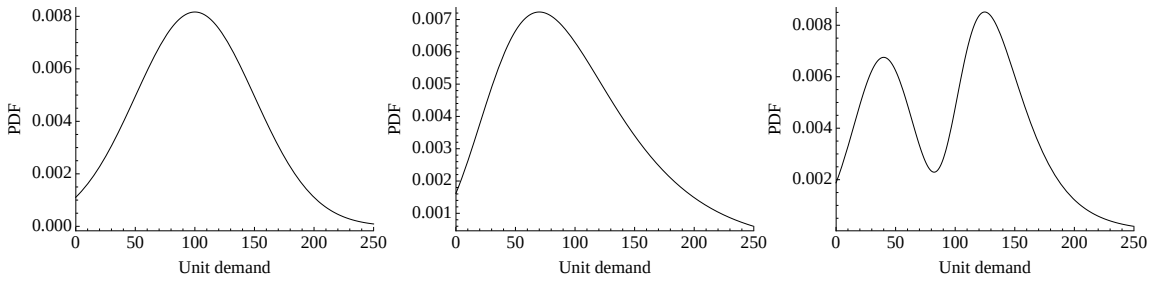
We now turn to an empirical study of Robust SAA as applied to specific problems in inventory and portfolio management. The cost functions are all specific cases of the examples studied in Section 4.5. Recall that in these examples the resulting formulations were all linear and second-order cone optimization problems.

4.7.1 Single-Item Newsvendor

We begin with an application to the classic newsvendor problem with continuous demand distribution, as studied in Example 4.28. We will consider both bounded and unbounded distributions. In the latter case we employ the Student's T-test to ensure a finite solution. We implement (4.3) in closed form for the KS test with bounded support using Proposition 4.29, using IPOPT 3.11 Wächter and Biegler (2006) for the AD test, and using GUROBI 5.5 Gurobi Optimization Inc. (2013) otherwise.

We consider a 95% service-level requirement ($b = 19$, $h = 1$) and each of the following distributions, truncated above at 250 units in the bounded case:

Figure 4-3: The PDFs of Demand Distributions for the Newsvendor Problem



(a) Normal (b) Gumbel (c) Mixed

1. Normal distribution with mean 100 and standard deviation 50, truncated to be nonnegative.

2. Right-skewed Gumbel distribution with location 70 and scale $30/\gamma$ (γ the Euler constant), truncated to be nonnegative.

3. Mixture model of 40% normal with mean 40 and standard deviation 25 and 60% right-skewed Gumbel with location 125 and scale $15/\gamma$, truncated to be nonnegative.

We plot their PDFs in Figure 4-3. In the bounded case we use a significance level of 20% (i.e., 80% confidence). In the unbounded case we use a significance level of 15% for the GoF test and 5% for the Student's T-test (yielding total significance of 20%).

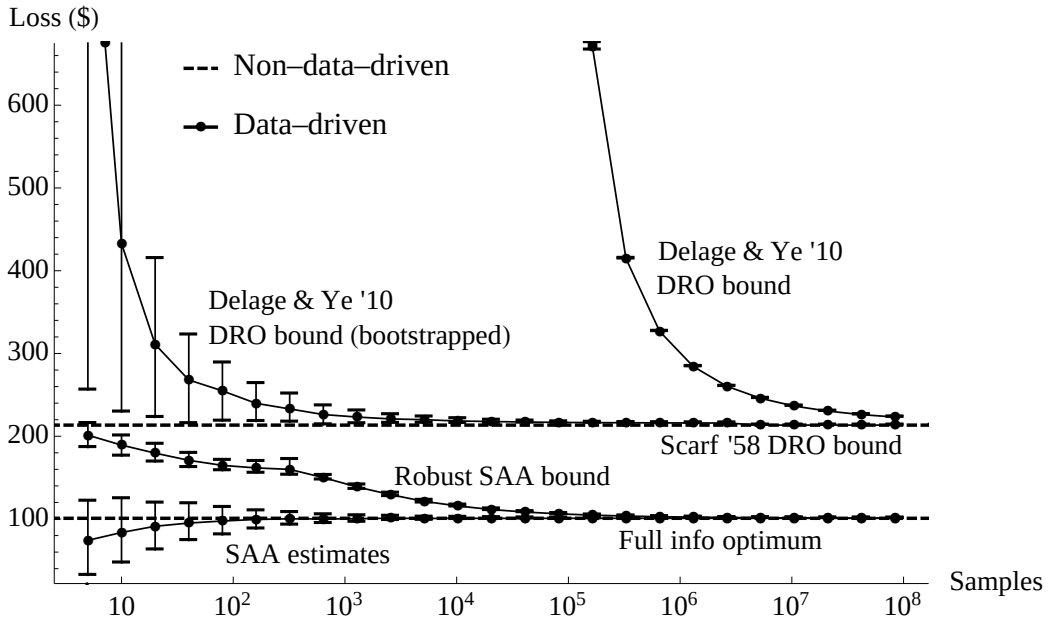
In Figure 4-4 we consider the bounded normal distribution and compare the values of the full-information problem (4.1), SAA estimates (4.2), Robust SAA bounds $\overline{z}$ using the KS test, the data-driven DRO bound of Delage and Ye (2010), and the non-data-driven DRO bound of Scarf (1958). We note that the SAA estimates converge to the true optimum, but very often underestimate the true costs of SAA's recommended control (e.g. 65% of time for $N = 100$), which is necessarily above the full-information optimum. These estimates are biased (the mean is below the full-information optimum) and have the peculiar property that the estimated costs grow with N, contradicting the value of data collection.

The data-driven guarantees of Delage and Ye (2010) do not converge to the full information optimum. Instead, they converge upon the bound of Scarf (1958), in which one restricts mean and variance to their exact true values and letting all else vary. These data-driven bounds, however, decrease with N, consistent with true costs improving as more data is gathered. Interpreting the DUS of Delage and Ye (2010) as a hypothesis test, we also attempt to apply the bootstrap to estimate valid thresholds $\gamma_{1,N}(\alpha)$, $\gamma_{2,N}(\alpha)\,\gamma_{3,N}(\alpha)$ (see (4.21)). The result is plotted in the same figure. We note that the bound is much smaller, but still non-convergent.

Our proposed method provides an order quantity, cost guarantee, and true costs that converges to the full information optimum control and value. In this particular case, computation of the bound and order quantity is done in closed form. The value of the bounds decrease with N, in agreement with the value of data collection, and their convergence is consistent with the notion that we discover F as we get more data. The magnitude of the effect of data collection on the bounds is what we termed the Price of Data, or PoD, in Section 4.6. In Figure 4-5, we compare the true PoD (4.40) and the approximation we developed (4.42). Notice the fit is quite tight.

In Figure 4-6, we consider the behavior of Robust SAA for a wider range of distributions and tests. First and foremost, the numerical results confirm the guarantees and convergence as $N \to \infty$ irrespective of what is the true, unknown distribution F. Different tests also seem to yield mostly comparable results, with the AD test providing slightly better results when N is at least 100. With small N, the Kuiper and Watson tests seem to perform the best. These observations should not, however, be taken as general conclusions about the relative performance of these tests for general problems. The conservatism of the guarantees depends both on the structure

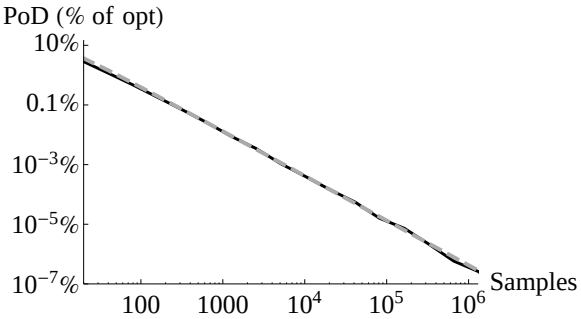
Figure 4-4: Convergence of Robust SAA guarantees and SAA estimates compared with the data-driven DRO of Delage and Ye (2010) and non-data-driven DRO of Scarf (1958)



Note: The error bars denote the 20^{th} and 80^{th} percentiles. All data-driven DRO guarantee bounds are solved at a significance of 20%. The horizontal axis is on a log scale.

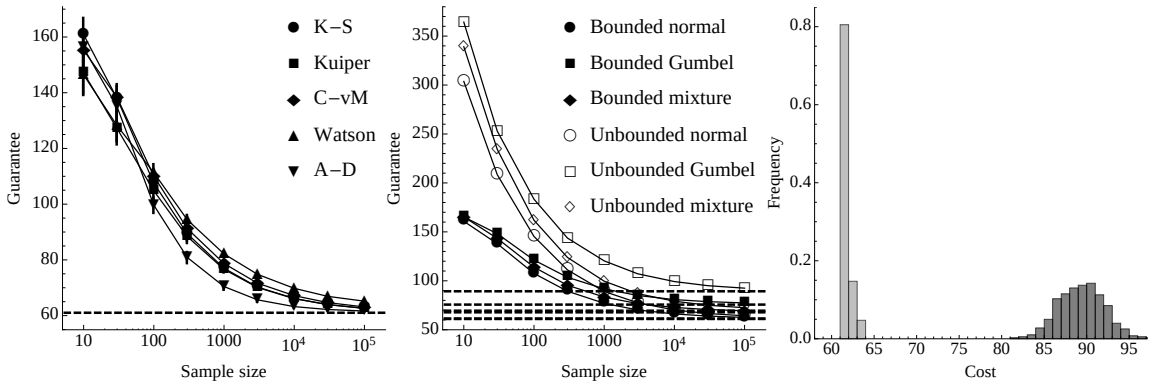
of the cost function as well as the true, unknown distribution and how we test it. For practical purposes, if the convergence rates are comparable as they are here, we recommend to choose the test that yields the simplest optimization problem, which in this case is the KS test.

Figure 4-5: The Price of Data in the Newsvendor Problem



Note: The average of true PoD is shown in solid black and the distribution-agnostic approximation (4.42) is shown in dashed grey. Both axes are on a log scale.

Figure 4-6: Probabilistic Guarantees of Robust SAA for the Singled-Item Newsvendor Problem



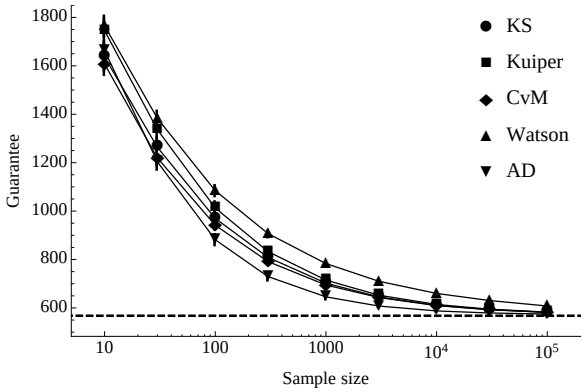
(a) Various tests (b) Various distributions (c) Compared to true cost

Note: Significance is set to 20%. Panel (a) displays the guarantees given by various tests for the bounded normal distribution as the sample size grows. The vertical lines in (a) denote the span from the 20^{th} to the 80^{th} percentile with respect to sampling. Panel (b) displays the guarantees given by the Kolmogorov-Smirnov test for various distributions. Dashed lines in (a) and (b) denote the full-information optimum. Panel (c) displays the distribution of true costs (light grey) and guarantees (dark grey) for the Kolmogorov-Smirnov test with $N = 300$ samples from the bounded normal distribution.

4.7.2 Multi-Item Newsvendor

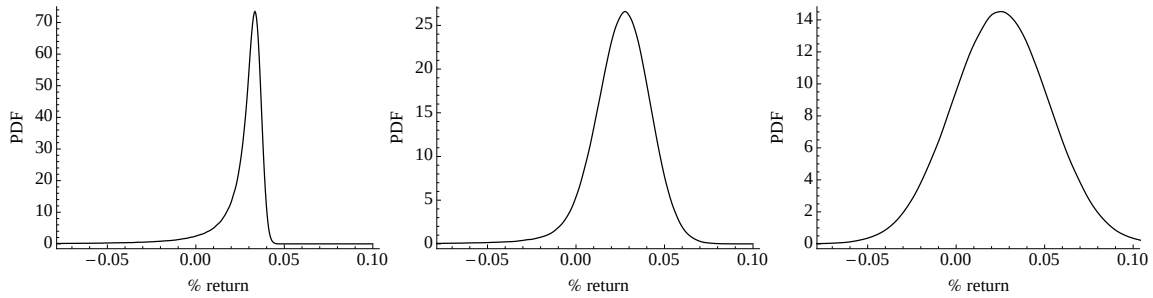
We now consider the multi-item newsvendor problem, which is a special case of a separable cost function as considered in Section 4.5.3. Recall that in the multi-item

Figure 4-7: The Probabilistic Guarantees of Robust SAA for the Multi-Item Newsvendor Problem



Note: Significance is set to 20%. The vertical lines denote the span from the 20^{th} to the 80^{th} percentile. The dashed line denotes the full-information optimum.

Figure 4-8: The PDFs of Security Returns Distributions for the Portfolio Allocation Problem



(a) $i = 1$ (b) $i = 5$ (c) $i = 10$

newsvendor we have $X = \{x \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i \leq \overline{x}\}$ for some capacity $\sum_i x_i \leq \overline{x}$ and

$$c(x;\xi) = \sum_{i=1}^{d} c_i(x_i;\xi_i),$$

where each c_i takes the form of a newsvendor cost function with its own parameters b_i, h_i.

We consider the case of three items, each having demand distributed as one of the three bounded distributions considered in the single-item case, with the parameters $\overline{x} = 250$, $r_1 = 15$, $r_2 = 10$, $r_3 = 5$, $c_1 = 6$, $c_2 = 4$, $c_3 = 2$, $b_1 = 3$, $b_2 = 2$, $b_3 = 1$. In our application of Robust SAA we employ the test based on marginals where, for different choices of univariate test, we use the same GoF test for each marginal, each at significance of 6.67% (total signifiance 20%).

We present the results in Figure 4-7. Again, we plot both the mean and 20^{th} and 80^{th} percentiles of probabilistic guarantees as the size of the sample grows and compare these to the full-information optimum. As predicted by the theory, we observe convergence of guarantees even though testing marginals is not generally a uniformly consistent test.

4.7.3 Portfolio Allocation

We now consider the minimum-CVaR portfolio allocation problem as studied in Example 4.33. We minimize the 10%-level CVaR of negative returns of a portfolio of $d = 10$ securities. The random returns are supported on the unbounded domain $\mathbb{R}^{10}$ and given by the factor model

$$\xi_i = \frac{i}{11}\tau + \frac{11-i}{11}\zeta_i \quad i = 1, \ldots, 10$$

where τ is a common market factor following a normal distribution with mean 2.5% and standard deviation 3% and ζ_i's are independent idiosyncratic contributions all following a negative Pareto distribution with upper support 3.7%, mean 2.5%, and

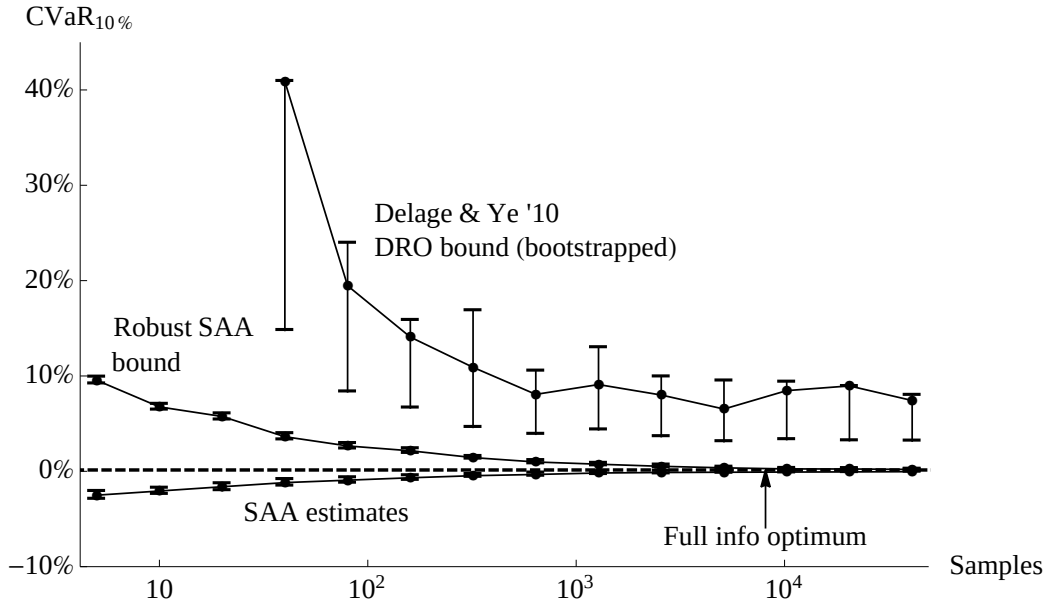
standard deviation 3.8% (i.e. $\zeta_i \sim 0.05 - \text{Pareto}(0.013, 2.05)$). All securities have the same average return. Lower indexed securities are more volatile but are also more diversified. We plot the PDFs of the returns of a few of the securities in Figure 4-8.

For samples drawn from this distribution, we consider data-driven solutions by the SAA, the DRO of Delage and Ye (2010), and our method using the test for LCX. We use the bootstrap to compute $Q_{C_N}(\alpha)$ (see Section C.1). Since the constants $\gamma_{1,N}(\alpha)$, $\gamma_{2,N}(\alpha)\,\gamma_{3,N}(\alpha)$ (see (4.21)) for the DRO of Delage and Ye (2010) are only developed therein for the case of known bounded support and in order to offer a fair comparison, we also bootstrap these thresholds. We implement the DRO (4.3) using GUROBI 5.5 Gurobi Optimization Inc. (2013).

We present the results in Figure 4-9. As can be seen the SAA underestimates the risk of its recommended portfolios. The method of Delage and Ye (2010) provides valid bounds but they do not appear to converge and are also highly variable. In comparison, our method using the LCX test provides apparently convergent guarantees and its guarantees are tightly concentrated.

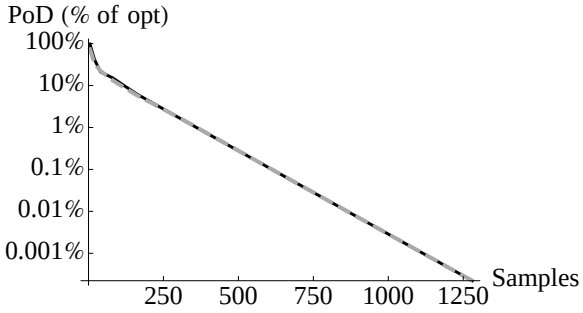
In Figure 4-10 we compare the true price of data (4.40) for the LCX-based DRO bound and the resampling based approximation of it (4.41).

Figure 4-9: Robust SAA Guarantees for the Portfolio Allocation Problem Compared to Other Data-Driven Approachs



Note: The vertical bars denote the span from the 20th to the 80th percentile. The full-information optimum is shown as a dashed line.

Figure 4-10: The Price of Data in Portfolio Allocation



Note: The average of true PoD is shown in solid black and the resampling-based approximation (4.41) is shown in dashed grey. The vertical axis is on a log scale.

4.8 Conclusions

In this chapter, we proposed a novel, tractable approach to data-driven optimization called robust sample average approximation (Robust SAA). Robust SAA enjoys the tractability and finite-sample performance guarantees of many existing data-driven methods, but, unlike those methods, additionally exhibits asymptotic behavior similar to traditional sample average approximation (SAA). The key to the approach is a novel connection between SAA, DRO, and statistical hypothesis testing.

In particular, we were able to link properties of a data-driven *optimization* problem, i.e., its finite sample and asymptotic performance, to *statistical* properties of an associated goodness-of-fit hypothesis test, i.e., its significance and consistency. As a theoretical consequence, this connection allow us to describe the finite sample and asymptotic performance of both Robust SAA and other data-driven DRO formulations. As a practical consequence, our hypothesis testing perspective first, sheds light on which data-driven DRO formulations are likely to perform well in particular applications and second, enables us to use powerful, numerical tools like bootstrapping to improve their performance. Numerical experiments in inventory management and portfolio allocation confirm that our new method Robust SAA is tractable and can outperform existing data-driven methods in these applications.

Chapter 5

Data-Driven Robust Optimization

The last decade witnessed an explosion in the availability of data for operations research applications. Motivated by this growing availability, we propose a novel schema for utilizing data to design uncertainty sets for robust optimization using statistical hypothesis tests. The approach is flexible and widely applicable, and robust optimization problems built from our new sets are computationally tractable, both theoretically and practically. Furthermore, optimal solutions to these problems enjoy a strong, finite-sample probabilistic guarantee. We describe concrete procedures for choosing an appropriate set for a given application and applying our approach to multiple uncertain constraints. Computational evidence in portfolio management and queuing confirm that our data-driven sets significantly outperform traditional robust optimization techniques whenever data is available.

5.1 Introduction

Robust optimization is a popular approach to optimization under uncertainty. The key idea 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 sets, robust models yield tractable optimization problems whose solutions perform as well or better than other approaches. With poorly chosen sets, however, robust models may be overly-conservative or computationally intractable. Choosing a good set is crucial. Fortunately, there are several theoretically motivated and experimentally validated proposals for constructing good uncertainty sets (Ben-Tal and Nemirovski 2000, Bertsimas and Sim 2004, Ben-Tal et al. 2009, Bandi and Bertsimas 2012). These proposals share a common paradigm; they combine a priori reasoning with mild assumptions on the uncertainty to motivate the construction of the set.

On the other hand, the last decade witnessed an explosion in the availability of data. Massive amounts of data are now routinely collected in many industries. 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 reasoning and assumptions and towards a new data-centered paradigm. A natural question, then, is how should robust optimization techniques be tailored to this new paradigm?

In this chapter, we propose a general schema for designing uncertainty sets for robust optimization from data. We consider uncertain constraints of the form $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ where $\mathbf{x} \in \mathbb{R}^k$ is the optimization variable, and $\tilde{\mathbf{u}} \in \mathbb{R}^d$ is an uncertain parameter. We model this constraint by choosing a set $\mathcal{U}$ and forming the corresponding robust constraint

$$f(\mathbf{u}, \mathbf{x}) \leq 0 \quad \forall \mathbf{u} \in \mathcal{U}. \tag{5.1}$$

We assume throughout that $f(\mathbf{u}, \mathbf{x})$ is concave in $\mathbf{u}$ for any $\mathbf{x}$.

In many applications, robust formulations decompose into a series constraints of the form (5.1) through an appropriate transformation of variables, including uncertain linear optimization and multistage adaptive optimization (see, e.g., Ben-Tal et al. (2009)). In this sense, (5.1) is a fundamental building block for more complex robust optimization models.

Many approaches (Bertsimas and Sim 2004, Ben-Tal et al. 2009, Chen et al. 2010) to constructing uncertainty sets for (5.1) assume $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known except for some assumed structural features. For example, they may assume that $\mathbb{P}^*$ has independent components, while its marginal distributions are not known. Given $\epsilon > 0$, these approaches seek sets $\mathcal{U}_\epsilon$ that satisfy two key properties:

(P1) The robust constraint (5.1) is *computationally tractable.*

(P2) The set $\mathcal{U}_\epsilon$ *implies a probabilistic guarantee for* $\mathbb{P}^*$ *at level* ϵ, that is, for any $\mathbf{x}^* \in \mathbb{R}^k$ and for every function $f(\mathbf{u}, \mathbf{x})$ concave in $\mathbf{u}$ for all $\mathbf{x}$, we have the implication:

$$\text{If } f(\mathbf{u}, \mathbf{x}^*) \leq 0 \quad \forall \mathbf{u} \in \mathcal{U}_\epsilon, \text{ then } \mathbb{P}^*(f(\tilde{\mathbf{u}}, \mathbf{x}^*) \leq 0) \geq 1 - \epsilon. \tag{5.2}$$

(P2) ensures that a feasible solution to the robust constraint will also be feasible with probability $1 - \epsilon$ with respect to $\mathbb{P}^*$, despite not knowing $\mathbb{P}^*$ exactly. Existing proposals achieve (P2) by leveraging the a priori structural features of $\mathbb{P}^*$. Some of these approaches, e.g., (Bertsimas and Sim 2004), only consider the special case when $f(\mathbf{u}, \mathbf{x})$ is bi-affine, but one can generalize them to (5.2) using techniques from Ben-Tal et al. (2012) (see also Sec. 5.2.1).

Like previous proposals, we also assume $\tilde{\mathbf{u}}$ is a random variable whose distribution $\mathbb{P}^*$ is not known exactly, and seek sets $\mathcal{U}_\epsilon$ that satisfy these properties. Unlike previous proposals – and this is critical – we assume that we have data $\mathcal{S} = \{\hat{\mathbf{u}}^1, \ldots, \hat{\mathbf{u}}^N\}$ drawn i.i.d. according to $\mathbb{P}^*$. By combining these data with the a priori structural features of $\mathbb{P}^*$, 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 using the confidence region of a statistical hypothesis test to quantify what we learn about $\mathbb{P}^*$ from the data. Specifically, our constructions depend on three ingredients: the a priori assumptions on $\mathbb{P}^*$, the data, and a hypothesis test. By pairing different a priori assumptions and tests, we obtain distinct data-driven uncertainty sets, each with its own geometric shape, computational properties, and modeling power. These sets can capture a variety of features of $\mathbb{P}^*$, including skewness, heavy-tails and correlations.

In principle, there is a multitude of possible pairings of a priori assumptions and tests. We focus on pairings we believe are most relevant to applied robust modeling. Specifically, we consider a priori assumptions that are common in practice and tests that lead to tractable uncertainty sets. Our list is non-exhaustive; there may exist other pairings that yield effective sets. Specifically, we consider situations where:

- $\mathbb{P}^*$ has known, finite discrete support (Sec. 5.4).
- $\mathbb{P}^*$ may have continuous support, and the components of $\tilde{\mathbf{u}}$ are independent (Sec. 5.5).
- $\mathbb{P}^*$ may have continuous support, but data are drawn from its marginal distributions asynchronously (Sec. 5.6). This situation models the case of missing values.
- $\mathbb{P}^*$ may have continuous support, and data are drawn from its joint distribution (Sec. 5.7). This is the general case.

Table 5.1 summarizes the a priori structural assumptions, hypothesis tests, and resulting uncertainty sets that we propose. Each set is convex and admits a tractable, explicit description; see the referenced equations.

For each of our sets, we provide an explicit, equivalent reformulation of (5.1). The complexity of optimizing over this reformulation depends both on the function $f(\mathbf{u}, \mathbf{x})$ and the set $\mathcal{U}$. For each of our sets, we show that this reformulation is polynomial time tractable for a large class of functions f including bi-affine functions, separable functions, conic-quadratic representable functions and certain sums of uncertain exponential functions. By exploiting special structure in some of our sets, we can provide specialized routines for directly separating over (5.1) for bi-affine f. In these cases, the column "Separation" in Table 5.1 roughly describes these routines. Utilizing this separation routine within a cutting-plane method may offer performance superior to reformulation based-approaches (Bertsimas et al. (2014a), Mutapcic and Boyd (2009)).

We are not the first to consider using hypothesis tests in data-driven optimization. Recently, Ben-Tal et al. (2013) proposed a class of data-driven uncertainty sets based on phi-divergences. (Phi divergences are closely related to some types of hypothesis tests.) They focus on the case where the uncertain parameter is a probability distribution with known, finite, discrete support. By contrast, we design uncertainty sets for general uncertain parameters with potentially continuous support such as future product demand, service times, and asset returns. Many existing robust optimization applications utilize similar general uncertain parameters. Consequently, retrofitting

Table 5.1: Summary of Data-Driven Uncertainty Sets Proposed

Assumptions on $\mathbb{P}^*$	Hypothesis Test	Geometric Description	Eqs.	Separation
Discrete support	χ^2-test	SOC	(5.10) (5.13)	
Discrete support	G-test	Polyhedral*	(5.10) (5.14)	
Independent marginals	KS Test	Polyhedral*	(5.20)	line search
Independent marginals	K Test	Polyhedral*	(D.23)	line search
Independent marginals	CvM Test	SOC*	(D.23)	
Independent marginals	W Test	SOC*	(D.23)	
Independent marginals	AD Test	EC	(D.23)	
Independent marginals	Chen et al. (2007)	SOC	(5.25)	closed-form
None	Marginal Samples	Box	(5.30)	closed-form
None	Linear Convex Ordering	Varies	(5.33)	linear optimization
None	Shawe-Taylor & Cristianini (2003)	SOC	(5.37)	closed-form
None	Delage & Ye (2010)	LMI	(5.38)	

Note: SOC, EC and LMI denote second-order cone representable sets, exponential cone representable sets, and linear matrix inequalities, respectively. The additional "*" notation indicates a set of of the above type with one additional, relative entropy constraint. KS, K, CvM, W, and AD denote the Kolmogorov-Smirnov, Kuiper, Cramer-von Mises, Watson and Anderson-Darling goodness of fit tests, respectively. In some cases, we can separate over the constraint (5.1) for bi-affine f with a specialized algorithm. In these cases, the column "Separation" roughly describes this algorithm.

these applications with our new data-driven sets to yield data-driven variants is perhaps more straightforward than using sets for uncertain probabilities. From a methodological perspective, treating general uncertain parameters requires combining ideas from a variety of hypothesis tests (not just those based on phi-divergences of discrete distributions) with techniques from convex analysis and risk theory. (See Sec. 5.3.)

Other authors have also considered more specialized applications of hypothesis testing in data-driven optimization. Klabjan et al. (2013) proposes a distributionally robust dynamic program based on Pearson's χ^2-test for a particular inventory problem. Goldfarb and Iyengar (2003) calibrate an uncertainty set for the mean and covariance of a distribution using linear regression and the t-test. It is not clear how to generalize these methods to other settings, e.g., distributions with continuous support in the first case or general parameter uncertainty in the second. By contrast, we offer a comprehensive study of the connection between hypothesis testing and uncertainty set design, addressing a number of cases with general machinery.

Moreover, our hypothesis testing perspective provides a unified view of many other data-driven methods from the literature. For example, Calafiore and El Ghaoui (2006) and Delage and Ye (2010) have proposed data-driven methods for chance-constrained and distributionally robust problems, respectively without using hypothesis testing. We show how these works can be reinterpreted through the lens of hypothesis testing. Leveraging this viewpoint enables us to apply state-of-the-art methods from statistics, such as the bootstrap, to refine these methods and improve their numerical performance. Moreover, applying our schema, we can design data-driven uncertainty sets for robust optimization based upon these methods. Although we focus on Calafiore and El Ghaoui (2006) and Delage and Ye (2010) in this chapter, this strategy applies equally well to a host of other methods, such as the likelihood estimation approach of Wang et al. (2009). In this sense, we believe hypothesis testing and uncertainty set design provide a common framework in which to compare and contrast different approaches.

Finally, we note that Campi and Garatti (2008) propose a very different data-driven method for robust optimization not based on hypothesis tests. In their approach, one replaces the uncertain constraint $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ with N sampled constraints over the data, $f(\hat{\mathbf{u}}^j, \mathbf{x}) \leq 0$, for $j = 1, \ldots, N$. For $f(\mathbf{u}, \mathbf{x})$ convex in $\mathbf{x}$ with arbitrary dependence in $\mathbf{u}$, they provide a tight bound $N(\epsilon)$ such that if $N \geq N(\epsilon)$, then, with high probability with respect to the sampling, any $\mathbf{x}$ which is feasible in the N sampled constraints satisfies $\mathbb{P}^*(f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0) \geq 1 - \epsilon$. Various refinements of this base method have also been proposed yielding smaller bounds $N(\epsilon)$, including incorporating ℓ_1-regularization (Campi and Carè 2013) and allowing $\mathbf{x}$ to violate a small fraction of the constraints (Calafiore and Monastero 2012). Compared to our approach, these methods are more generally applicable and provide a similar probabilistic guarantee. In the special case we treat where $f(\tilde{\mathbf{u}}, \mathbf{x})$ is concave in $\mathbf{u}$, however, our proposed approach offers some advantages. First, because it leverages the concave structure of $f(\mathbf{u}, \mathbf{x})$, our approach generally yields less conservative solutions (for the same N and ϵ) than Campi and Garatti (2008). (See Sec. 5.3.) Second, for fixed $\epsilon > 0$, our approach is applicable even if $N < N(\epsilon)$, while theirs is not. This distinction is important when ϵ is very small and there may not exist enough data. Finally, as

we will show, our approach reformulates (5.1) as a series of (relatively) sparse convex constraints, while the Campi and Garatti (2008) approach will in general yield N dense constraints which may be numerically challenging when N is large. For these reasons, practitioners may prefer our proposed approach in certain applications.

We summarize our contributions:

1. We propose a new, systematic schema for constructing uncertainty sets from data using statistical hypothesis tests. When the data are 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 5.1.

3. We prove that robust optimization problems over each of our sets are generally tractable. Specifically, for each set, we derive an explicit robust counterpart to (5.1) and show that for a large class of functions $f(\mathbf{u}, \mathbf{x})$ optimizing over this counterpart can be accomplished in polynomial time using off-the-shelf software.

4. We unify several existing data-driven methods through the lens of hypothesis testing. Through this lens, we motivate the use of common numerical techniques from statistics such as bootstrapping and gaussian approximation to improve their performance. Moreover, we apply our schema to derive new uncertainty sets for (5.1) inspired by the refined versions of these methods.

5. We propose a new approach to modeling multiple uncertain constraints simultaneously with our sets by optimizing the parameters chosen for each individual constraint. We prove that this technique is tractable and yields solutions which will satisfy all the uncertain constraints simultaneously for any desired level ϵ.

6. We provide guidelines for practitioners on choosing an appropriate set and calibrating its parameters by leveraging techniques from model selection in machine learning.

7. Through applications in queueing and portfolio allocation, we assess the relative strengths and weaknesses of our sets. Overall, we find that although all of our sets shrink in size as $N \to \infty$, they differ in their ability to represent features of $\mathbb{P}^*$. Consequently, they may perform very differently in a given application. In the above two settings, we find that our model selection technique frequently identifies a good set choice, and a robust optimization model built with this set performs as well or better than other robust data-driven approaches.

The remainder of the chapter is structured as follows. Sec. 5.2 reviews background to keep the chapter self-contained. Sec. 5.3 presents our schema for constructing uncertainty sets. Sec. 5.4-5.7 describe the various constructions in Table 5.1. Sec. 5.8 reinterprets several techniques in the literature through the lens of hypothesis testing

and, subsequently, uses them to motivate new uncertainty sets. Sec. 5.9 and Sec. 5.10 discuss modeling multiple constraints and choosing the right set for an application, respectively. Sec. 5.11 presents numerical experiments, and Sec. 5.12 concludes. All proofs are in the electronic companion.

5.1.1 Notation and Setup

Boldfaced lowercase letters ($\mathbf{x}$, $\boldsymbol{\theta}$, ...) denote vectors, boldfaced capital letters ($\mathbf{A}$, $\mathbf{C}$, ...) denote matrices, and ordinary lowercase letters (x, θ) denote scalars. Calligraphic type ($\mathcal{P}, \mathcal{S} \ldots$) denotes sets. The i^{th} coordinate vector is $\mathbf{e}_i$, and the vector of all ones is $\mathbf{e}$. We 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 a generic probability measure for $\tilde{\mathbf{u}}$, and $\mathbb{P}^*$ denotes its true (unknown) measure. Moreover, $\mathbb{P}_i$ denotes the marginal measure of $\tilde{u}_i$. 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 $\mathbb{P}^*$. Finally, $\hat{\mathbb{P}}$ denotes the empirical distribution with respect to $\mathcal{S}$.

5.2 Background

To keep the chapter self-contained, we recall some results needed to prove our sets are tractable and imply a probabilistic guarantee.

5.2.1 Tractability of Robust Nonlinear Constraints

Ben-Tal et al. (2012) study constraint (5.1) and prove that for nonempty, convex, compact $\mathcal{U}$ satisfying a mild, regularity condition[1], (5.1) is equivalent to

$$\exists \mathbf{v} \in \mathbb{R}^d \text{ s.t. } \delta^*(\mathbf{v} |\, \mathcal{U}) - f_*(\mathbf{v}, \mathbf{x}) \leq 0. \tag{5.3}$$

Here, $f_*(\mathbf{v}, \mathbf{x})$ denotes the partial concave-conjugate of $f(\mathbf{u}, \mathbf{x})$ and $\delta^*(\mathbf{v} |\, \mathcal{U})$ denotes the support function of $\mathcal{U}$, defined respectively as

$$f_*(\mathbf{v}, \mathbf{x}) \equiv \sup_{\mathbf{u} \in \mathbb{R}^d} \mathbf{u}^T \mathbf{v} - f(\mathbf{u}, \mathbf{x}), \qquad \delta^*(\mathbf{v} |\, \mathcal{U}) \equiv \sup_{\mathbf{u} \in \mathcal{U}} \mathbf{v}^T \mathbf{u}.$$

For many $f(\mathbf{u}, \mathbf{x})$, $f_*(\mathbf{v}, \mathbf{x})$ admits a simple, explicit description. For example, for bi-affine $f(\mathbf{u}, \mathbf{x}) = \mathbf{u}^T \mathbf{F} \mathbf{x} + \mathbf{f}_{\mathbf{u}}^T \mathbf{u} + \mathbf{f}_{\mathbf{x}}^T \mathbf{x} + f_0$, we have

$$f_*(\mathbf{v}, \mathbf{x}) = \begin{cases} -\mathbf{f}_{\mathbf{x}}^T \mathbf{x} - f_0 & \text{if } v = \mathbf{F}\mathbf{x} + \mathbf{f}_{\mathbf{u}} \\ -\infty & \text{otherwise,} \end{cases}$$

[1] An example of a sufficient regularity condition is that $ri(\mathcal{U}) \cap ri(dom(f(\cdot, \mathbf{x}))) \neq \emptyset$, $\forall \mathbf{x} \in \mathbb{R}^k$. Here $ri(\mathcal{U})$ denotes the *relative interior* of $\mathcal{U}$. Recall that for any non-empty convex set $\mathcal{U}$, $ri(\mathcal{U}) \equiv \{\mathbf{u} \in \mathcal{U} \ : \ \forall \mathbf{z} \in \mathcal{U}, \ \exists \lambda > 1 \text{ s.t. } \lambda \mathbf{u} + (1 - \lambda)\mathbf{z} \in \mathcal{U}\}$ (cf. Bertsekas et al. (2003)).

and (5.3) yields

$$\delta^*(\mathbf{F}\mathbf{x} + \mathbf{f_u} |\, \mathcal{U}) + \mathbf{f_x}^T\mathbf{x} + f_0 \leq 0. \tag{5.4}$$

In what follows, we concentrate on proving we can separate over $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}) \leq t\}$ in polynomial time for each of our sets $\mathcal{U}$, usually by representing this set as a small number of convex inequalities suitable for off-the-shelf solvers. From (5.4), this representation will imply that (5.1) is tractable for each of our sets whenever $f(\mathbf{u}, \mathbf{x})$ is bi-affine.

On the other hand, Ben-Tal et al. (2012) provide a number of other examples of $f(\mathbf{u}, \mathbf{x})$ for which $f_*(\mathbf{v}, \mathbf{x})$ is tractable, including:

Separable Concave: $f(\mathbf{u}, \mathbf{x}) = \sum_{i=1}^k f_i(\mathbf{u})x_i$, for $f_i(\mathbf{u})$ concave and $x_i \geq 0$.

Uncertain Exponentials: $f(\mathbf{u}, \mathbf{x}) = -\sum_{i=1}^k x_i^{u_i}$, for $x_i > 1$ and $0 < u_i \leq 1$.

Conic Quadratic Representable: $f(\mathbf{u}, \mathbf{x})$ such that the set $\{(t, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}) \geq t\}$ conic quadratic representable (cf. Ben-Tal and Nemirovski 2001).

Consequently, by providing a representation of $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}) \leq t\}$ for each of our sets, we will also have proven that (5.1) is tractable for each of these functions via (5.3). In other words, proving $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}) \leq t\}$ is tractable implies that (5.1) is tractable not only for bi-affine functions, but for many other concave functions as well.

For some sets, our formulation of $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}) \leq t\}$ will involve complex nonlinear constraints, such as exponential cone constraints (cf. Table 5.1). Although it is possible to optimize over these constraints directly in (5.3), this approach may be numerically challenging. As mentioned, an alternative is to use cutting-plane or bundle methods as in Bertsimas et al. (2014a), Mutapcic and Boyd (2009). To this end, when appropriate, we provide specialized algorithms for separating over $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}) \leq t\}$.

5.2.2 Hypothesis Testing

We briefly review hypothesis testing as it relates to our set constructions. See Lehmann and Romano (2010) for a more complete treatment.

Given a null-hypothesis H_0 that makes a claim about an unknown distribution $\mathbb{P}^*$, a hypothesis test seeks to use data $\mathcal{S}$ drawn from $\mathbb{P}^*$ to either declare that H_0 is false, or, else, that there is insufficient evidence to determine its validity. For a given significance level $0 < \alpha < 1$, a typical test prescribes a statistic $T \equiv T(\mathcal{S}, H_0)$, depending on the data and H_0, and a threshold $\Gamma \equiv \Gamma(\alpha, \mathcal{S}, H_0)$, depending on α, $\mathcal{S}$, and H_0. If $T > \Gamma$, we reject H_0. Since T depends on $\mathcal{S}$, it is random. The threshold Γ is chosen so that the probability with respect to the sampling of *incorrectly* rejecting H_0 is at most α. The appropriate α is often application specific, although values of $\alpha = 1\%, 5\%$ and 10% are common (cf., Lehmann and Romano 2010, Chapt. 3.1).

As an example, consider the two-sided Student's t-test (Lehmann and Romano 2010, Chapt. 5). Given $\mu_0 \in \mathbb{R}$, the t-test considers the null-hypothesis $H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{u}] =$

μ_0 using the statistic $T = |(\hat{\mu} - \mu_0)/(\hat{\sigma}\sqrt{N})|$ and threshold $\Gamma = t_{N-1,1-\alpha/2}$. Here $\hat{\mu}, \hat{\sigma}$ are the sample mean and sample standard deviation, respectively, and $t_{N-1,1-\alpha}$ is the $1-\alpha$ quantile of the Student t-distribution with $N-1$ degrees of freedom. Under the a priori assumption that $\mathbb{P}^*$ is Gaussian, the test guarantees that we will incorrectly reject H_0 with probability at most α.

Many of the tests we consider are common in applied statistics, and tables for their thresholds are widely available. Several of our tests, however, are novel (e.g., the deviations test in Sec. 5.5.2.) In these cases, we propose using the *bootstrap* to approximate a threshold (cf. Algorithm 1). N_B should be chosen to be fairly large; we take $N_B = 10^4$ in our experiments. The bootstrap is a well-studied and widely-used technique in statistics (Efron and Tibshirani 1993, Lehmann and Romano 2010). Strictly speaking, hypothesis tests based on the bootstrap are only asymptotically valid for large N. (See the references for a precise statement.) Nonetheless, they are routinely used in applied statistics, even with N as small as 100, and a wealth of practical experience suggests they are extremely accurate. Consequently, we believe practitioners can safely use bootstrapped thresholds in the above tests.

Algorithm 1 Bootstrapping a Threshold

Input: $\mathcal{S}$, T, H_0, $0 < \alpha < 1$, $N_B \in \mathbb{Z}_+$
Output: Approximate Threshold Γ
  for $j = 1 \dots N_B$ **do**
    $\mathcal{S}^j \leftarrow$ Resample $|\mathcal{S}|$ data points from $\mathcal{S}$ with replacement
    $T^j \leftarrow T(\mathcal{S}^j, H_0)$
  end for
  return $\lceil N_B(1-\alpha) \rceil$-largest value of $T^1, \dots, T^{N_B}$.

Finally, we introduce the confidence region of a test, which will play a critical role in our construction. Given data $\mathcal{S}$, the $1-\alpha$ confidence region of a test is the set of null-hypotheses that would not be rejected for $\mathcal{S}$ at level $1-\alpha$. For example, the $1-\alpha$ confidence region of the t-test is $\left\{\mu \in \mathbb{R} : \left|\frac{\hat{\mu}-\mu}{\hat{\sigma}\sqrt{N}}\right| \leq t_{N-1,1-\alpha/2}\right\}$. In what follows, however, we commit a slight abuse of nomenclature and instead use the term confidence region to refer to the set of all measures that are consistent with any a priori assumptions of the test and also satisfy a null-hypothesis that would not be rejected. In the case of the t-test, the confidence region in the context of this chapter is

$$\mathcal{P}^t \equiv \left\{\mathbb{P} \in \Theta(-\infty, \infty) : \mathbb{P} \text{ is Gaussian with mean } \mu, \text{ and } \left|\frac{\hat{\mu}-\mu}{\hat{\sigma}\sqrt{N}}\right| \leq t_{N-1,1-\alpha/2}\right\}, \tag{5.5}$$

where $\Theta(-\infty, \infty)$ is the set of Borel probability measures on $\mathbb{R}$.

By construction, the probability (with respect to the sampling procedure) that $\mathbb{P}^*$ is a member of its confidence region is at least $1-\alpha$ as long as all a priori assumptions are valid. 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 that contains $\mathbb{P}^*$ for any specified probability.

5.3 Designing Data-Driven Uncertainty Sets

5.3.1 Geometric Characterization of the Probabilistic Guarantee

As a first step towards our schema, we provide a geometric characterization of (P2). One might intuit that a set $\mathcal{U}$ implies a probabilistic guarantee at level ϵ only if $\mathbb{P}^*(\tilde{\mathbf{u}} \in \mathcal{U}) \geq 1 - \epsilon$. As noted by other authors (cf. pg. 32-33 Ben-Tal et al. 2009)), however, this intuition is false. Often, sets that are much smaller than the $1 - \epsilon$ support will still imply a probabilistic guarantee at level ϵ, and such sets should be preferred because they are less conservative.

The crux of the issue is that there may be many realizations $\tilde{\mathbf{u}} \notin \mathcal{U}$ where nonetheless $f(\tilde{\mathbf{u}}, \mathbf{x}^*) \leq 0$. Thus, $\mathbb{P}^*(\tilde{\mathbf{u}} \in \mathcal{U})$ is in general an underestimate of $\mathbb{P}^*(f(\tilde{\mathbf{u}}, \mathbf{x}^*) \leq 0)$. One needs to exploit the dependence of f on $\mathbf{u}$ to refine the estimate. We note in passing that many existing data-driven approaches for robust optimization, e.g., Campi and Garatti (2008), do not leverage this dependence. Consequently, although these approaches are general purpose, they may yield overly conservative uncertainty sets for (5.1).

In order to tightly characterize (P2), we introduce the Value at Risk. For any $\mathbf{v} \in \mathbb{R}^d$ and measure $\mathbb{P}$, the Value at Risk at level ϵ with respect to $\mathbf{v}$ is

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \equiv \inf \left\{ t : \mathbb{P}(\tilde{\mathbf{u}}^T \mathbf{v} \leq t) \geq 1 - \epsilon \right\}. \tag{5.6}$$

Value at Risk is positively homogenous (in $\mathbf{v}$), but typically non-convex. (Recall a function $g(\mathbf{v})$ is positively homogenous if $g(\lambda \mathbf{v}) = \lambda g(\mathbf{v})$ for all $\lambda > 0$.) The critical result underlying our method is, then,

Theorem 5.1.

a) Suppose $\mathcal{U}$ is nonempty, convex and compact. Then, $\mathcal{U}$ implies a probabilistic guarantee at level ϵ for $\mathbb{P}$ for every $f(\mathbf{u}, \mathbf{x})$ concave in $\mathbf{u}$ for every $\mathbf{x}$ if

$$\delta^*(\mathbf{v} |\ \mathcal{U}) \geq VaR_\epsilon^{\mathbb{P}}(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbb{R}^d.$$

b) Suppose $\exists \mathbf{v} \in \mathbb{R}^d$ such that $\delta^*(\mathbf{v} |\ \mathcal{U}^*) < VaR_\epsilon^{\mathbb{P}}(\mathbf{v})$. Then, there exists bi-affine functions $f(\mathbf{u}, \mathbf{x})$ for which (5.2) does not hold.

The first part generalizes a result implicitly used in (Ben-Tal et al. 2009, Chen et al. 2007) when designing uncertainty sets for the special case of bi-affine functions. To the best of our knowledge, the extension to general concave functions f is new.

5.3.2 Our Schema

The principal challenge in applying Theorem 5.1 to designing uncertainty sets is that $\mathbb{P}^*$ is not known. Recall, however, that the confidence region $\mathcal{P}$ of a hypothesis test, will contain $\mathbb{P}^*$ with probability at least $1 - \alpha$. This motivates the following schema: Fix $0 < \alpha < 1$ and $0 < \epsilon < 1$.

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

2. Construct a convex, positively homogenous (in $\mathbf{v}$) upperbound $g(\mathbf{v}, \mathcal{S}, \epsilon, \alpha)$ to the worst-case Value at Risk:

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

3. Identify the convex set $\mathcal{U}(\mathcal{S}, \epsilon, \alpha)$ such that $g(\mathbf{v}, \mathcal{S}, \epsilon, \alpha) = \delta^*(\mathbf{v} | \, \mathcal{U}(\mathcal{S}, \epsilon, \alpha))$.[2]

Theorem 5.2. *With probability at least $1 - \alpha$ with respect to the sampling, the resulting set $\mathcal{U}(\mathcal{S}, \epsilon, \alpha)$ implies a probabilistic guarantee at level ϵ for $\mathbb{P}^*$.*

Remark 5.3. We note in passing that $\delta^*(\mathbf{v} | \, \mathcal{U}(\mathcal{S}, \epsilon, \alpha)) \leq t$ is a safe-approximation to the ambiguous chance constraint $\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{S}, \alpha, \epsilon)} \mathbb{P}(\mathbf{v}^T \tilde{\mathbf{u}} \leq t) \geq 1 - \epsilon$ as defined in Ben-Tal et al. (2009). Ambiguous chance-constraints are closely related to sets which imply a probabilistic guarantee. We refer the reader to Ben-Tal et al. (2009) for more details.

Theorem 5.2 ensures that with probability at least $1 - \alpha$ with respect to the sampling, a robust feasible solution $\mathbf{x}$ will satisfy a *single* uncertain constraint $f(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0$ with probability at least $1 - \epsilon$. Often, however, we face $m > 1$ uncertain constraints $f_j(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0, \quad j = 1, \ldots, m$, and seek $\mathbf{x}$ that will simultaneously satisfy these constraints, i.e.,

$$\mathbb{P}\left(\max_{j=1,\ldots,m} f_j(\tilde{\mathbf{u}}, \mathbf{x}) \leq 0 \right) \geq 1 - \bar{\epsilon}, \tag{5.7}$$

for some given $\bar{\epsilon}$. In this case, one approach is to replace each uncertain constraint with a corresponding robust constraint

$$f_j(\mathbf{u}, \mathbf{x}) \leq 0, \quad \forall \mathbf{u} \in \mathcal{U}(\mathcal{S}, \epsilon_j, \alpha), \tag{5.8}$$

where $\mathcal{U}(\mathcal{S}, \epsilon_j, \alpha)$ is constructed via our schema at level $\epsilon_j = \epsilon/m$. By the union bound and Theorem 5.2, with probability at least $1 - \alpha$ with respect to the sampling, any $\mathbf{x}$ which satisfies (5.8) will satisfy (5.7).

The choice $\epsilon_j = \epsilon/m$ is somewhat arbitrary. We would prefer to treat the ϵ_j as decision variables and optimize over them, i.e., replace the m uncertain constraints by

$$\min_{\epsilon_1 + \ldots + \epsilon_m \leq \bar{\epsilon}, \boldsymbol{\epsilon} \geq \mathbf{0}} \left\{ \max_{j=1,\ldots,m} \left\{ \max_{\mathbf{u} \in \mathcal{U}(\mathcal{S}, \epsilon_j, \alpha)} f_j(\mathbf{u}, \mathbf{x}) \right\} \right\} \leq 0$$

$$\text{or, equivalently,} \tag{5.9}$$

$$\exists \epsilon_1 + \ldots \epsilon_m \leq \bar{\epsilon}, \ \boldsymbol{\epsilon} \geq \mathbf{0} \ : f_j(\mathbf{u}, \mathbf{x}) \leq 0 \ \ \forall \mathbf{u} \in \mathcal{U}(\mathcal{S}, \epsilon_j, \alpha), \quad j = 1, \ldots, m.$$

[2]The existence of such a set in Step 3 by the bijection between closed, positively homogenous convex functions and closed convex sets in convex analysis (see Bertsekas et al. (2003)).

Unfortunately, we cannot use Theorem 5.2 to claim that with probability at least $1-\alpha$ with respect to the sampling, any feasible to solution to (5.9) will satisfy (5.7). Indeed, in general, this implication will hold with probability much less than $1-\alpha$. The issue is that Theorem 5.2 requires selecting ϵ independently of $\mathcal{S}$, whereas the optimal ϵ_j's in (5.9) *will* depend on $\mathcal{S}$, creating an in-sample bias. Consequently, we next extend Theorem 5.2 to lift this requirement.

Given a family of sets indexed by ϵ, $\{\mathcal{U}(\epsilon) : 0 < \epsilon < 1\}$, we say this family *simultaneously* implies a probabilistic guarantee for $\mathbb{P}^*$ if, for all $0 < \epsilon < 1$, each $\mathcal{U}(\epsilon)$ implies a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ. Then,

Theorem 5.4. *Suppose $\mathcal{P}(\mathcal{S}, \alpha, \epsilon) \equiv \mathcal{P}(\mathcal{S}, \alpha)$ does not depend on ϵ in Step 1 above. Let $\{\mathcal{U}(\mathcal{S}, \epsilon, \alpha) : 0 < \epsilon < 1\}$ be the resulting family of sets obtained from the our schema.*

a) With probability at least $1-\alpha$ with respect to the sampling, $\{\mathcal{U}(\mathcal{S}, \epsilon, \alpha) : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$.

b) With probability at least $1-\alpha$ with respect to the sampling, any $\mathbf{x}$ which satisfies (5.9) will satisfy (5.7).

In what follows, all of our constructions will simultaneously imply a probabilistic guarantee with the exception of $\mathcal{U}_\epsilon^M$ in Sec. 5.6. We provide numerical evidence in Sec. 5.11 that (5.9) offers significant benefit over (5.8). In some special cases, we can optimize the ϵ_j's in (5.9) exactly (see Sec. 5.11.2). More generally, we must approximate this outer optimization numerically. We postpone a treatment of this optimization problem until Sec. 5.9 after we have introduced our sets.

The next four sections apply this schema to create uncertainty sets. Often, ϵ, α and $\mathcal{S}$ are typically fixed, so we may suppress some or all of them in the notation.

5.4 Uncertainty Sets Built from Discrete Distributions

In this section, we assume $\mathbb{P}^*$ has known, finite support $\text{supp}(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$. We consider two hypothesis tests for this setup: Pearson's χ^2 test and the G test (Rice 2007). Both tests consider the hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$ where $\mathbb{P}_0$ is some specified measure. Specifically, let $p_i = \mathbb{P}_0(\tilde{\mathbf{u}} = \mathbf{a}_i)$ be the specified null-hypothesis, and let $\hat{\mathbf{p}}$ denote the empirical probability distribution , i.e.,

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

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

$$\mathcal{P}^{\chi^2} = \left\{ \mathbf{p} \in \Delta_n : \sum_{i=0}^{n-1} \frac{(p_i - \hat{p}_i)^2}{2p_i} \leq \frac{1}{2N} \chi^2_{n-1,1-\alpha} \right\}, \tag{5.10}$$

$$\mathcal{P}^{G} = \left\{ \mathbf{p} \in \Delta_n : D(\hat{\mathbf{p}}, \mathbf{p}) \leq \frac{1}{2N} \chi^2_{n-1,1-\alpha} \right\}. \tag{5.11}$$

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 a fixed measure $\mathbb{P}$, and vector $\mathbf{v} \in \mathbb{R}^d$, recall the Conditional Value at Risk:

$$\text{CVaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \equiv \min_t \left\{ t + \frac{1}{\epsilon} \mathbb{E}^{\mathbb{P}}[(\tilde{\mathbf{u}}^T \mathbf{v} - t)^+] \right\}. \tag{5.12}$$

Conditional Value at Risk is well-known to be a convex upper bound to Value at Risk (Acerbi and Tasche 2002, Rockafellar and Uryasev 2000) for a fixed $\mathbb{P}$. We can compute a bound in Step 2 by considering the worst-case Conditional Value at Risk over the above confidence regions, yielding

Theorem 5.5. *Suppose supp*$(\mathbb{P}^*) \subseteq \{\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}\}$. *With probability* $1 - \alpha$ *over the sample, the families* $\{\mathcal{U}_\epsilon^{\chi^2} : 0 < \epsilon < 1\}$ *and* $\{\mathcal{U}_\epsilon^{G} : 0 < \epsilon < 1\}$ *simultaneously imply a probabilistic guarantee for* $\mathbb{P}^*$, *where*

$$\mathcal{U}_\epsilon^{\chi^2} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j, \;\; \mathbf{q} \in \Delta_n, \;\; \mathbf{q} \leq \frac{1}{\epsilon} \mathbf{p}, \;\; \mathbf{p} \in \mathcal{P}^{\chi^2} \right\}, \tag{5.13}$$

$$\mathcal{U}_\epsilon^{G} = \left\{ \mathbf{u} \in \mathbb{R}^d : \mathbf{u} = \sum_{j=0}^{n-1} q_j \mathbf{a}_j, \;\; \mathbf{q} \in \Delta_n, \;\; \mathbf{q} \leq \frac{1}{\epsilon} \mathbf{p}, \;\; \mathbf{p} \in \mathcal{P}^{G} \right\}. \tag{5.14}$$

Their support functions are given by

$$\begin{array}{rll} \delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^{\chi^2}) = \min\limits_{\mathbf{w},\eta,\lambda,\mathbf{t}} & \beta + \dfrac{1}{\epsilon} \left(\eta + \dfrac{\lambda \chi^2_{n-1,1-\alpha}}{N} + 2\lambda - 2 \displaystyle\sum_{i=0}^{n-1} \hat{p}_i s_i \right) \\ s.t. & \mathbf{0} \leq \mathbf{w} \leq (\lambda + \eta)\mathbf{e}, \quad \lambda \geq 0, \quad \mathbf{s} \geq \mathbf{0}, \\ & \left\| \begin{matrix} 2s_i \\ w_i - \eta \end{matrix} \right\| \leq 2\lambda - w_i + \eta, \quad i = 0, \ldots, n-1 \\ & \mathbf{a}_i^T \mathbf{v} - w_i \leq \beta, \quad i = 0, \ldots, n-1, \end{array} \tag{5.15}$$

$$\begin{array}{rll} \delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^{G}) = \min\limits_{\mathbf{w},\eta,\lambda} & \beta + \dfrac{1}{\epsilon} \left(\eta + \dfrac{\lambda \chi^2_{n-1,1-\alpha}}{2N} - \lambda \displaystyle\sum_{i=0}^{n-1} \hat{p}_i \log \left(1 - \frac{w_i - \eta}{\lambda} \right) \right) \\ s.t & \mathbf{0} \leq \mathbf{w} \leq (\lambda + \eta)\mathbf{e}, \quad \lambda \geq 0, \\ & \mathbf{a}_i^T \mathbf{v} - w_i \leq \beta, \quad i = 0, \ldots, n-1. \end{array} \tag{5.16}$$

Remark 5.6. The sets $\mathcal{U}_\epsilon^{\chi^2}$, $\mathcal{U}_\epsilon^G$ strongly resemble the uncertainty set for $\text{CVaR}_\epsilon^{\hat{\mathbb{P}}}$ in Bertsimas and Brown (2009). In fact, as $N \to \infty$, all three of these sets converge almost surely to the set $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ defined by $\delta^*(\mathbf{v}|\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}) = \text{CVaR}_\epsilon^{\mathbb{P}^*}(\mathbf{v})$. The key difference is that for finite N, $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$ imply a probabilistic guarantee for $\mathbb{P}^*$ at level ϵ, while $\mathcal{U}^{\text{CVaR}_\epsilon^{\hat{\mathbb{P}}}}$ does not.

Remark 5.7. Theorem 5.5 exemplifies the distinction drawn in the introduction between uncertainty sets for discrete probability distributions – such as $\mathcal{P}^{\chi^2}$ or $\mathcal{P}^G$ which have been proposed in Ben-Tal et al. (2013) – and uncertainty sets for general uncertain parameters like $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$. The relationship between these two types of sets is explicit in eqs. (5.13) and (5.14) because we have known, finite support. For continuous support and our other sets, the relationship is implicit and must be understood through worst-case value-at-risk in Step 2 of our schema.

Remark 5.8. When considering $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}_\epsilon^{\chi^2}) \leq t\}$ or $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\, \mathcal{U}_\epsilon^G) \leq t\}$, we may drop the minimum in the formulation (5.15) or (5.16). Thus, these sets are second-order-cone representable and exponential-cone representable, respectively. Although theoretically tractable, the exponential cone can be numerically challenging.

Because of these numerical issues, modeling with $\mathcal{U}_\epsilon^{\chi^2}$ is perhaps preferable to modeling with $\mathcal{U}_\epsilon^G$. Fortunately, for large N, the difference between these two sets is negligible:

Proposition 5.9. *With arbitrarily high probability, for any* $\mathbf{p} \in \mathcal{P}^G$, $|D(\hat{\mathbf{p}}, \mathbf{p}) - \sum_{j=0}^{n-1} \frac{(\hat{p}_j - p_j)^2}{2p_j}| = O(nN^{-3})$.

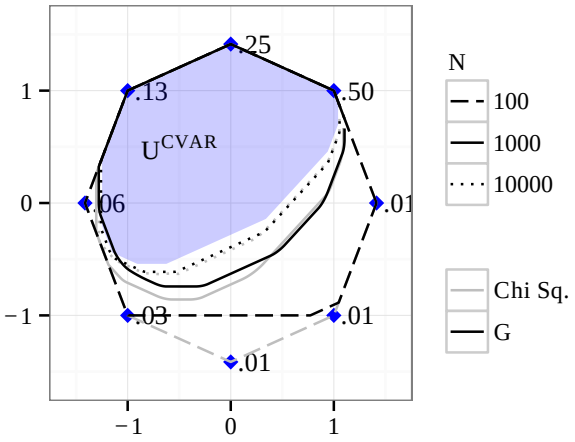
Thus, for large N, $\mathcal{P}^G$ is approximately equal to $\mathcal{P}^{\chi^2}$, whereby $\mathcal{U}_\epsilon^G$ is approximately equal to $\mathcal{U}_\epsilon^{\chi^2}$. For large N, then, $\mathcal{U}_\epsilon^{\chi^2}$ should be preferred for its computational tractability.

5.4.1 A Numerical Example of $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$

Figure 5-1 illustrates the sets $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$ 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 construct the sets $\mathcal{U}_\epsilon^{\chi^2}$ (grey line) and $\mathcal{U}_\epsilon^G$ (black line) for $\alpha = \epsilon = 10\%$ for various N. For reference, we also plot $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ (shaded region) which is the limit of both sets as $N \to \infty$. For small N, our data-driven sets are equivalent to the convex hull of $\text{supp}(\mathbb{P}^*)$, however, as N increases, our sets shrink considerably. For large N, as predicted by Propostion 5.9, $\mathcal{U}_\epsilon^G$ and $\mathcal{U}_\epsilon^{\chi^2}$ are very similarly shaped.

Remark 5.10. Fig. 5-1 also enables us to contrast our approach to that of Campi and Garatti (2008). Namely, suppose that $f(\mathbf{u}, \mathbf{x})$ is linear in $\mathbf{u}$. In this case, $\mathbf{x}$ satisfies $f(\hat{\mathbf{u}}^j, \mathbf{x}) \leq 0$ for $j = 1, \ldots, N$, if and only if $f(\mathbf{u}, \mathbf{x}) \leq 0$ for all $\mathbf{u} \in \text{conv}(\mathcal{A})$ where

Figure 5-1: The Uncertainty Sets $\mathcal{U}_\epsilon^{\chi^2}$ and $\mathcal{U}_\epsilon^G$



Note: We set $\alpha = \epsilon = 10\%$. When $N = 0$, the smallest set which implies a probabilistic guarantee is $\text{supp}(\mathbb{P}^*)$, the given octagon. As N increases, both sets shrink to the $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ given by the shaded region.

$\mathcal{A} \equiv \{\mathbf{a} \in \text{supp}(\mathbb{P}^*) : \exists 1 \leq j \leq N \text{ s.t. } \mathbf{a} = \hat{\mathbf{u}}^j\}$. As $N \to \infty$, $\mathcal{A} \to \text{supp}(\mathbb{P}^*)$ almost surely. In other words, as $N \to \infty$, the method of Campi and Garatti (2008) in this case is equivalent to using the entire support as an uncertainty set, which is much larger than $\mathcal{U}^{\text{CVaR}_\epsilon^{\mathbb{P}^*}}$ above. Similar examples can be constructed with continuous distributions or the method of Calafiore and Monastero (2012). In each case, the critical observation is that these methods do not explicitly leverage the concave (or, in this case, linear) structure of $f(\mathbf{u}, \mathbf{x})$.

5.5 Independent Marginal Distributions

We next consider the case where $\mathbb{P}^*$ may have continuous support, but the marginal distributions $\mathbb{P}_i^*$ are known to be independent. Our strategy is to build up a multivariate test by combining univariate tests for each marginal distribution.

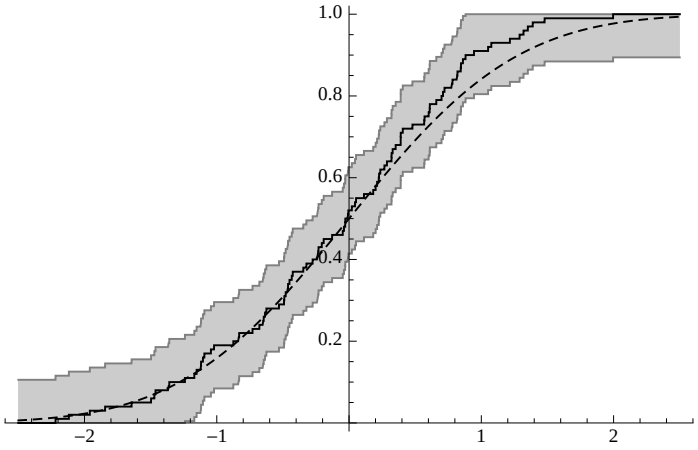
5.5.1 Uncertainty Sets Built from the Kolmogorov-Smirnov Test

For this section, we assume that $\text{supp}(\mathbb{P}^*)$ is contained in a known, bounded box

$$[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] \equiv \{\mathbf{u} \in \mathbb{R}^d : \hat{u}_i^{(0)} \leq u_i \leq \hat{u}_i^{(N+1)}, \quad i = 1, \ldots, d\}.$$

Given a univariate measure $\mathbb{P}_{0,i}$, the Kolmogorov-Smirnov (KS) goodness-of fit

Figure 5-2: The Empirical Distribution Function and Confidence Region Corresponding to the KS Test



test applied to marginal i considers the null-hypothesis $H_0 : \mathbb{P}_i^* = \mathbb{P}_{0,i}$. It rejects this hypothesis if

$$\max_{j=1,\ldots,N} \max\left(\frac{j}{N} - \mathbb{P}_{0,i}(\tilde{u} \le \hat{u}_i^{(j)}), \mathbb{P}_{0,i}(\tilde{u} < \hat{u}_i^{(j)}) - \frac{j-1}{N}\right) > \Gamma^{KS}.$$

where $\hat{u}_i^{(j)}$ is the j^{th} largest element among $\hat{u}_i^1, \ldots, \hat{u}_i^N$. Tables for the threshold Γ^{KS} are widely available (Stephens 1974, Thas 2009).

The confidence region of the above test for the i-th marginal distribution is

$$\mathcal{P}_i^{KS} = \left\{ \mathbb{P}_i \in \Theta[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}] : \begin{array}{ll} \mathbb{P}_i(\tilde{u}_i \le \hat{u}_i^{(j)}) \ge \frac{j}{N} - \Gamma^{KS}, & j = 1, \ldots, N \\ \mathbb{P}_i(\tilde{u}_i < \hat{u}_i^{(j)}) \le \frac{j-1}{N} + \Gamma^{KS}, & j = 1, \ldots, N \end{array} \right\},$$

where $\Theta[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}]$ is the set of all Borel probability measures on $[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}]$. Unlike $\mathcal{P}^{\chi^2}$ and $\mathcal{P}^G$, this confidence region is infinite dimensional.

Figure 5-2 illustrates an example. The true distribution is a standard normal whose cumulative distribution function (cdf) is the dotted line. We draw $N = 100$ data points and form the empirical cdf (solid black line). The 80% confidence region of the KS test is the set of measures whose cdfs are more than Γ^{KS} above or below this solid line, i.e. the grey region.

Now consider the multivariate null-hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$. Since $\mathbb{P}^*$ has independent components, the test which rejects if $\mathbb{P}_i$ fails the KS test at level $\alpha' = 1 - \sqrt[d]{1-\alpha}$ for any i is a valid test. Namely, $\mathbb{P}_{\mathcal{S}}^*(\mathbb{P}_i^*$ is accepted by KS at level α' for all $i = 1, \ldots, d) = \prod_{i=1}^d \sqrt[d]{1-\alpha'} = 1 - \alpha$ by independence. The confidence region of this multivariate test is

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

("I" in $\mathcal{P}^I$ is to emphasize independence). We use this confidence region in Step 1 of our schema.

When the marginals are independent, Nemirovski and Shapiro (2006) proved

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \leq \inf_{\lambda \geq 0} \left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log \mathbb{E}^{\mathbb{P}_i}[e^{v_i \tilde{u}_i/\lambda}] \right).$$

We use the worst-case value of this bound over $\mathcal{P}^I$ in Step 2 of our schema. By passing the supremum through the infimum and logarithm, we obtain

$$\sup_{\mathbb{P} \in \mathcal{P}^I} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \leq \inf_{\lambda \geq 0} \left(\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log \sup_{\mathbb{P}_i \in \mathcal{P}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}[e^{v_i \tilde{u}_i/\lambda}] \right). \tag{5.17}$$

Despite the infinite dimensionality, we can solve in the inner-most supremum explicitly by leveraging the simple geometry of $\mathcal{P}_i^{KS}$. Intuitively, the worst-case distribution will either be the lefthand boundary or the righthand boundary of the region in Fig. 5-2 depending on the sign of v_i.

Specifically, define

$$q_j^L(\Gamma) = \begin{cases} \Gamma & \text{if } j = 0, \\ \frac{1}{N} & \text{if } 1 \leq j \leq \lfloor N(1-\Gamma) \rfloor, \\ 1 - \Gamma - \frac{\lfloor N(1-\Gamma) \rfloor}{N} & \text{if } j = \lfloor N(1-\Gamma) \rfloor + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{5.18}$$

$$q_j^R(\Gamma) = q_{N+1-j}^L(\Gamma), \quad j = 0, \ldots, N+1. \tag{5.19}$$

Both $\mathbf{q}^L(\Gamma), \mathbf{q}^R(\Gamma) \in \Delta_{N+2}$ so that each vector can be interpreted as a discrete probability distribution on the points $\hat{u}_i^{(0)}, \ldots, \hat{u}_i^{(N+1)}$. One can check that the distributions corresponding to these vectors are precisely the lefthand side and righthand side of the grey region in Fig. 5-2. Then, we have

Theorem 5.11. *Suppose $\mathbb{P}^*$ has independent components, with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. With probability at least $1-\alpha$ with respect to the sampling, $\{\mathcal{U}_\epsilon^I : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\begin{aligned} \mathcal{U}_\epsilon^I = \Bigg\{ & \mathbf{u} \in \mathbb{R}^d : \exists \theta_i \in [0,1],\ \mathbf{q}^i \in \Delta_{N+2},\ i = 1 \ldots, d, \\ & \sum_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i = u_i,\ i = 1, \ldots, d, \\ & \sum_{i=1}^{d} D(\mathbf{q}_i, \theta_i \mathbf{q}^L(\Gamma^{KS}) + (1-\theta_i)\mathbf{q}^R(\Gamma^{KS})) \leq \log(1/\epsilon) \Bigg\}. \end{aligned} \tag{5.20}$$

Moreover,

$$\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^I) = \inf_{\lambda\geq 0}\Bigg\{\lambda\log(1/\epsilon) + \lambda\sum_{i=1}^{d}\log\left[\max\left(\sum_{j=0}^{N+1} q_j^L(\Gamma^{KS})e^{v_i\hat{u}_i^{(j)}/\lambda}, \sum_{j=0}^{N+1} q_j^R(\Gamma^{KS})e^{v_i\hat{u}_i^{(j)}/\lambda}\right)\right]\Bigg\} \tag{5.21}$$

Remark 5.12. Because $\mathbf{q}^L(\Gamma)$ (resp. $\mathbf{q}^R(\Gamma)$) is decreasing (resp. increasing) in its components, the lefthand branch of the innermost maximum in (5.21) will be attained when $v_i \leq 0$ and the righthand branch is attained otherwise. Thus, for fixed $\mathbf{v}$, the optimization problem in λ is convex and differentiable and can be efficiently solved with a line search.

Remark 5.13. When representing $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U}^I) \leq t)\}$, we can drop the infimum in (5.21). Thus, this set is exponential cone representable, which, again, may be numerically challenging. Using the above line search, however, we can separate over this set: Given $\mathbf{v}\in\mathbb{R}^d, t\in\mathbb{R}$ such that $\delta^*(\mathbf{v}|\,\mathcal{U}^I) > t$, solve (5.21) by line search, and let λ^* be an optimal solution. Define

$$\mathbf{p}^i = \begin{cases}\mathbf{q}^L & \text{if } v_i \leq 0,\\ \mathbf{q}^R & \text{otherwise,}\end{cases} \qquad q_j^i = \frac{p_j^i e^{v_i\hat{u}_i^{(j)}/\lambda}}{\sum_{j=0}^{N+1} p_j^i e^{v_i\hat{u}_i^{(j)}/\lambda}}, \quad j=0,\ldots,N+1, \quad i=1,\ldots,d,$$

$$u_i = \sum_{j=0}^{N+1} q_j^i\hat{u}_i^{(j)}, \quad i=1\ldots,d.$$

Then $\mathbf{u}\in\mathcal{U}_\epsilon^I$ and $\mathbf{u}^T\mathbf{v}\leq t$ is a violated cut for $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^I)\leq t\}$. That this procedure is valid follows from the proof of Theorem 5.11, see appendix.

Remark 5.14. The KS test 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 (Thas 2009, Chapt. 5). We can define analogues of $\mathcal{U}_\epsilon^I$ for each of these tests, each having slightly different shape. Separating over $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U})\leq t\}$ is polynomial time tractable for each these sets, but we no longer have a simple algorithm for generating violated cuts. Thus, these sets are considerably less attractive from a computational point of view. Fortunately, through simulation studies with a variety of different distributions, we have found that the version of $\mathcal{U}_\epsilon^I$ based on the KS test generally performs as well as or better than the other EDF tests. Consequently, we recommend using the sets $\mathcal{U}_\epsilon^I$ as described. For completeness, we present the constructions for the analogous tests in Appendix D.6.

5.5.2 Uncertainty Sets Motivated by Forward and Backward Deviations

In Chen et al. (2007), the authors propose an uncertainty set based on the forward and backward deviations of a distribution. They focus on a non-data-driven setting, where the mean and support of $\mathbb{P}^*$ are known a priori, and show how to upper bound these deviations to calibrate their set. In a setting where one has data *and a priori knows the mean of* $\mathbb{P}^*$ *precisely*, they propose a method based on sample average approximation to estimate these deviations. Unfortunately, the precise statistical behavior of these estimators is not known, so it is not clear that this set calibrated from data implies a probabilistic guarantee with high probability with respect to the sampling.

In this section, we use our schema to generalize the set of Chen et al. (2007) to a data-driven setting where *neither the mean of the distribution nor its support are known.* Our set differs in shape and size from their proposal, and, our construction, unlike their original proposal, will simultaneously imply a probabilistic guarantee for $\mathbb{P}^*$.

We begin by specifying an appropriate multivariate hypothesis test based on combining univariate tests. Specifically, for a known (univariate) distribution $\mathbb{P}_i$ define its forward and backward deviations by

$$\sigma_{fi}(\mathbb{P}_i) = \sup_{x>0} \sqrt{-\frac{2\mu_i}{x} + \frac{2}{x^2}\log(\mathbb{E}^{\mathbb{P}_i}[e^{x\tilde{u}_i}])}, \quad \sigma_{bi}(\mathbb{P}_i) = \sup_{x>0} \sqrt{\frac{2\mu_i}{x} + \frac{2}{x^2}\log(\mathbb{E}^{\mathbb{P}_i}[e^{-x\tilde{u}_i}])}, \tag{5.22}$$

where $\mathbb{E}^{\mathbb{P}_i}[\tilde{u}_i] = \mu_i$. Notice the optimizations defining $\sigma_{fi}(\mathbb{P}_i), \sigma_{bi}(\mathbb{P}_i)$ are one dimensional, convex problems which can be solved by a line search. A sufficient, but not necessary, condition for $\sigma_{fi}(\mathbb{P}_i), \sigma_{bi}(\mathbb{P}_i)$ to be finite is that $\mathbb{P}_i$ has bounded support (c.f. Chen et al. 2007). To streamline the exposition, we assume throughout this section $\mathbb{P}^*$ has bounded (but potentially unknown) support.

For a given $\mu_{0,i}, \sigma_{0,fi}, \sigma_{0,bi} \in \mathbb{R}$, consider the following three null-hypotheses:

$$H_0^1 : \mathbb{E}^{\mathbb{P}_i^*}[\tilde{u}] = \mu_{0,i}, \quad H_0^2 : \sigma_{fi}(\mathbb{P}_i^*) \leq \sigma_{0,fi}, \quad H_0^3 : \sigma_{bi}(\mathbb{P}_i^*) \leq \sigma_{0,bi}. \tag{5.23}$$

We can test these hypotheses (separately) using $|\hat{\mu}_i - \mu_{0,i}|$, $\sigma_{fi}(\hat{\mathbb{P}}_i)$ and $\sigma_{bi}(\hat{\mathbb{P}}_i)$, respectively, as test statistics. Since these are not common hypothesis tests in applied statistics, there are no tables for their thresholds. Instead, we compute approximate thresholds t_i, $\overline{\sigma}_{fi}$ and $\overline{\sigma}_{bi}$ at the $\alpha/2$, $\alpha/4$ and $\alpha/4$ significance level, respectively, using the bootstrap procedure in Algorithm 1.

By the union bound, the univariate test which rejects if any of these thresholds is exceeded is a valid test at level α for the three hypotheses above to hold simultaneously. The confidence region of this test is

$$\mathcal{P}_i^{FB} = \{\mathbb{P}_i \in \Theta(-\infty, \infty) : m_{bi} \leq \mathbb{E}_i^{\mathbb{P}}[\tilde{u}_i] \leq m_{fi}, \quad \sigma_{fi}(\mathbb{P}_i) \leq \overline{\sigma}_{fi}, \quad \sigma_{bi}(\mathbb{P}_i) \leq \overline{\sigma}_{bi}\},$$

where $m_{bi} = \hat{\mu}_i - t_i$ and $m_{fi} = \hat{\mu}_i + t_i$.

Next, consider the multivariate null-hypothesis that all three null-hypotheses in (5.23) hold simultaneously for all $i = 1, \ldots, d$. As in Sec. 5.5, the test which rejects if the above univariate test rejects at level $\alpha' = 1 - \sqrt[d]{1-\alpha}$ for any i is a valid test. Its confidence region is $\mathcal{P}^{FB} = \{\mathbb{P} : \mathbb{P}_i \in \mathcal{P}_i^{FB}\ i = 1, \ldots, d\}$. We will use this confidence region in Step 1 of our schema.

When the mean and deviations for $\mathbb{P}$ are known and the marginals are independent, Chen et al. (2007) prove

$$\mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \leq \sum_{i=1}^{d} \mathbb{E}^{\mathbb{P}}[\tilde{u}_i] v_i + \sqrt{2\log(1/\epsilon)\left(\sum_{i:v_i<0} \sigma_{bi}^2(\mathbb{P}) v_i^2 + \sum_{i:v_i\geq 0} \sigma_{fi}^2(\mathbb{P}) v_i^2\right)}. \tag{5.24}$$

Computing the worst-case value of this bound over the above confidence region in Step 2 of our schema yields:

Theorem 5.15. *Suppose $\mathbb{P}^*$ has independent components and bounded support. With probability $1-\alpha$ with respect to the sample, the family $\{\mathcal{U}_\epsilon^{FB} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\mathcal{U}_\epsilon^{FB} = \left\{ \mathbf{y}_1 + \mathbf{y}_2 - \mathbf{y}_3 : \begin{array}{l} \mathbf{y}_2, \mathbf{y}_2 \in \mathbb{R}_+^d, \\ \sum_{i=1}^d \frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2} + \frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2} \leq \log(1/\epsilon), \\ m_{bi} \leq y_{1i} \leq m_{fi}, \quad i = 1, \ldots, d \end{array} \right\}. \tag{5.25}$$

Moreover,

$$\delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^{FB}) = \sum_{i:v_i\geq 0} m_{fi} v_i + \sum_{i:v_i<0} m_{bi} v_i + \sqrt{2\log(1/\epsilon)\left(\sum_{i:v_i\geq 0} \overline{\sigma}_{fi}^2 v_i^2 + \sum_{i:v_i<0} \overline{\sigma}_{bi}^2 v_i^2\right)} \tag{5.26}$$

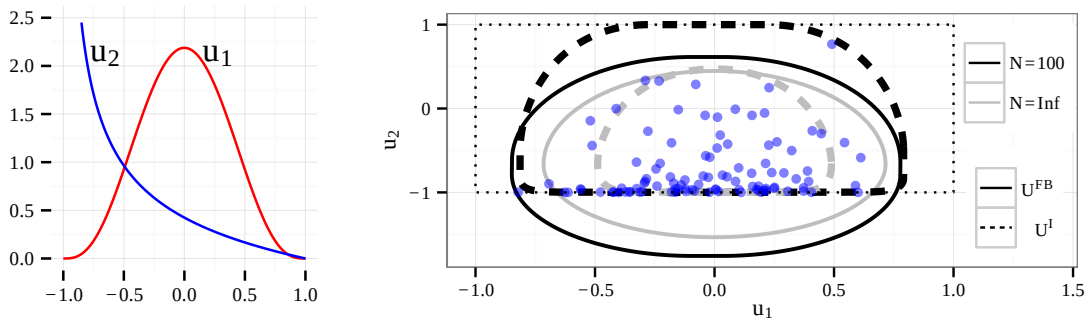
Remark 5.16. From (5.26), $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^{FB}) \leq t\}$ is second order cone representable. We can separate over this constraint in closed-form: Given $\mathbf{v}, t$, use (5.26) to check if $\delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^{FB}) > t$. If so, let

$$\lambda = \sqrt{\frac{\sum_{i:v_i>0} v_i^2 \overline{\sigma}_{fi}^2 + \sum_{i:v_i\leq 0} v_i^2 \overline{\sigma}_{bi}^2}{2\log(1/\epsilon)}}, \qquad u_i = \begin{cases} m_{fi} + \frac{v_i \overline{\sigma}_{fi}^2}{\lambda} & \text{if } v_i > 0 \\ m_{bi} + \frac{v_i \overline{\sigma}_{bi}^2}{\lambda} & \text{otherwise.} \end{cases}$$

Then, $\mathbf{u}^T \mathbf{v} \leq t$ is a violated constraint. The correctness of this procedure follows from the proof of Theorem 5.15.

Remark 5.17. There is no guarantee that $\mathcal{U}_\epsilon^{FB} \subseteq \mathrm{supp}(\mathbb{P}^*)$. Consequently, if we have a priori information of the support, we can use this to refine $\mathcal{U}_\epsilon^{FB}$. Specifically, let $\mathcal{U}_0$ be convex, compact such that $\mathrm{supp}(\mathbb{P}^*) \subseteq \mathcal{U}_0$. Then, the family $\{\mathcal{U}_\epsilon^{FB} \cap \mathcal{U}_0 : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee. Moreover, for common $\mathcal{U}_0$, optimizing over (5.3) with $\mathcal{U}_\epsilon^{FB} \cap \mathcal{U}_0$ is computationally similar to optimizing with $\mathcal{U}_\epsilon^{FB}$. More precisely, from (Ben-Tal et al. 2012, Lemma A.4), $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon(\mathcal{S}) \cap \mathcal{U}_0)\}$ is

Figure 5-3: Comparison of $\mathcal{U}_\epsilon^I$ and $\mathcal{U}_\epsilon^{FB}$



Note: The left panel shows the marginal densities. The right panel shows $\mathcal{U}_\epsilon^I$ (dashed black line) and $\mathcal{U}_\epsilon^{FB}$ (solid black line) built from $N = 100$ data points (blue circles) and in the limit as $N \to \infty$ (corresponding gray lines).

equivalent to

$$\left\{(\mathbf{v}, t) : \exists \mathbf{w}, \in \mathbb{R}^d,\, t_1, t_2 \in \mathbb{R} \text{ s.t. } \delta^*(\mathbf{v} - \mathbf{w} | \mathcal{U}_\epsilon(\mathcal{S})) \le t_1,\, \delta^*(\mathbf{w} |\, \mathcal{U}_0) \le t_2,\, t_1 + t_2 \le t\right\}, \tag{5.27}$$

so that (5.3) with $\mathcal{U}_\epsilon^{FB} \cap \mathcal{U}_0$ will be tractable whenever $\{(\mathbf{v}, t) : \delta^*(\mathbf{v} |\, \mathcal{U}_0) \le t\}$ is tractable, examples of which include when $\mathcal{U}_0$ is a norm-ball, ellipse, or polyhedron (see Ben-Tal et al. (2012)).

5.5.3 Comparing $\mathcal{U}_\epsilon^I$ and $\mathcal{U}_\epsilon^{FB}$

Figure 5-3 illustrates the sets $\mathcal{U}_\epsilon^I$ and $\mathcal{U}_\epsilon^{FB}$ numerically. The marginal distributions of $\mathbb{P}^*$ are independent and their densities are given in the left panel. Notice that the first marginal is symmetric while the second is highly skewed.

In the absence of any data, knowing only supp($\mathbb{P}^*$) and that $\mathbb{P}^*$ has independent components, the smallest uncertainty which implies a probabilistic guarantee is the unit square (dotted line). With $N = 100$ data points from this distribution (blue circles), however, we can construct both $\mathcal{U}_\epsilon^I$ (dashed black line) and $\mathcal{U}_\epsilon^{FB}$ (solid black line) with $\epsilon = \alpha = 10\%$, as shown. We also plot the limiting shape of these two sets as $N \to \infty$ (corresponding grey lines).

Several features are evident from the plots. First, both sets are able to learn that $\mathbb{P}^*$ is symmetric in its first coordinate (the sets display vertical symmetry) and that $\mathbb{P}^*$ is skewed downwards in its second coordinate (the sets taper more sharply towards the top). Both sets *learn* these features from the data. Second, although $\mathcal{U}_\epsilon^I$ is a strict subset of supp($\mathbb{P}^*$), $\mathcal{U}_\epsilon^{FB}$ is not (see also Remark 5.17). Finally, neither set is a subset of the other, and, although for $N = 100$, $\mathcal{U}_\epsilon^{FB} \cap$ supp($\mathbb{P}^*$) has smaller volume than $\mathcal{U}_\epsilon^I$, the reverse holds for larger N. Consequently, it is not clear which set to prefer in a given application, and the best choice likely depends on N.

5.6 Uncertainty Sets Built from Marginal Samples

In this section, we observe samples from the marginal distributions of $\mathbb{P}^*$ separately, but do not assume these marginals are independent. This happens, e.g., when samples are drawn asynchronously, or when there are many missing values. In these cases, it is impossible to learn 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.

In the univariate case, David and Nagaraja (1970) develop a hypothesis test for the $1-\epsilon/d$ quantile, or equivalently $\mathrm{VaR}^{\mathbb{P}_i}_{\epsilon/d}(\mathbf{e}_i)$ of a distribution $\mathbb{P}$. Namely, given $\overline{q}_{i,0} \in \mathbb{R}$, consider the hypothesis $H_{0,i} : \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) \geq \overline{q}_{i,0}$. Define the index s by

$$s = \min\left\{ k \in \mathbb{N} : \sum_{j=k}^{N} \binom{N}{j} (\epsilon/d)^{N-j}(1-\epsilon/d)^j \leq \frac{\alpha}{2d} \right\}, \tag{5.28}$$

and let $s = N+1$ if the corresponding set is empty. Then, the test which rejects if $q_{i,0} > \hat{u}_i^{(s)}$ is valid at level $\alpha/2d$ (David and Nagaraja 1970, Sec. 7.1). David and Nagaraja (1970) also prove that $\frac{s}{N} \downarrow (1-\epsilon/d)$.

The above argument applies symmetrically to the hypothesis $H_{0,i} : \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\mathbf{e}_i) \geq \underline{q}_{i,0}$ where the rejection threshold now becomes $\hat{u}_i^{(N-s+1)}$. In the typical case when ϵ/d is small, $N-s+1 < s$ so that $\hat{u}_i^{(N-s+1)} \leq \hat{u}_i^{(s)}$.

Next given $\overline{q}_{i,0}, \underline{q}_{i,0} \in \mathbb{R}$ for $i = 1, \ldots, d$, consider the multivariate hypothesis:

$$H_0 : \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(\mathbf{e}_i) \geq \overline{q}_{i,0} \text{ and } \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon/d}(-\mathbf{e}_i) \geq \underline{q}_{i,0} \text{ for all } i = 1, \ldots, d.$$

By the union bound, the test which rejects if $\hat{u}_i^{(s)} < \overline{q}_i$ or $-\hat{u}_i^{(N-s+1)} < \underline{q}_i$, i.e., the above tests fail for the i-th component, is valid at level α. Its confidence region is

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

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

When the marginals of $\mathbb{P}$ are known, Embrechts et al. (2003) proves

$$\mathrm{VaR}^{\mathbb{P}}_{\epsilon}(\mathbf{v}) \leq \min_{\boldsymbol{\lambda}: \mathbf{e}^T\boldsymbol{\lambda} = \epsilon} \sum_{i=1}^{d} \mathrm{VaR}^{\mathbb{P}}_{\lambda_i}(v_i \mathbf{e}_i). \tag{5.29}$$

Since the minimization on the right-hand side can be difficult, we will use the weaker bound $\mathrm{VaR}^{\mathbb{P}}_{\epsilon}(\mathbf{v}) \leq \sum_{i=1}^{d} \mathrm{VaR}^{\mathbb{P}}_{\epsilon/d}(v_i \mathbf{e}_i)$ obtained by letting $\lambda_i = \epsilon/d$ for all i.

We compute the worst case value of this bound over $\mathcal{P}^M$, yielding:

Theorem 5.18. *If s defined by Eq. (5.28) satisfies $N-s+1 < s$, then, with probability*

at least $1-\alpha$ *over the sample, the set*

$$\mathcal{U}_\epsilon^M = \left\{ \mathbf{u} \in \mathbb{R}^d : \hat{u}_i^{(N-s+1)} \leq u_i \leq \hat{u}_i^{(s)} \quad i = 1, \ldots, d \right\}. \tag{5.30}$$

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

$$\delta^*(\mathbf{v} |\, \mathcal{U}_\epsilon^M) = \sum_{i=1}^{d} \max(v_i \hat{u}_i^{(N-s+1)}, v_i \hat{u}_i^{(s)}). \tag{5.31}$$

Remark 5.19. Notice that the family $\{\mathcal{U}_\epsilon^M : 0 < \epsilon < 1\}$, may *not* simultaneously imply a probabilistic guarantee for $\mathbb{P}^*$ because the confidence region $\mathcal{P}^M$ depends on ϵ.

Remark 5.20. The set $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\mathcal{U}^M) \leq t\}$ is a simple box, representable by linear inequalities. We can separate over this set in closed form via (5.31).

5.7 Uncertainty Sets for Potentially Non-independent Components

In this section, we assume we observe samples drawn from the joint distribution of $\mathbb{P}^*$ which may have unbounded support. We consider a goodness-of-fit hypothesis test based on linear-convex ordering proposed in Bertsimas et al. (2014b). Specifically, given some multivariate $\mathbb{P}_0$, consider the null-hypothesis $H_0 : \mathbb{P}^* = \mathbb{P}_0$. Bertsimas et al. (2014b) prove that the test which rejects H_0 if $\exists(\mathbf{a}, b) \in \mathcal{B} \equiv \{\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R} : \|\mathbf{a}\|_1 + |b| \leq 1\}$ such that

$$\mathbb{E}^{\mathbb{P}_0}[(\mathbf{a}^T\tilde{\mathbf{u}} - b)^+] - \frac{1}{N}\sum_{j=1}^{N}(\mathbf{a}^T\hat{\mathbf{u}}^j - b)^+ > \Gamma_{LCX} \quad \text{or} \quad \frac{1}{N}\sum_{j=1}^{N}(\hat{\mathbf{u}}^j)^T\hat{\mathbf{u}}^j - \mathbb{E}^{\mathbb{P}_0}[\tilde{\mathbf{u}}^T\tilde{\mathbf{u}}] > \Gamma_\sigma$$

for appropriate thresholds $\Gamma_{LCX}, \Gamma_\sigma$ is a valid test at level α. The authors provide an explicit bootstrap algorithm to compute $\Gamma_{LCX}, \Gamma_\sigma$.

The confidence region of this test is

$$\begin{aligned}\mathcal{P}^{LCX} = \Bigg\{ \mathbb{P} \in \Theta(\mathbb{R}^d) : \mathbb{E}^{\mathbb{P}}[(\mathbf{a}^T\tilde{\mathbf{u}} - b)^+] \leq \frac{1}{N}\sum_{j=1}^{N}(\mathbf{a}^T\hat{\mathbf{u}}_j - b)^+ + \Gamma_{LCX} \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\ \sum_{i=1}^{d}\mathbb{E}^{\mathbb{P}}[\|\tilde{\mathbf{u}}\|^2] \geq \frac{1}{N}\sum_{j=1}^{N}\|\hat{\mathbf{u}}_j\|^2] - \Gamma_\sigma \Bigg\},\end{aligned} \tag{5.32}$$

We will use this confidence region in Step 1 of our schema.

Combining techniques from semi-infinite optimization with our schema (see electronic companion for proof), we obtain

Theorem 5.21. *The family* $\{\mathcal{U}_\epsilon^{LCX} : 0 < \epsilon < 1\}$ *simultaneously implies a probabilistic guarantee for* $\mathbb{P}^*$ *where*

$$\mathcal{U}_\epsilon^{LCX} = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \ \exists \mathbf{r} \in \mathbb{R}^d, \ 1 \le z \le 1/\epsilon, \quad s.t. \tag{5.33a}$$

$$(\mathbf{a}^T\mathbf{r} - b(z-1))^+ + (\mathbf{a}^T\mathbf{u} - b)^+ \le \frac{z}{N}\sum_{j=1}^N (\mathbf{a}^T\hat{\mathbf{u}}_j - b)^+ + \Gamma_{LCX}, \ \forall (\mathbf{a}, b) \in \mathcal{B} \Bigg\}. \tag{5.33b}$$

Moreover,

$$\begin{aligned}
\delta^*(\mathbf{v} |\ \mathcal{U}_\epsilon^{LCX}) = \sup_{\mathbb{P} \in \mathcal{P}^{LCX}} VaR_\epsilon^{\mathbb{P}}(\mathbf{v}) = \min_{\tau,\theta,y_1,y_2,\lambda} \quad & \frac{1}{\epsilon}\tau - \theta - \int_{\mathcal{B}} b dy_1(\mathbf{a}, b) + \int_{\mathcal{B}} b dy_2(\mathbf{a}, b) \\
s.t. \quad & \theta + \int_{\mathcal{B}} b dy_1(\mathbf{a}, b) + \int_{\mathcal{B}} \Gamma(\mathbf{a}, b) d\lambda(\mathbf{a}, b) \le \tau \\
& 0 \le dy_1(\mathbf{a}, b) \le d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& 0 \le dy_2(\mathbf{a}, b) \le d\lambda(\mathbf{a}, b) \quad \forall (\mathbf{a}, b) \in \mathcal{B}, \\
& \int_{\mathcal{B}} \mathbf{a}\ dy_1(\mathbf{a}, b) = 0, \quad \mathbf{v} = \int_{\mathcal{B}} \mathbf{a}\ dy_2(\mathbf{a}, b), \\
& \theta, \tau \ge 0.
\end{aligned} \tag{5.34}$$

Remark 5.22. As the intersection of convex constraints, $\mathcal{U}_\epsilon^{LCX}$ is convex.

Remark 5.23. It is possible to separate over (5.33b) efficiently. Specifically, fix $\mathbf{u}, \mathbf{r} \in \mathbb{R}^d$ and $1 \le z \le 1/\epsilon$. We identify the worst-case $(\mathbf{a}, b) \in \mathcal{B}$ in (5.33b) by solving three auxiliary optimization problems:

$$\begin{aligned}
\xi_1 = \max_{(\mathbf{a},b) \in \mathcal{B}, \mathbf{t} \ge \mathbf{0}} \quad & \mathbf{a}^T\mathbf{r} - b(z-1)) + (\mathbf{a}^T\mathbf{u} - b) - \frac{z}{N}\sum_{j=1}^N t_j \\
\text{s.t.} \quad & t_j \ge \mathbf{a}^T\hat{\mathbf{u}}_j - b, \quad \mathbf{a}^T\mathbf{u} - b \ge 0, \quad \mathbf{a}^T\mathbf{r} - b(z-1) \ge 0, \\
\xi_2 = \max_{(\mathbf{a},b) \in \mathcal{B}, \mathbf{t} \ge \mathbf{0}} \quad & \mathbf{a}^T\mathbf{r} - b(z-1)) - \frac{z}{N}\sum_{j=1}^N t_j \\
\text{s.t.} \quad & t_j \ge \mathbf{a}^T\hat{\mathbf{u}}_j - b, \quad \mathbf{a}^T\mathbf{u} - b \le 0, \quad \mathbf{a}^T\mathbf{r} - b(z-1) \ge 0, \\
\xi_3 = \max_{(\mathbf{a},b) \in \mathcal{B}, \mathbf{t} \ge \mathbf{0}} \quad & (\mathbf{a}^T\mathbf{u} - b) - \frac{z}{N}\sum_{j=1}^N t_j \\
\text{s.t.} \quad & t_j \ge \mathbf{a}^T\hat{\mathbf{u}}_j - b, \quad \mathbf{a}^T\mathbf{u} - b \ge 0, \quad \mathbf{a}^T\mathbf{r} - b(z-1) \le 0,
\end{aligned}$$

corresponding to the potential signs of $\mathbf{a}^T\mathbf{r} - b(z-1)$ and $\mathbf{a}^T\mathbf{u} - b$ at the worst-case value. (The fourth case, where both terms are negative is trivial since $\Gamma_{LCX} > 0$.) Each of these optimization problems can be written as linear optimizations. If

$\max(\xi_1, \xi_2, \xi_3) \leq \Gamma_{LCX}$, then $\mathbf{u}, \mathbf{r}$ and z are feasible in (5.33b). Otherwise, the optimal $\mathbf{a}, b$ in the maximizing subproblem yields a violated cut.

Remark 5.24. The representation of $\delta^*(\mathbf{v}|\ \mathcal{U}^{LCX})$ is not particularly convenient. Nonetheless, we can separate over $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\,\mathcal{U}^{LCX}) \leq t\}$ in polynomial time by using the above separation routine with the ellipsoid algorithm to solve $\max_{\mathbf{u}\in\mathcal{U}^{LCX}} \mathbf{v}^T\mathbf{u}$. Alternatively, combining the above separation routine with the dual-simplex algorithm yields a practically efficient algorithm for large-scale instances

5.8 Hypothesis Testing: A Unifying Perspective

Several data-driven methods in the literature create families of measures $\mathcal{P}(\mathcal{S})$ that contain $\mathbb{P}^*$ with high probability. These methods do not explicitly reference hypothesis testing. In this section, we provide a hypothesis testing interpretation of two such methods (Shawe-Taylor and Cristianini 2003, Delage and Ye 2010). Leveraging this new perspective, we show how standard techniques for hypothesis testing, such as the bootstrap, can be used to improve upon these methods. Finally, we illustrate how our schema can be applied to these improved family of measures to generate new uncertainty sets. To the best of our knowledge, generating uncertainty sets for (5.1) is a new application of both (Shawe-Taylor and Cristianini 2003, Delage and Ye 2010).

The key idea in both cases is to recast $\mathcal{P}(\mathcal{S})$ as the confidence region of a hypothesis test. This correspondence is not unique to these methods. There is a one-to-one correspondence between families of measures which contain $\mathbb{P}^*$ with probability at least $1-\alpha$ 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 (Rice 2007). It implies that any data-driven method predicated on a family of measures that contain $\mathbb{P}^*$ with probability $1-\alpha$ can be interpreted in the light of hypothesis testing.

This observation is interesting for two reasons. First, it provides a unified framework to compare distinct methods in the literature and ties them to the well-established theory of hypothesis testing in statistics. Secondly, there is a wealth of practical experience with hypothesis testing. In particular, 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 strengthen these methods, and then derive uncertainty sets corresponding to these hypothesis tests to facilitate comparison between the approaches.

5.8.1 Uncertainty Set Motivated by Cristianini and Shawe-Taylor, 2003

Let $\|\cdot\|_F$ denote the Frobenius norm of matrices. As part of a particular machine learning application, Shawe-Taylor and Cristianini (2003) prove

Theorem 5.25 (Cristianini and Shawe-Taylor, 2003)**.** *Suppose that supp*$(\mathbb{P}^*)$ *is contained within the ball of radius* R *and that* $N > (2+2\log(2/\alpha))^2$*. Then, with probability at least* $1-\alpha$ *with respect to the sampling,* $\mathbb{P}^* \in \mathcal{P}^{CS}$ *for*

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

where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ *denote the sample mean and covariance,* $\Gamma_1(\alpha, N) = \frac{R}{\sqrt{N}}\left(2 + \sqrt{2\log 1/\alpha}\right)$, $\Gamma_2(\alpha, N) = \frac{2R^2}{\sqrt{N}}\left(2 + \sqrt{2\log 2/\alpha}\right)$*, and* $\Theta(R)$ *denotes the set of Borel probability measures supported on the ball of radius* R*.*

The key idea of their proof is to use a general purpose concentration inequality (McDiarmid's inequality) to compute $\Gamma_1(\alpha, N)$, $\Gamma_2(\alpha, N)$.

We observe that $\mathcal{P}^{CS}$ is the $1-\alpha$ confidence region of a hypothesis test for the mean and covariance of $\mathbb{P}^*$. Namely, the test considers

$$H_0 : \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}_0 \text{ and } \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}\tilde{\mathbf{u}}^T] - \mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}]\mathbb{E}^{\mathbb{P}^*}[\tilde{\mathbf{u}}^T] = \boldsymbol{\Sigma}_0, \tag{5.35}$$

using statistics $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|$ and $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_0\|$ and thresholds $\Gamma_1(\alpha/2, N), \Gamma_2(\alpha/2, N)$.

Practical experience in applied statistics suggests, however, that tests whose thresholds are computed as above using general purpose concentration inequalities, while valid, are typically very conservative for reasonable values of α, N. They reject H_0 when it is false only when N is very large. The standard remedy is to use the bootstrap (Algorithm 1) to calculate alternate thresholds Γ_1^B, Γ_2^B. These bootstrapped thresholds are typically much smaller, but still (approximately) valid at level $1-\alpha$. The first five columns of Table 5.2 illustrates the magnitude of the difference with a particular example. Entries of ∞ indicate that the threshold as derived in Shawe-Taylor and Cristianini (2003) does not apply for this value of N. The data are drawn from a standard normal distribution with $d = 2$ truncated to live in a ball of radius 9.2. We take $\alpha = 10\%$, $N_B = 10,000$. We can see that the reduction can be a full-order of magnitude, or more.

Reducing the thresholds Γ_1^B, Γ_2^B shrinks $\mathcal{P}^{CS}$, in turn reducing the ambiguity in $\mathbb{P}^*$. This reduction ameliorates the potential over-conservativeness of any method using $\mathcal{P}^{CS}$, including the original machine learning application of Shawe-Taylor and Cristianini (2003) and our own schema for developing uncertainty sets.

We next use $\mathcal{P}^{CS}$ in Step 1 of our schema to construct an uncertainty set. Bounding Value at Risk for regions like $\mathcal{P}^{CS}$ was studied by Calafiore and El Ghaoui (2006). Their results imply

$$\sup_{\mathbb{P} \in \mathcal{P}^{CS}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) = \hat{\boldsymbol{\mu}}^T\mathbf{v} + \Gamma_1\|\mathbf{v}\|_2 + \sqrt{\frac{1-\epsilon}{\epsilon}}\sqrt{\mathbf{v}^T(\hat{\boldsymbol{\Sigma}} + \Gamma_2\mathbf{I})\mathbf{v}}. \tag{5.36}$$

We translate this bound into an uncertainty set.

Theorem 5.26. *With probability at least* $1-\alpha$ *with respect to the sampling, the family*

Table 5.2: Comparing Thresholds With and Without Bootstrap

	Shawe-Taylor & Cristianini (2003)				Delage & Ye (2010)			
N	Γ_1	Γ_2	Γ_1^B	Γ_2^B	γ_1	γ_2	γ_1^B	γ_2^B
10	∞	∞	0.805	1.161	∞	∞	0.526	5.372
50	∞	∞	0.382	0.585	∞	∞	0.118	1.684
100	3.814	75.291	0.262	0.427	∞	∞	0.061	1.452
500	1.706	33.671	0.105	0.157	∞	∞	0.012	1.154
50000	0.171	3.367	0.011	0.018	∞	∞	1e-4	1.015
100000	0.121	2.381	0.008	0.013	0.083	5.044	6e-5	1.010

Note: We use $N_B = 10,000$ replications and $\alpha = 10\%$.

$\{\mathcal{U}_\epsilon^{CS} : 0 < \epsilon < 1\}$ *simultaneously implies a probabilistic guarantee for* $\mathbb{P}^*$*, where*

$$\mathcal{U}_\epsilon^{CS} = \left\{ \hat{\boldsymbol{\mu}} + \mathbf{y} + \mathbf{C}^T\mathbf{w} : \exists \mathbf{y}, \mathbf{w} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{y}\| \leq \Gamma_1^B, \quad \|\mathbf{w}\| \leq \sqrt{\frac{1}{\epsilon} - 1} \right\}, \tag{5.37}$$

where $\mathbf{C}^T\mathbf{C} = \hat{\boldsymbol{\Sigma}} + \Gamma_2^B\mathbf{I}$ *is a cholesky decomposition. Moreover,* $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS})$ *is given explicitly by the right-hand side of Eq.* (5.36) *with* (Γ_1, Γ_2) *replaced by the bootstrapped thresholds* Γ_1^B, Γ_2^B*.*

Remark 5.27. Notice that (5.36) is written with an *equality*. The robust constraint $\max_{\mathbf{u}\in\mathcal{U}_\epsilon^{CS}} \mathbf{v}^T\mathbf{x} \leq 0$ is exactly equivalent to the ambiguous chance-constraint $\operatorname{supp}_{\mathbb{P}\in\mathcal{P}^{CS}} \operatorname{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \leq 0$ where $\mathcal{P}^{CS}$ is defined with the smaller (bootstrapped) thresholds.

Remark 5.28. From (5.36), $\{(\mathbf{v}, t) : \delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) \leq t\}$ is second order cone representable. Moreover, we can separate over this constraint in closed-form. Given $\mathbf{v}, t$ such that $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) > t$, let $\mathbf{u} = \boldsymbol{\mu} + \frac{\Gamma_1^B}{\|\mathbf{v}\|}\mathbf{v} + \sqrt{\frac{1}{\epsilon} - 1}\frac{\mathbf{C}\mathbf{v}}{\|\mathbf{C}\mathbf{v}\|}$. Then $\mathbf{u} \in \mathcal{U}_\epsilon^{CS}$ and $\mathbf{u}^T\mathbf{v} \leq t$ is a violated inequality (cf. Proof of Theorem 5.26.)

Remark 5.29. Like $\mathcal{U}_\epsilon^{FB}$, there is no guarantee that $\mathcal{U}_\epsilon^{CS} \subseteq \operatorname{supp}(\mathbb{P}^*)$. Consequently, when a priori knowledge of the support is available, we can refine this set as in Remark 5.17.

To emphasize the benefits of bootstrapping when constructing uncertainty sets, Fig. D-1 in the electronic companion illustrates the set $\mathcal{U}_\epsilon^{CS}$ for the example considered in Fig. 5-3 with thresholds computed with and without the bootstrap.

5.8.2 Uncertainty Set Motivated by Delage and Ye, 2010

Delage and Ye (2010) propose a data-driven approach for solving distributionally robust optimization problems. Their method relies on a slightly more general version

of the following:[3]

Theorem 5.30 (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, $\gamma_1 \equiv \frac{\beta_2}{1-\beta_1-\beta_2}$, $\gamma_2 \equiv \frac{1+\beta_2}{1-\beta_1-\beta_2}$, $\beta_2 \equiv \frac{R^2}{N}\left(2+\sqrt{2\log(\frac{2}{\alpha})}\right)^2$, $\beta_1 \equiv \frac{R^2}{\sqrt{N}}\left(\sqrt{1-\frac{d}{R^4}}+\sqrt{\log(\frac{4}{\alpha})}\right)$, and suppose also that N is large enough so that $1-\beta_1-\beta_2 > 0$. Finally suppose $supp(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Then with probability at least $1-\alpha$ with respect to the sampling, $\mathbb{P}^* \in \mathcal{P}^{DY}$ where*

$$\mathcal{P}^{DY} \equiv \left\{ \mathbb{P} \in \Theta[\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] : \begin{array}{l} (\mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}]-\hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbb{E}^{\mathbb{P}}[\tilde{\mathbf{u}}]-\hat{\boldsymbol{\mu}}) \leq \gamma_1, \\ \mathbb{E}^{\mathbb{P}}[(\tilde{\mathbf{u}}-\hat{\boldsymbol{\mu}})(\tilde{\mathbf{u}}-\hat{\boldsymbol{\mu}})^T] \preceq \gamma_2\hat{\boldsymbol{\Sigma}} \end{array} \right\}.$$

The key idea is again to compute the thresholds using a general purpose concentration inequality. The condition on N is required for the confidence region to be well-defined.

We again observe that $\mathcal{P}^{DY}$ is the $1-\alpha$ confidence region of a hypothesis test. Specifically, it considers the hypothesis (5.35) using the statistics $(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_0)$ and $\max_{\boldsymbol{\lambda}} \frac{\boldsymbol{\lambda}^T(\boldsymbol{\Sigma}_0+(\boldsymbol{\mu}_0-\hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_0-\hat{\boldsymbol{\mu}})^T)\boldsymbol{\lambda}}{\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}}$ with thresholds γ_1, γ_2.

Since the thresholds are, again, potentially overly conservative, we approximate new thresholds using the bootstrap. Table 5.2 shows the reduction in magnitude. Observe that the bootstrap thresholds exist for all N, not just N sufficiently large. Moreover, they are significantly smaller. This reduction translates to a reduction in the potential over conservatism of any method using $\mathcal{P}^{DY}$, including those presented within Delage and Ye (2010) while retaining the same probabilistic guarantee.

We next consider using $\mathcal{P}^{DY}$ in Step 1 of our schema to generate an uncertainty set $\mathcal{U}$ that "corresponds" to this method.

Theorem 5.31. *Suppose $supp(\mathbb{P}^*) \subset [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$. Then, with probability at least $1-\alpha$ with respect to the sampling, the family $\{\mathcal{U}_\epsilon^{DY} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee for $\mathbb{P}^*$, where*

$$\begin{aligned} \mathcal{U}_\epsilon^{DY} = \Big\{ \mathbf{u} \in [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] : & \exists \lambda \in \mathbb{R},\ \mathbf{w}, \mathbf{m} \in \mathbb{R}^d,\ \mathbf{A}, \hat{\mathbf{A}} \succeq \mathbf{0} \text{ s.t.} \\ & \lambda \leq \frac{1}{\epsilon}, \quad (\lambda-1)\hat{\mathbf{u}}^{(0)} \leq \mathbf{m} \leq (\lambda-1)\hat{\mathbf{u}}^{(N+1)}, \\ & \begin{pmatrix} \lambda-1 & \mathbf{m}^T \\ \mathbf{m} & \mathbf{A} \end{pmatrix} \succeq \mathbf{0}, \quad \begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \hat{\mathbf{A}} \end{pmatrix} \succeq \mathbf{0}, \\ & \lambda\hat{\boldsymbol{\mu}} = \mathbf{m}+\mathbf{u}+\mathbf{w}, \quad \|\mathbf{C}\mathbf{w}\| \leq \lambda\sqrt{\gamma_1^B}, \\ & \lambda(\gamma_2^B\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) - \mathbf{A} - \hat{\mathbf{A}} - \mathbf{w}\hat{\boldsymbol{\mu}}^T - \hat{\boldsymbol{\mu}}\mathbf{w}^T \succeq \mathbf{0} \Big\}, \end{aligned} \tag{5.38}$$

$C^TC = \hat{\boldsymbol{\Sigma}}^{-1}$ is a Cholesky-decomposition, and γ_1^B, γ_2^B are computed by bootstrap.

[3]Specifically, since R is typically unknown, the authors describe an estimation procedure for R and prove a modified version of the Theorem 5.30 using this estimate and different constants. We treat the simpler case where R is known here. Extensions to the other case are straightforward.

Moreover,

$$\begin{aligned}
\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^{DY}) &= \sup_{\mathbb{P}\in\mathcal{P}^{DY}} VaR_\epsilon^{\mathbb{P}}(\mathbf{v}) \\
&= \inf \quad t \\
&\quad s.t. \quad r+s \le \theta\epsilon, \\
&\qquad \begin{pmatrix} r+\mathbf{y}_1^{+T}\hat{\mathbf{u}}^{(0)}-\mathbf{y}_1^{-T}\hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}(\mathbf{q}-\mathbf{y}_1)^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_1) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
&\qquad \begin{pmatrix} r+\mathbf{y}_2^{+T}\hat{\mathbf{u}}^{(0)}-\mathbf{y}_2^{-T}\hat{\mathbf{u}}^{(N+1)}+t-\theta & \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\mathbf{v})^T, \\ \frac{1}{2}(\mathbf{q}-\mathbf{y}_2-\mathbf{v}) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
&\qquad s \ge (\gamma_2^B\hat{\mathbf{\Sigma}}+\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T)\circ\mathbf{Z}+\hat{\boldsymbol{\mu}}^T\mathbf{q}+\sqrt{\gamma_1^B}\|\mathbf{q}+2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
&\qquad \mathbf{y}_1=\mathbf{y}_1^+-\mathbf{y}_1^-, \quad \mathbf{y}_2=\mathbf{y}_2^+-\mathbf{y}_2^-, \quad \mathbf{y}_1^+,\mathbf{y}_1^-,\mathbf{y}_2^+,\mathbf{y}_2^-,\theta\ge\mathbf{0}.
\end{aligned}$$

Remark 5.32. Similar to $\mathcal{U}_\epsilon^{CS}$, the robust constraint $\max_{\mathbf{u}\in\mathcal{U}_\epsilon^{DY}} \mathbf{v}^T\mathbf{u} \le 0$ is equivalent to the ambiguous chance constraint $\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \le 0$.

Remark 5.33. The set $\{(\mathbf{v},t) : \delta^*(\mathbf{v}|\,\mathcal{U}^{DY}) \le t\}$ is representable as a linear matrix inequality. At time of writing, solvers for linear matrix inequalities are not as developed as those for second order cone programs. Consequently, one may prefer $\mathcal{U}_\epsilon^{CS}$ to $\mathcal{U}_\epsilon^{DY}$ in practice for its simplicity.

5.8.3 Comparing $\mathcal{U}_\epsilon^M$, $\mathcal{U}_\epsilon^{LCX}$, $\mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$

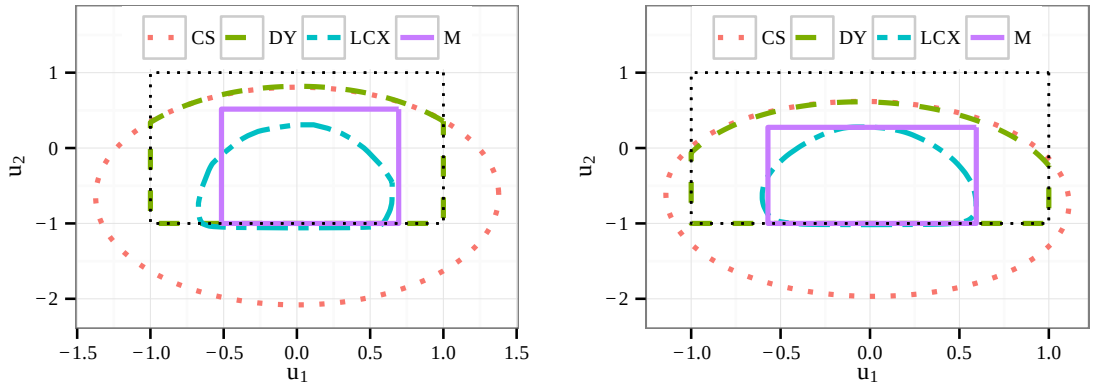
One of the benefits of deriving uncertainty sets corresponding to the methods of Shawe-Taylor and Cristianini (2003) and Delage and Ye (2010) is that it facilitates comparisons between these methods and our own proposals. In Fig. 5-4, we illustrate the sets $\mathcal{U}_\epsilon^M$, $\mathcal{U}_\epsilon^{LCX}$, $\mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$ for the same numerical example from Fig. 5-3. Because $\mathcal{U}^M$ does not leverage the joint distribution $\mathbb{P}^*$, it does not learn that its marginals are independent. Consequently, $\mathcal{U}^M$ has pointed corners permitting extreme values of both coordinates simultaneously. The remaining sets do learn the marginal independence from the data and, hence, have rounded corners.

The set $\mathcal{U}_\epsilon^{CS}$ is not contained in $\mathrm{supp}(\mathbb{P}^*)$. Interestingly, the intersection $\mathcal{U}_\epsilon^{CS} \cap \mathrm{supp}(\mathbb{P}^*)$ is very similar to $\mathcal{U}_\epsilon^{DY}$ for this example (indistinguishable in picture). Since $\mathcal{U}^{CS}$ and $\mathcal{U}^{DY}$ only depend on the first two moments of $\mathbb{P}^*$, neither is able to capture the skewness in the second coordinate. Finally, $\mathcal{U}^{LCX}$ is contained within $\mathrm{supp}(\mathbb{P}^*)$ and displays symmetry in the first coordinate and skewness in the second. In this example it is also the smallest set (in terms of volume). All sets shrink as N increases.

5.8.4 Refining $\mathcal{U}_\epsilon^{FB}$

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. 5.2.2, for which there is a great deal of experimental evidence to suggest that the test is still approximately

Figure 5-4: Comparing $\mathcal{U}_\epsilon^M$, $\mathcal{U}_\epsilon^{LCX}$, $\mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$



Note: The true distributions are as in Fig. 5-3 and $\epsilon = 10\%$, $\alpha = 20\%$. The left panel uses $N = 100$ data points, while the right panel uses $N = 1,000$ data points.

valid when the underlying data is non-Gaussian (Lehmann and Romano 2010, 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 (Lehmann and Romano 2010, 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 $\mathcal{U}_\epsilon^{FB}$. We replace m_{fi}, m_{bi} in Eq. (5.25), with the upper and lower thresholds of a t-test at level $\alpha'/2$. We expect these new thresholds to correctly bound the true mean μ_i with probability approximately $1 - \alpha'/2$ with respect to the data. We then use the bootstrap to calculate bounds on the forward and backward deviations $\overline{\sigma}_{fi}, \overline{\sigma}_{bi}$.

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.

5.9 Optimizing over Multiple Constraints

In this section, we propose an approach for solving (5.9). The key observation is

Theorem 5.34.

a) The constraint $\delta^*(\mathbf{v} | \, \mathcal{U}_\epsilon^{CS}) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* ϵ*, for* $0 < \epsilon < .75$.

b) The constraint $\delta^*(\mathbf{v} | \, \mathcal{U}_\epsilon^{FB}) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* ϵ*, for* $0 < \epsilon < 1/\sqrt{e}$.

c) The constraint $\delta^*(\mathbf{v} | \, \mathcal{U}_\epsilon) \leq t$ *is bi-convex in* $(\mathbf{v}, t)$ *and* ϵ*, for* $0 < \epsilon < 1$*, and* $\mathcal{U}_\epsilon \in \{\mathcal{U}_\epsilon^{\chi^2}, \mathcal{U}_\epsilon^{G}, \mathcal{U}_\epsilon^{I}, \mathcal{U}_\epsilon^{LCX}, \mathcal{U}_\epsilon^{DY}\}$.

This observations suggests a heuristic: Fix the values of ϵ_j, and solve the robust optimization problem in the original decision variables. Then fix this solution and optimize over the ϵ_j. Repeat until some stopping criteria is met or no further improvement occurs. Chen et al. (2010) suggested a similar heuristic for multiple chance-constraints in a different context. In Appendix D.3 we propose a refinement of this approach that solves a linear optimization problem to obtain the next iterates for ϵ_j, incorporating dual information from the overall optimization and other constraints. Our proposal ensures the optimization value is non-increasing between iterations and that the procedure is finitely convergent.

5.10 Choosing the "Right" Set and Tuning α, ϵ

Often several of our data-driven sets may be consistent with the a priori knowledge of $\mathbb{P}^*$. Choosing an appropriate set from amongst our proposals is a non-trivial task that depends on the application and the data. One may be tempted to use the intersection of all eligible sets. We caution that the intersection of two sets which imply a probabilistic guarantee at level ϵ need not imply a probabilistic guarantee at level ϵ. Similarly, one may be tempted to solve the robust optimization model for each eligible set separately and select the set and solution with best objective value. We caution that a set chosen in this way will suffer from an in-sample bias. Specifically, the probability with respect to the sampling that this set does not imply a probabilistic guarantee at level ϵ may be much larger than α.

Drawing an analogy to model selection in machine learning, we propose a different approach to set selection. Specifically, split the data into two parts, a training set and a hold-out set. Use the training set to construct each potential uncertainty set, in turn, and solve the robust optimization problem. Test each of the corresponding solutions out-of-sample on the hold-out set, and select the best solution and corresponding uncertainty set. Since the two halves of the data are independent, it follows that with probability at least $1 - \alpha$ with respect to the sampling, the set so selected will correctly imply a probabilistic guarantee at level ϵ.

The drawback of this approach is that only half the data is used to calibrate the uncertainty set. When N is only moderately large, this may be impractical. In these cases, k-fold cross-validation can be used to select a set. (See Hastie et al. (2001) for a review of cross-validation.) Unlike the above procedure, we cannot prove that the set chosen by k-fold cross-validation satisfies the appropriate guarantee. Nevertheless, experience in model selection suggests that this procedure frequently identifies a good model, and, thus, we expect it will identify a good set. We use 5-fold cross-validation in our numerical experiments.

In applications where there is not a natural choice for α or ϵ, we suggest tuning these parameters in an entirely analogous way. Namely, we propose selecting a grid of potential values for α and/or ϵ and then selecting the best value either using a hold-out set or cross-validation. Since the optimal value likely depends on the choice of uncertainty set, we suggest choosing them jointly.

5.11 Applications

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

- In these two applications, our data-driven sets outperform traditional, non-data driven uncertainty sets, and, moreover, robust models built with our sets perform as well or better than other data-driven approaches.
- Although our data-driven sets all shrink as $N \to \infty$, they learn different features of $\mathbb{P}^*$, such as correlation structure and skewness. Consequently, different sets may be better suited to different applications, and the right choice of set may depend on N. Cross-validation and other model selection techniques effectively identify the best set.
- Optimizing the ϵ_j's in the case of multiple constraints can significantly improve performance.

5.11.1 Portfolio Management

Portfolio management has been well-studied in the robust optimization literature (e.g., Goldfarb and Iyengar 2003, Natarajan et al. 2008, Calafiore and Monastero 2012). For simplicity, we will consider the one period allocation problem:

$$\max_{\mathbf{x}} \left\{ \min_{\mathbf{r} \in \mathcal{U}} \quad \mathbf{r}^T \mathbf{x} : \quad \mathbf{e}^T \mathbf{x} = 1, \quad \mathbf{x} \geq \mathbf{0} \right\}, \tag{5.39}$$

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 model from Natarajan et al. (2008):

$$\tilde{r}_i = \begin{cases} \frac{\sqrt{(1-\beta_i)\beta_i}}{\beta_i} & \text{with probability } \beta_i \\ -\frac{\sqrt{(1-\beta_i)\beta_i}}{1-\beta_i} & \text{with probability } 1-\beta_i \end{cases}, \quad \beta_i = \frac{1}{2}\left(1 + \frac{i}{11}\right), \quad i = 1, \ldots, 10. \tag{5.40}$$

In this model, all assets have the same mean return (0%), the same standard deviation (1.00%), but have different skew and support. Higher indexed assets are highly skewed; they have a small probability of achieving a very negative return. Returns for different assets are independent. We simulate $N = 500$ returns to use as data.

We will utilize our sets $\mathcal{U}_\epsilon^M$ and $\mathcal{U}_\epsilon^{LCX}$ in this application. We do not consider the sets $\mathcal{U}_\epsilon^I$ or $\mathcal{U}_\epsilon^{FB}$ since we do not know a priori that the returns are independent. To contrast to the methods of (Shawe-Taylor and Cristianini 2003) and (Delage and Ye 2010) we also construct the sets $\mathcal{U}_\epsilon^{CS}$ and $\mathcal{U}_\epsilon^{DY}$. Recall from Remarks 5.27 and 5.32

Table 5.3: Portfolio Statistics for Each of Our Methods

	$N = 500$				$N = 2000$			
	z_{In}	CV	z_{Out}	z_{Avg}	z_{In}	CV	z_{Out}	z_{Avg}
M	-1.095	-1.095	-1.095	-1.095	-1.095	-1.095	-1.095	-1.095
LCX	-0.699	-0.373	-0.373	-0.411	-0.89	-0.428	-0.395	-0.411
CS	-1.125	-0.403	-0.416	-0.397	-1.306	-0.400	-0.417	-0.396
CM	-0.653	-0.495	-0.425	-0.539	-0.739	-0.426	-0.549	-0.451

Note: $\mathcal{U}_\epsilon^{DY}$ and $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ perform identically to $\mathcal{U}_\epsilon^{M}$. "CM" refers to the method of Calafiore and Monastero (2012).

that robust linear constraints over these sets are equivalent to ambiguous chance-constraints in the original methods, but with improved thresholds. As discussed in Remark 5.29, we also construct $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ for comparison. We use $\alpha = \epsilon = 10\%$ in all of our sets. Finally, we will also compare to the method of Calafiore and Monastero (2012) (denoted "CM" in our plots), which is not an uncertainty set based method. We calibrate this method to also provide a bound on the 10% worst-case return that holds with at least 90% with respect to the sampling so as to provide a fair comparison.

We first consider the problem of selecting an appropriate set via 5-fold cross-validation. The top left panel in Fig. 5-5 shows the out-of-sample 10% worst-case return for each of the 5 runs (blue dots), as well as the average performance on the 5 runs for each set (black square). Sets $\mathcal{U}_\epsilon^{M}$, $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ and $\mathcal{U}_\epsilon^{DY}$ yield identical portfolios (investing everything in the first asset) so we only include $\mathcal{U}^M$ in our graphs. The average performance is also shown in Table 5.3 under column CV (for "cross-validation.") The optimal objective value of (5.39) for each of our sets (trained with the entire data set) is shown in column z_{In}.

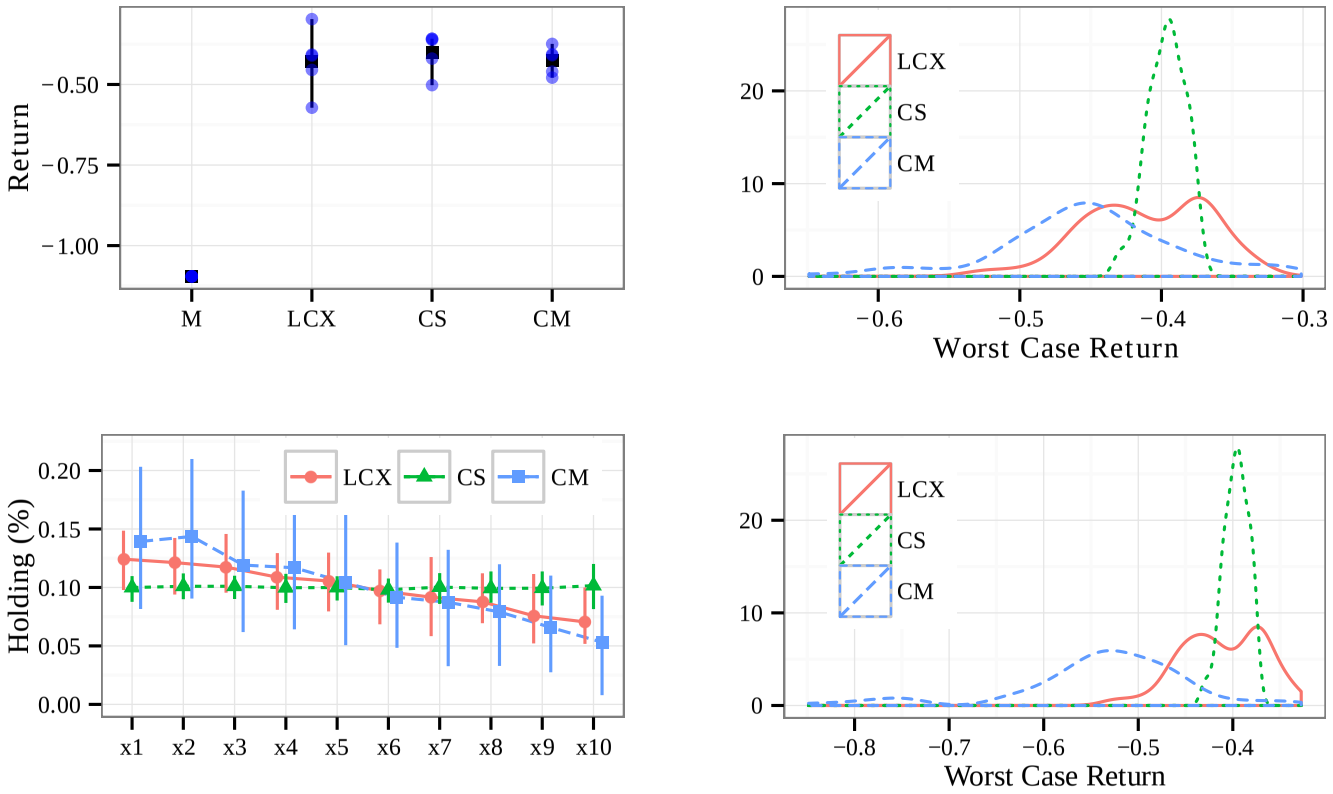
Based on the top left panel of Fig. 5-5, it is clear that $\mathcal{U}_\epsilon^{LCX}$ and $\mathcal{U}_\epsilon^{CS}$ significantly outperform the remaining sets. They seem to perform similarly to the CM method. Consequently, we would choose one of these two sets in practice.

We can assess the quality of this choice by using the ground-truth model (5.40) to calculate the true 10% worst-case return for each of the portfolios. These are shown in Table 5.3 under column z_{Out}. Indeed, these sets perform better than the alternatives, and, as expected, the cross-validation estimates are reasonably close to the true out-of-sample performance. By contrast, the in-sample objective value z_{In} is a loose bound. We caution against using this in-sample value to select the best set.

Interestingly, we point out that while $\mathcal{U}_\epsilon^{CS} \cap \text{supp}(\mathbb{P}^*)$ is potentially smaller (with respect to subset containment) than $\mathcal{U}_\epsilon^{CS}$, it performs much worse out-of-sample (it performs identically to $\mathcal{U}_\epsilon^{M}$). This experiment highlights the fact that size calculations alone cannot predict performance. Cross-validation or similar techniques are required.

One might ask if these results are specific to the particular draw of 500 data points we use. We repeat the above procedure 100 times. The resulting distribution of 10%

Figure 5-5: Portfolio Performance by Method



Note: $\alpha = \epsilon = 10\%$. Top left: Cross-validation results. Top right: Out-of-sample distribution of the 10% worst-case return over 100 runs. Bottom left: Average portfolio holdings by method. Bottom right: Out-of-sample distribution of the 10% worst-case return over 100 runs. The bottom right panel uses $N = 2000$. The remainder use $N = 500$.

worst-case return is shown in the top right panel of Fig. 5-5 and the average of these runs is shown Table 5.3 under column z_{Avg}. As might have been guessed from the cross-validation results, $\mathcal{U}_\epsilon^{CS}$ delivers more stable and better performance than either $\mathcal{U}_\epsilon^{LCX}$ or CM. $\mathcal{U}_\epsilon^{LCX}$ slightly outperforms CM, and its distribution is shifted right.

We next look at the distribution of actual holdings between these methods. We show the average holding across these 100 runs as well as 10% and 90% quantiles for each asset in the bottom left panel of Fig. 5-5. Since $\mathcal{U}_\epsilon^{M}$ does not use the joint distribution, it sees no benefit to diversification. Portfolios built from $\mathcal{U}_\epsilon^{M}$ consistently holds all their wealth in the first asset over all the runs, hence, omitted from graphs. The set $\mathcal{U}_\epsilon^{CS}$ depends only on the first two moments of the data, and, consequently, cannot distinguish between the assets. It holds a very stable portfolio of approximately the same amount in each asset. By contrast, $\mathcal{U}^{LCX}$ is able to learn the asymmetry in the distributions, and holds slightly less of the higher indexed (toxic) assets. CM is similar to $\mathcal{U}^{LCX}$, but demonstrates more variability in the holdings.

We point out that the performance of each method depends slightly on N. We repeat the above experiments with $N = 2000$. Results are summarized in Table 5.3. The bottom right panel of Fig. 5-5 shows the distribution of the 10% worst-case return. (Additional plots are also available in Appendix D.4.) Both $\mathcal{U}^{LCX}$ and CM perform noticeably better with the extra data, but $\mathcal{U}^{LCX}$ now noticeably outperforms CM and its distribution is shifted significantly to the right.

5.11.2 Queueing Analysis

One of the strengths of our approach is the ability to retrofit existing robust optimization models by replacing their uncertainty sets with our proposed sets, thereby creating new data-driven models that satisfy strong guarantees. In this section, we illustrate this idea with a robust queueing model as in Bertsimas et al. (2011a) and Bandi et al. (2012). Bandi et al. (2012) use robust optimization to generate *approximations* to a performance metric of a queuing network. We will combine their method with our new sets to generate *probabilistic upper bounds* to these metrics. For concreteness, we focus on the waiting time in a G/G/1 queue. Extending our analysis to more complex queueing networks can likely be accomplished similarly. We stress that we do not claim that our new bounds are the best possible – indeed there exist extremely accurate, specialized techniques for the G/G/1 queue – but, rather, that the retrofitting procedure is general purpose and yields reasonably good results. These features suggest that a host of other robust optimization applications in information theory (Bandi and Bertsimas 2012), supply-chain management (Ben-Tal et al. 2005) and revenue management (Rusmevichientong and Topaloglu 2012) might benefit from this retrofitting.

Let $\tilde{\mathbf{u}}_i = (\tilde{x}_i, \tilde{t}_i)$ for $i = 1, \ldots, n$ denote the uncertain service times and interarrival times of the first n customers in a queue. We assume that $\tilde{\mathbf{u}}_i$ is i.i.d. for all i and has independent components, and that there exists $\hat{\mathbf{u}}^{(N+1)} \equiv (\overline{x}, \overline{t})$ such that $0 \leq \tilde{x}_i \leq \overline{x}$ and $0 \leq \tilde{t}_i \leq \overline{t}$ almost surely.

From Lindley's recursion (Lindley 1952), the waiting time of the n^{th} customer is

$$\tilde{W}_n = \max_{1\le j\le n}\left(\max\left(\sum_{l=j}^{n-1}\tilde{x}_l - \sum_{l=j+1}^{n}\tilde{t}_l, 0\right)\right) = \max\left(0, \max_{1\le j\le n}\left(\sum_{l=j}^{n-1}\tilde{x}_l - \sum_{l=j+1}^{n}\tilde{t}_l\right)\right). \tag{5.41}$$

Motivated by Bandi et al. (2012), we consider a worst-case realization of a Lindley recursion

$$\max\left(0, \max_{1\le j\le n}\max_{(\mathbf{x},\mathbf{t})\in\mathcal{U}}\left(\sum_{l=j}^{n-1}\tilde{x}_l - \sum_{l=j+1}^{n}\tilde{t}_l\right)\right). \tag{5.42}$$

Taking $\mathcal{U} = \mathcal{U}^{FB}_{\bar{\epsilon}/n}$ and applying Theorem 5.15 to the inner-most optimization yields

$$\max_{1\le j\le n}(m_{f1} - m_{b2})(n-j) + \sqrt{2\log(n/\bar{\epsilon})(\sigma_{f1}^2 + \sigma_{b2}^2)}\sqrt{n-j} \tag{5.43}$$

Relaxing the integrality on j, this optimization can be solved closed-form yielding

$$W_n^{1,FB} \equiv \begin{cases} (m_{f1} - m_{b2})n + \sqrt{2\log(\frac{n}{\bar{\epsilon}})(\sigma_{f1}^2 + \sigma_{b2}^2)}\sqrt{n} & \text{if } n < n_c^{1,FB} \text{ or } m_{f1} > m_{b2}, \\ \frac{\log(\frac{n}{\bar{\epsilon}})(\sigma_{f1}^2+\sigma_{b2}^2)}{2(m_{b2}-m_{f1})} & \text{otherwise,} \end{cases} \tag{5.44}$$

where $n_c^{1,FB} = \frac{\log(\frac{n}{\bar{\epsilon}})(\sigma_{f1}^2+\sigma_{b2}^2)}{2(m_{b2}-m_{f1})^2}$. From (5.42), with probability at least $1-\alpha$ with respect to the sampling, each of the inner-most optimizations upper bound their corresponding random quantity with probability $1-\bar{\epsilon}/n$ with respect to $\mathbb{P}^*$. Thus, by union bound, $\mathbb{P}^*(\tilde{W}_n \le W_n^{1,FB}) \ge 1-\bar{\epsilon}$.

On the other hand, since $\{\mathcal{U}_\epsilon^{FB} : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee, we can also optimize the choice of ϵ_j in (5.43), yielding

$$\begin{aligned} W_n^{2,FB} \equiv \min_{w,\boldsymbol{\epsilon}} \quad & w \\ \text{s.t.} \quad & w \ge (m_{f1} - m_{b2})(n-j) + \sqrt{2\log(1/\epsilon_j)(\sigma_{f1}^2 + \sigma_{b2}^2)}\sqrt{n-j}, \quad j = 1,\dots,n-1, \\ & w \ge 0, \quad \boldsymbol{\epsilon} \ge \mathbf{0}, \quad \sum_{j=1}^{n-1}\epsilon_j \le \bar{\epsilon}. \end{aligned} \tag{5.45}$$

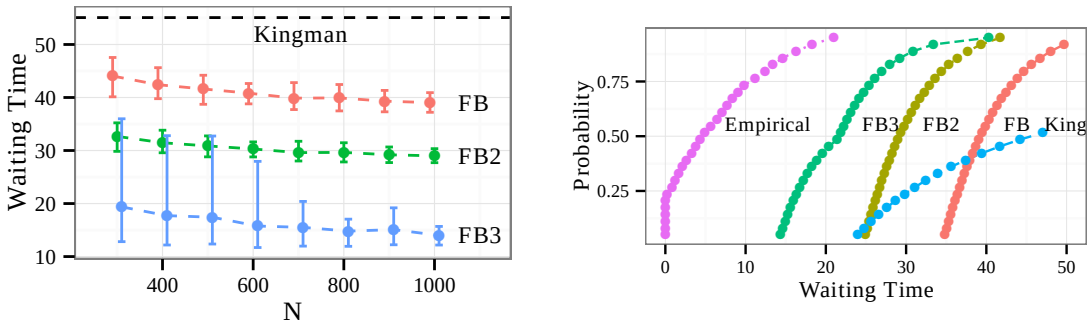
From the KKT conditions, the constraint (5.45) will be tight for all j, so that $W_n^{2,FB}$ satisfies

$$\sum_{j=1}^{n-1}\exp\left(-\frac{(W_n^{2,FB} - (m_{f1} - m_{b2}))^2}{2(n-j)(\sigma_{f1}^2 + \sigma_{b2}^2)^2}\right) = \bar{\epsilon}, \tag{5.46}$$

which can be solved by line search. Again, with probability $1-\alpha$ with respect to the sampling, $\mathbb{P}^*(\tilde{W}_n \le W_n^{2,FB}) \ge 1-\bar{\epsilon}$, and $W_n^{2,FB} \le W_n^{1,FB}$ by construction.

We can further refine our bound by truncating the recursion (5.41) at customer

Figure 5-6: Results for the Queueing Analysis Example



Note: The left panel shows various bounds on the median waiting time ($\epsilon = .5$) for $n = 10$ and various values of N. The right panel bounds the entire cumulative distribution of the waiting time for $n = 10$ and $N = 1000$. using $W_n^{FB,3}$. In both cases, $\alpha = 20\%$.

$\min(n, n^{(k)})$ where, with high probability, $\tilde{n} \leq n^{(k)}$. A formal derivation of the resulting bound, which we denote $W_n^{3,FB}$, can be found in Appendix D.5. Therein we also prove that with probability at least $1 - \alpha$ with respect to the sampling, $\mathbb{P}^*(\tilde{W}_n \leq W_n^{3,FB}) \geq 1 - \bar{\epsilon}$.

Finally, our choice of $\mathcal{U}_\epsilon^{FB}$ was somewhat arbitrary. Similar analysis can be performed for many of our sets. To illustrate, Appendix D.5 also contains corresponding bounds for the set $\mathcal{U}_\epsilon^{CS}$.

We illustrate these ideas numerically. Let service times follow a Pareto distribution with parameter 1.1 truncated at 15, and the interarrival times follow an exponential distribution with rate 3.05 truncated at 15.25. The resulting truncated distributions have means of approximately 3.029 and 3.372, respectively, yielding an approximate 90% utilization.

As a first experiment, we bound the median waiting time ($\epsilon = 50\%$) for the $n = 10$ customer, using each of our bounds with differing amounts of data. We repeat this procedure 100 times to study the variability of our bounds with respect to the data. The left panel of Fig. 5-6 shows the average value of the bound and error bars for the 10% and 90% quantiles. As can be seen, all of the bounds improve as we add more data. Moreover, optimizing the ϵ_j's (the difference between $W_n^{FB,1}$ and $W_n^{FB,2}$ is significant.

For comparison purposes, we include a sample analogue of Kingman's bound (Kingman 1962) on the $1 - \epsilon$ quantile of the waiting time, namely,

$$W^{King} \equiv \frac{\hat{\mu}_x(\hat{\sigma}_a^2\hat{\mu}_x^2 + \hat{\sigma}_x^2\hat{\mu}_t^2)}{2\bar{\epsilon}\hat{\mu}_t^2(\hat{\mu}_t - \hat{\mu}_x)},$$

where $\hat{\mu}_t, \hat{\sigma}_t^2$ are the sample mean and sample variance of the arrivals, $\hat{\mu}_x, \hat{\sigma}_x^2$ are the sample mean and sample variance of the service times, and we have applied Markov's

Table 5.4: Summary Statistics for Various Bounds on Median Waiting Time

	Mean	St. Dev	10%	90%
$W_n^{FB,1}$	34.6	0.4	34.0	35.2
$W_n^{FB,2}$	25.8	0.3	25.4	26.2
$W_n^{FB,3}$	14.4	1.2	13.5	15.5
W^{King}	55.1	8.7	46.0	67.4

Note: $N = 10,000$, $n = 10$, $\alpha = 10\%$. The last two columns refer to upper and lower quantiles over the simulation.

inequality. Unfortunately, this bound is extremely unstable, even for large N. The dotted line in the left-panel of Fig. 5-6 is the average value over the 100 runs of this bound for $N = 10,000$ data points (the error-bars do not fit on graph.) Sample statistics for this bound and our bounds can also be seen in Table 5.4. As shown, our bounds are both significantly better (with less data), and exhibit less variability.

As a second experiment, we use our bounds to calculate a probabilistic upper bound on the entire CDF of $\tilde{W}_n$ for $n = 10$ with $N = 1,000$, $\alpha = 20\%$. Results can be seen in the right panel of Fig. 5-6. We have included the empirical CDF of the waiting time and the sampled version of the Kingman bound comparison. As seen, our bounds significantly improve upon the sampled Kingman bound, and the benefit of optimizing the ϵ_j's is again, significant. We remark that the ability to simultaneously bound the entire CDF for any n, whether transient or steady-state, is an important strength of this type of analysis.

5.12 Conclusions

The prevalence of high quality data is reshaping operations research. Indeed, a new data-centered paradigm is emerging. In this work, 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 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.

Part III

The Interface Between Controlled Experimentation and Modern Optimization

Chapter 6

The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples

Random assignment, typically seen as the standard in controlled trials, aims to make experimental groups statistically equivalent before treatment. However, with a small sample, which is a practical reality in many disciplines, randomized groups are often too dissimilar to be useful. We propose an approach based on discrete linear optimization to create groups whose discrepancy in their means and variances is several orders of magnitude smaller than with randomization. We provide theoretical and computational evidence that groups created by optimization have exponentially lower discrepancy than those created by randomization and that this allows for more powerful statistical inference.

6.1 Introduction

Experimentation on groups of subjects, similar in all ways but for the application of an experimental treatment, is a cornerstone of modern scientific inquiry. In any controlled experiment, the quality, interpretability, and validity of the measurements and inferences drawn depends upon the degree to which the groups are similar at the outset.

For close to a century, randomization of subjects into different groups has been relied upon to generate statistically equivalent groups. Where group size is large relative to variability, randomization robustly generates groups that are well-matched with respect to any statistic. However, when group sizes are small, the expected discrepancy in any covariate under randomization can be surprisingly large, hindering inference. This problem is further aggravated as the number of groups one needs to populate becomes larger.

This is the situation faced in numerous disciplines in which the rarity or expense of subjects makes assembly of large groups impractical. For example, in the field of oncology research, experimental chemotherapy agents are typically tested first in

mouse models of cancer, in which tumor-bearing mice are segregated into groups and dosed with experimental compounds. Since these mouse models are laborious and expensive, group size is kept small (typically 8-10), while the number of groups is relatively large, to accommodate comparison of multiple compounds and doses with standard-of-care compounds and untreated control groups. In this case, it is clear that initial tumor weight is highly correlated with the post-treatment tumor weight, in which we measure the effect of treatment. A typical experiment might consist of 40-60 mice segregated into four to six groups of ten, though experiments using fewer mice per group and many more groups are performed as well. Given that the implanted tumors grow quite heterogeneously, a coefficient of variation of 50% or more in pre-treatment tumor size is not unusual.

In such circumstances, common in nearly all research using animal models of disease as well as many other endeavors, simple randomization fails to reliably generate statistically equivalent groups, and therefore fails to generate reliable inference. It is clearly more desirable that experiments be conducted with groups that are similar, in particular in mean and variance of relevant baseline covariates. Here we treat the composition of small statistically equivalent groups as a mathematical optimization problem in which the goal is to minimize the maximum difference in both mean and variance between any two groups. We report one treatment of this problem as well as a study of the size of the discrepancy when group enrollment is optimized compared to other common designs including complete randomization.

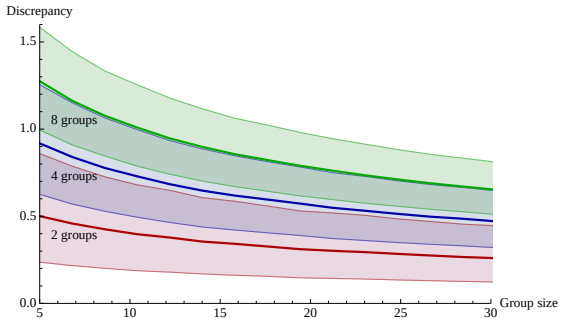
Block and orthogonal designs (see Fisher (1935)) have been a common way to reduce variability when baseline covariates are categorical, but do not apply to mixed (discrete and continuous) covariates, which is the main focus of our work. For such cases, apart from randomization, two prominent methods are pairwise matching for controlled trials (see Rosenbaum and Rubin (1985) and Greevy et al. (2004)) and re-randomization as proposed in Morgan et al. (2012).[1] The finite selection model (FSM) proposed by Morris (1979) can also be used for this purpose. In comparisons explored in Section 6.4, we find that the balance produced by our proposed optimization-based approach greatly improves on both randomization and these methods.

Pairwise matching is most common in observational studies, where assignment to treatment cannot be controlled (see Rubin (1979) and Rosenbaum and Rubin (1983) for a thorough discussion of the application of pairwise matching and other methods to observational studies). A large impediment to the existing practices is that they are based on subject pairs. When sample sizes are small and random there will hardly be any well-matched pairs. We will see that such matching does little to eliminate bias in the statistics that measure the overall average effect size. Instead we consider matching the experimental groups in order to minimize the en-masse discrepancies in means and variances among groups as formulated in (6.1).

When discrepancy is minimized, statistics such as the mean difference in subject responses are far more precise, concentrated tightly around their nominal value, while

[1]The work of Morgan et al. (2012) can be seen as formalizing and reinterpreting the common informal practice of cherry-picking from several randomizations as a principled heuristic method for matching.

Figure 6-1: Average Maximal Pairwise Discrepancy in Means Among Randomly Assigned Groups of Normal Variates



Note: The vertical axis is in units of standard deviation. The band denotes the average over- and under-shoot: $\mathbb{E}[X|X \geq \mathbb{E}X]$ and $\mathbb{E}[X|X \leq \mathbb{E}X]$ where X is maximal pairwise discrepancy.

still being unbiased estimates. Indeed, under optimization, these statistics will no longer follow their usual distributions, which are wider, and traditional tests that rely on knowledge of this distribution, like the Student T test, no longer apply. Beyond estimation, we propose a hypothesis test based on the bootstrap to draw inferences on the differences between treatments – inferences which experimental evidence shows are much more powerful than is usually possible.

In this chapter, we provide theoretical and computational evidence that groups created by optimization have exponentially lower discrepancy in pre-treatment covariates than those created by randomization or by existing matching methods.

6.2 Limitations of Randomization

Three factors can impair successful matching of the independent variable means of groups assembled using randomization. These are: (a) the group size, (b) the variance of the data and (c) the number of groups being populated. The specific influence of these three factors is shown graphically in Figure 6-1. The plot shows the average maximal pair-wise discrepancy in means between groups under the conditions indicated for the normal distribution. Average discrepancy is proportional to standard deviation and is therefore reported in units of standard deviations.

It can be seen from the plot that discrepancy increases with the number of groups involved and decreases with increasing group size. When all three factors come into play: small group size, high standard deviation, and numerous groups, the degree of discrepancy can be substantial. For example, a researcher using randomization to create four groups of ten mice each will be left with an average discrepancy of 0.66 standard deviations between some two of the groups. Since statistical significance is often declared at a mean difference of 1.96 standard deviations ($p \leq 0.05$), this introduces enough noise into the experiment to conceal an effect in comparisons between

the mismatched groups or to severely skew the apparent magnitude and statistical significance of a larger effect. Examination of Figure 6-1 makes it clear that when multiple groups are involved, even apparently large group size can still result in a substantial discrepancy in means between some groups. Doubling the group sizes to twenty each still leaves the researcher with a discrepancy of 0.47 standard deviations.

One solution to this problem is simply to increase group size until discrepancies decrease to acceptable levels. However, the size of the groups needed to do so can be surprisingly large. To reduce the expected discrepancy to below 0.1 standard deviations would require more than 400 subjects per group in the above experiment. For 0.01 standard deviations, more than 40,000 subjects per group would be necessary. With diminishing returns in the reduction of discrepancy with additional subjects, larger increases in the number of subjects enrolled are needed to conduct experiments studying subtler effects.

When considering the effects of this on post-treatment measurements such as mean differences or T statistic, it is clear that a more precise measurement could be made when groups are well-matched at the onset. As we discuss below, well-matched groups yield a measurement that is much closer to the nominal (average or mode) measurement of pure randomization. Indeed, that this distribution of measurements is different (tighter) means that a naïve application of the Student T test would result in an underestimate of confidence and power, but that the distribution is tighter should allow for much more powerful inference.

6.3 Optimization Approach

Our proposal is to assign subjects so to minimize the discrepancies in centered first and second moments, where this assignment is gleaned via integer optimization. After assignment, we randomize which group is given which treatment, which ensures unbiased estimation as discussed in Section 6.5.

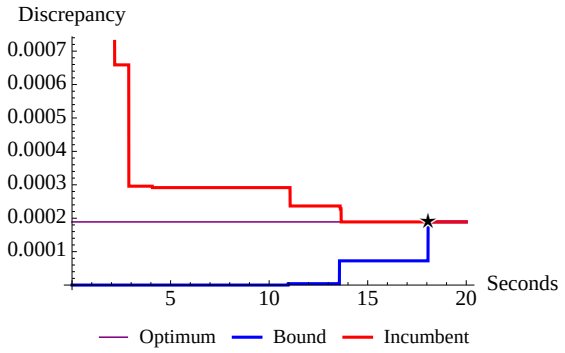
Given pre-treatment values of subjects $w_i,\ i = 1, \ldots, n = mk$, we are interested in creating m groups each containing k subjects in such a way that the discrepancy in means and ρ times the discrepancy in second moments is minimized between any two groups. We first preprocess the full sample by normalizing it so that it has zero sample mean and unit sample variance. We set

$$w_i' = (w_i - \hat{\mu})/\hat{\sigma}, \quad \text{where} \quad \hat{\mu} = \sum_{i=1}^{n} w_i/n \ \text{ and } \ \hat{\sigma}^2 = \sum_{i=1}^{n} (w_i - \hat{\mu})^2/n.$$

After construction of k groups, we randomize which treatment is given to which group. Algorithmically, we number the treatments and the groups in any way, shuffle the numbers $1, \ldots, m$ and treat the group in position j with treatment number j. This does not affect the objective value.

The parameter ρ controls the tradeoff between the discrepancy of first moments and of second moments and is chosen by the researcher. We introduce the decision variable $x_{ip} = 0$ or 1 to denote the assignment of subject i to group p. Using

Figure 6-2: The Progress of Solving an Instance of Problem (6.1) with $n = 40$, $m = 4$



continuous auxilliary variable d and letting

$$\mu_p(x) = \frac{1}{k}\sum_{i=1}^{n} w_i' x_{ip} \quad \text{and} \quad \sigma_p^2(x) = \frac{1}{k}\sum_{i=1}^{n} (w_i')^2 x_{ip},$$

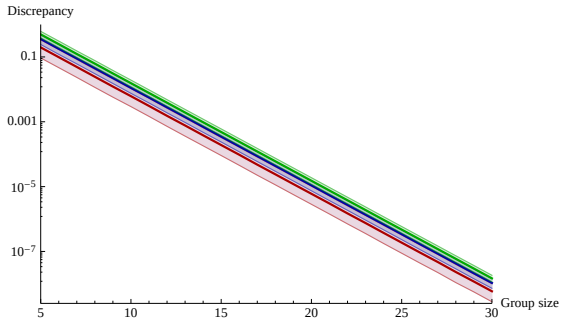
we formulate the problem as follows:

$$\begin{aligned}
Z_m^{\text{opt}}(\rho) &= \min_{x} \max_{p\neq q} \left(|\mu_p(x) - \mu_q(x)| + \rho\left|\sigma_p^2(x) - \sigma_q^2(x)\right|\right) \\
&= \min_{x,d} d \qquad (6.1)\\
&\quad \text{s.t. } \forall p < q = 1,\ldots,m: \\
&\qquad d \geq \mu_p(x) - \mu_q(x) + \rho\sigma_p^2(x) - \rho\sigma_q^2(x) \\
&\qquad d \geq \mu_p(x) - \mu_q(x) + \rho\sigma_q^2(x) - \rho\sigma_p^2(x) \\
&\qquad d \geq \mu_q(x) - \mu_p(x) + \rho\sigma_p^2(x) - \rho\sigma_q^2(x) \\
&\qquad d \geq \mu_q(x) - \mu_p(x) + \rho\sigma_q^2(x) - \rho\sigma_p^2(x) \\
&\quad x_{ip} \in \{0,1\} \\
&\quad \sum_{i=1}^{n} x_{ip} = k \;\; \forall p = 1,\ldots,m \\
&\quad \sum_{p=1}^{m} x_{ip} = 1 \;\; \forall i = 1,\ldots,n \\
&\quad x_{ip} = 0 \qquad \forall i < p.
\end{aligned}$$

As formulated, problem (6.1) is a mixed integer linear optimization problem with $m(1 + 2n - m)/2$ binary variables and 1 continuous variable. The last constraint reduces the redundancy in the branch-and-bound tree due to permutation symmetry. Further symmetry reduction is possible by methods described in Kaibel et al. (2011). Symmetry is reintroduced by randomizing which group receives which treatment.

We implement this optimization model in Gurobi v5.6 For values $n = 40$ and $m = 4$ problem (6.1) can be solved to full optimality in under twenty seconds on a personal

Figure 6-3: Discrepancy in Means Among Optimally Assigned Groups of Normal Variates with $\rho = 0$



Note: The colors are as in Figure 6-1. Notice the vertical log scale compared to the absolute scale of Figure 6-1.

computer with 8 processor cores. Gurobi also has built-in symmetry detection to avoid redundant computations in the branch-and-bound tree. We plot the progress of the branch and bound procedure for one example in Figure 6-2. For larger instances, Gurobi generally finds a solution with objective value that is near optimal within a few minutes, while finding the optimum can take longer and proving its optimality even longer.

The formulation of optimization problem (6.1) extends to multiple covariates. Suppose we are interested in matching the first and second moments in a vector of r covariates where w_{is} denotes the s^{th} covariate of subject i. Again, we normalize the sample to have zero sample mean and identity sample covariance by setting $\mathbf{w}_i' = \Gamma(\mathbf{w}_i - \hat{\mu})$, where Γ is the matrix square root of the (pseudo-)inverse of the sample covariance $\hat{\Sigma} = \sum_{i=1}^{n} (\mathbf{w}_i - \hat{\mu})(\mathbf{w}_i - \hat{\mu})^T / n$. Given the tradeoff parameter ρ, we rewrite the optimization problem for this case using $m(1 + 2n - m)/2$ binary variables and $1 + m(m-1)r(r+3)/4$ continuous variables as follows:

$$
\begin{aligned}
&\min\ d \\
&\text{s.t. } x \in \{0,1\}^{n \times m},\ x_{ip} = 0\ \forall i < p,\ d \geq 0 \\
&\quad \sum_{i=1}^{n} x_{ip} = k \ \ \forall p = 1, \ldots, m \\
&\quad \sum_{p=1}^{m} x_{ip} = 1 \ \ \forall i = 1, \ldots, n \\
&\quad x_{ip} = 0 \qquad \forall i < p \\
&\quad M \in \mathbb{R}^{\frac{m(m-1)}{2} \times r},\ V \in \mathbb{R}^{\frac{m(m-1)}{2} \times \frac{r(r+1)}{2}} \\
&\quad \forall p = 1, \ldots, m,\ q = p+1, \ldots, m: \\
&\qquad d \geq \sum_{s=1}^{r} M_{pqs} + \rho \sum_{s=1}^{r} V_{pqss} + 2\rho \sum_{s=1}^{r} \sum_{s'=s+1}^{r} V_{pqss'}
\end{aligned}
$$

$$
\begin{aligned}
&\forall s = 1, \ldots, r: \\
&\quad M_{pqs} \geq \frac{1}{k} \sum_{i=1}^{n} w'_{is} \left(x_{ip} - x_{iq} \right) \\
&\quad M_{pqs} \geq \frac{1}{k} \sum_{i=1}^{n} w'_{is} \left(x_{iq} - x_{ip} \right) \\
&\forall s = 1, \ldots, r,\ s' = s, \ldots, r: \\
&\quad V_{pqss'} \geq \frac{1}{k} \sum_{i=1}^{n} w'_{is} w'_{is'} \left(x_{ip} - x_{iq} \right) \\
&\quad V_{pqss'} \geq \frac{1}{k} \sum_{i=1}^{n} w'_{is} w'_{is'} \left(x_{iq} - x_{ip} \right).
\end{aligned}
$$

The potential extension to even higher moments is straightforward. More generally, such optimization procedures, along with complete randomization and pairwise matching, can all be interpreted under the unifying lens of minimizing worst-case variance; see Kallus (2014b).

6.4 Optimization vs. Randomization in Reducing Discrepancies

Using the above optimization model implemented in Gurobi v5.6, we conducted a series of simulations comparing the results of group assembly using randomization and optimization. Our key finding is that optimization is *starkly* superior to randomization in matching group means under all circumstances tested.

Figure 6-3 provides the analogue of Figure 6-1 for optimization and Figure 6-4 compares side-by-side the mismatch achieved in the first two moments by optimization and by randomization. In particular we show for various numbers of groups and group sizes the achievable range of feasible matchings as ρ varies. For all values of ρ, the pre-treatment discrepancy is significantly reduced compared to that seen under randomization, essentially eliminating population variance as a significant source of noise for all but the most extreme circumstances. Noting that discrepancy in either moment is minuscule under optimization using any of the values of ρ shown, we arbitrarily choose $\rho = 0.5$ for all further numerical examples unless otherwise noted. To revisit the example used to illustrate the limitations of randomization, the researcher assembling four groups of ten mice each under optimization with $\rho = 0.5$ would end up with 0.0005 standard deviations of discrepancy in first moment (or a twentieth of that for $\rho = 0$, not shown in figure), compared with 0.66 standard deviations under randomization.

There is some theoretical backing to the experimental evidence that optimization eliminates all discrepancies to such an extreme degree. When $\rho = 0$ and $m = 2$ the problem, scaled by $1/n$, reduces to the well-studied balanced number partitioning problem (see Karmarkar and Karp (1982)). Let Z_2^{rand} denote the discrepancy in means

under randomization. When pre-treatment covariates are random with variance σ^2, we have by Jensen's inequality that

$$\mathbb{E}\left[Z_2^{\text{rand}}\right] \leq \sqrt{\mathbb{E}\left[\left(Z_2^{\text{rand}}\right)^2\right]} = \sqrt{\frac{2}{k}}\sigma$$

and if they are normally distributed then

$$E[Z_2^{\text{rand}}] = \frac{2}{\sqrt{\pi k}}\sigma.$$

In comparison, an analysis of balanced number partitioning with random weights (see Karmarkar et al. (1986)) yields that there is a $C > 0$ such that

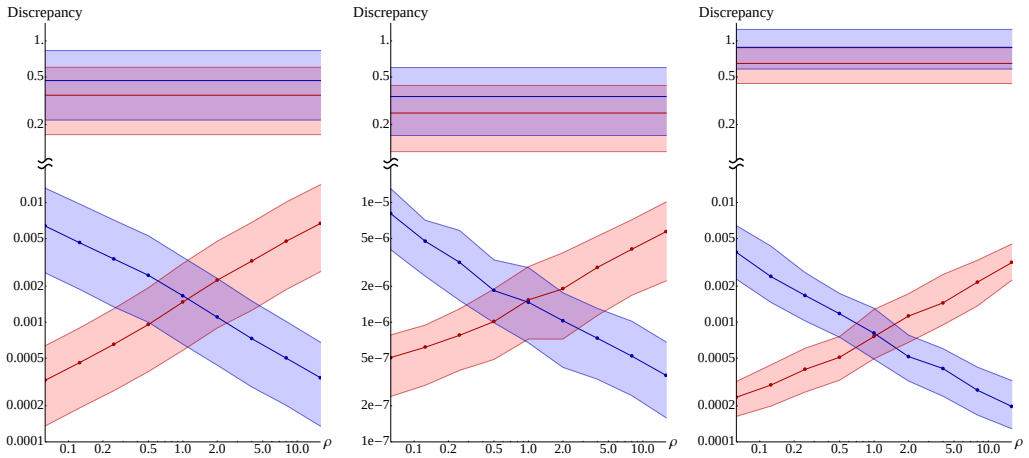
$$\text{median}\left(Z_2^{\text{opt}}(0)\right) \leq \frac{C}{2^{2k}}$$

and heuristic arguments from spin-glass theory (see Mertens (2001)) provide the prediction

$$E[Z_2^{\text{opt}}(0)] = \frac{2\pi\sigma}{2^k},$$

which agrees with our experimental results for large k. Comparing the asymptotic

Figure 6-4: The Range of Achievable Discrepancies Under Optimization and Under Randomization



(a) m=2 groups n=20 subjects (b) m=2 groups n=40 subjects (c) m=4 groups n=40 subjects

Note: The upper halves of the plots correspond to randomization and the lower ones to optimization. Red denotes discrepancy in mean and blue variance. The bands depict average under- and over-shoot. Notice the log scales and the break in the vertical axis.

Table 6.1: The Number of Subjects Per Group Needed to Guarantee an Expected Discrepancy No More Than $\epsilon\sigma$ for $m = 2$ and $\rho = 0$

ϵ	k^{Opt}	k^{Rand}	k^{Pair}	k^{RR}
0.1	3	128	9	4
0.01	5	12833	65	83
0.001	7	1273240	514	8130
0.0001	8	127323955	4354	820143

orders of Z_2^{rand} and $Z_2^{\text{opt}}(0)$, we see an *exponential reduction* in discrepancies by optimization versus randomization.

Matching done on a subject-pair-wise basis such as caliper matching as done in propensity score matching (see Rubin (1979)) does not close this gap either even when the sample-based optimal caliper width is chosen. Consider for simplicity uniformly-distributed pre-treatment covariates so that any subsequent difference of two nearest neighbors are on average $(n+1)^{-1}$. If assignment within each pair is randomized independently a simple calculation then shows that the average discrepancy is of order $k^{-3/2}$ whereas if assignment is alternating among the sorted covariates then the average discrepancy is of order k^{-1}. The case is worse for normally-distributed covariates as reported below.

Following the average predictions for the normal distribution, if we want to limit discrepancy to some fraction of the standard deviation, $\epsilon\sigma$, we see a dramatic difference in the necessary number of subjects per group, k:

$$k^{\text{Opt}} = \left\lceil \log_2 \frac{2\pi}{\epsilon} \right\rceil, \qquad k^{\text{Rand}} = \left\lceil \frac{4}{\pi\epsilon^2} \right\rceil.$$

In Table 6.1 we report specific values of k^{Opt} and k^{Rand}, as well as k^{PW} corresponding to optimal pairwise matching and k^{RR} corresponding to the Mahalanobis-distance re-randomization method of Morgan et al. (2012) with a fixed acceptance probability of 5%.[2] This is a clear example of the power of optimization for experiments hindered by small samples. While pairwise matching and re-randomization improve upon randomization, they are significantly outperformed by optimization especially when small discrepancy is desired.

A concern may be that by optimizing only the first two moments and not others those higher moments may become mismatched. We find, however, that this is not the case even when compared to all the other methods considered above. In Table 6.2 we tabulate the mismatch in the first five moments and in the generalized moment of log for the various methods when assigning $2k$ subjects with baseline covariates drawn from a standard normal population. In Table 6.3 we tabulate the mistmatch of multivariate moments for the various methods when assigning $2k$ subjects with multivariate baseline covariates drawn from a three-dimensional standard

[2]Simulation is used to glean k^{opt} for these values of ϵ, for which the asymptotic predictions yield overestimates. Simulation also shows that for FSM, $k^{\text{FSM}} \approx k^{\text{Rand}}$.

Table 6.2: The Discrepancy in Various Moments Under Different Assignment Mechanisms

		Moment					
k	Method	1	2	3	4	5	log
5	Opt	0.0513	0.286	1.43	2.67	9.75	0.498
	Rand	0.510	0.689	1.79	3.81	10.3	0.544
	Pair	0.184	0.498	1.27	3.29	8.93	0.345
	Re-rand	0.047	0.711	1.09	3.88	8.47	0.572
	FSM	0.508	0.553	1.76	3.33	10.2	0.440
10	Opt	0.00174	0.0145	0.906	1.47	6.87	0.338
	Rand	0.352	0.504	1.30	2.88	7.79	0.399
	Pair	0.0839	0.259	0.759	2.09	6.06	0.176
	Re-rand	0.0298	0.497	0.764	2.93	6.20	0.389
	FSM	0.374	0.334	1.33	2.26	7.90	0.264
20	Opt	1.23e-6	2.34e-6	0.600	1.04	5.23	0.221
	Rand	0.258	0.345	0.947	2.13	6.13	0.276
	Pair	0.0379	0.140	0.445	1.40	4.24	0.286
	Re-rand	0.0207	0.356	0.565	2.16	4.99	0.284
	FSM	0.249	0.190	0.896	1.50	5.89	0.146

Note: Column ℓ corresponds to the average mismatch in the ℓ^{th} moments between the two groups and the last column corresponds to the mismatch in the generalized moments in $\log|w|$.

normal population. For pairwise matching we use the Mahalanobis pairwise distance, for re-randomization we use an acceptance probability of 5%, for FSM we use the method implied by equation (2.11) of Morris (1979) with $c_i = 1$, $T = I$, and for our method we use $\rho = 0.5$. We notice that optimal assignment yields superior balance in the moments considered and that all methods result in similar balance for those moments not directly considered in the optimization problem.

6.5 Optimization, Randomization, and Bias

Randomization has traditionally been used to address two kinds of bias in experimental design. The first is investigator bias, or the possibility that an investigator may subconsciously or consciously construct experimental groups in a manner that biases toward achieving a particular result. As a fixed, mechanical process, optimization guards against this possibility at least as well as randomization. Indeed it does better because any manual manipulation of the optimized results would make the result less well-matched than the reproducible optimum, which is checkable, whereas no one

Table 6.3: The Discrepancy in Various Multivariate Moments Under Different Assignment Mechanisms

		Moment					
k	Method	w_1	w_1^2	w_1w_2	w_1^3	$w_1^2w_2$	$w_1w_2w_3$
10	Opt	0.0701	0.145	0.183	0.93	0.508	0.337
	Rand	0.360	0.492	0.344	1.29	0.58	0.333
	Pair	0.179	0.383	0.271	0.964	0.478	0.299
	Re-rand	0.141	0.493	0.357	0.883	0.484	0.34
	FSM	0.368	0.606	0.503	1.30	0.574	0.340
15	Opt	0.0230	0.0450	0.117	0.718	0.411	0.292
	Rand	0.292	0.400	0.286	1.05	0.489	0.289
	Pair	0.125	0.290	0.201	0.748	0.38	0.247
	Re-rand	0.113	0.409	0.289	0.714	0.414	0.293
	FSM	0.289	0.597	0.491	1.05	0.488	0.281
25	Opt	0.00302	0.00497	0.0780	0.547	0.315	0.227
	Rand	0.226	0.325	0.222	0.842	0.384	0.227
	Pair	0.0849	0.196	0.143	0.547	0.276	0.172
	Re-rand	0.0863	0.326	0.230	0.566	0.314	0.220
	FSM	0.219	0.592	0.494	0.823	0.388	0.224

Note: Column w_1w_2, for example, corresponds to the average mismatch in the moments of w_1w_2 between the two groups, which by symmetry is the same as that of w_1w_3 or w_2w_3 on average.

grouping can ever be verified to truly be the result of pure randomization.

The second sort of bias is the incidental disproportionate assignment of variables, measured or hidden, that directly affect the treatment. Randomization, given large enough samples, will tend to equalize the apportionment of any one factor. However, just as with the measured covariates w_i, randomization cannot be counted upon to eliminate discrepancies in hidden factors when samples are relatively small. Optimization considers the measured covariates w_i when allocating a subject to a particular group. For all factors that are independent with this variable, the allocation remains just as random. Variables that are correlated with the measured covariates in ways such as joint normality will be just as well balanced as the measured covariates and variables with a higher order dependence, such as having a polynomial conditional expectation in w, would be as balanced as seen in Tables 6.2 and 6.3.

In general, the observed difference in treatment effects after optimizing the assignment as described herein will always be an *unbiased estimator* of the true population average difference, as in a randomized experiment. This is a consequence of randomizing the identity of treatments (while optimizing the partition of subjects) so that the assignment of a single subject is marginally independent of its potential responses

to different treatments.[3] Unbiasedness in estimation means that were the experiment to be repeated many times and the results recorded, the average result would coincide with the true value. In particular, there is no omitted variable bias. That is, neglecting to take into consideration a relevant covariate does not introduce bias in estimation.

6.6 Optimization vs. Randomization in Making a Conclusion

As we have shown in the previous sections, optimization eliminates nearly all noise due to pre-treatment covariates. One would then expect that it can also offer superior precision in estimating the differences between treatments and superior power in making statistical inferences on these differences.

In randomized trials, randomization tests (see Edgington and Onghena (2007)) can be used to draw inferences based directly on the randomness of assignment without normality assumptions, which often fail for small samples. However, for optimization the assignment is not random enough and this test is not applicable. For the purpose of testing differences of treatments in an optimized trial, we propose the following test based on the bootstrap (see Efron and Tibshirani (1993)).

Comparing between two treatments, we would like to test the null hypothesis that every subject $i = 1, \ldots, n$ would have had the same response to treatment whether either of the two treatments were assigned (this is known as the sharp null hypothesis; see Rubin (1980)). Let v_i denote the response measured for subject i after it was administered the treatment to which it was assigned. Given subjects with covariates $w_1, \ldots, w_n$, the test we propose is as follows:

1. Find an optimal assignment of these to two groups (permuting randomly):

$$\{i_1, \ldots, i_{n/2}\} \text{ and } \{i_{n/2+1}, \ldots, i_n\}.$$

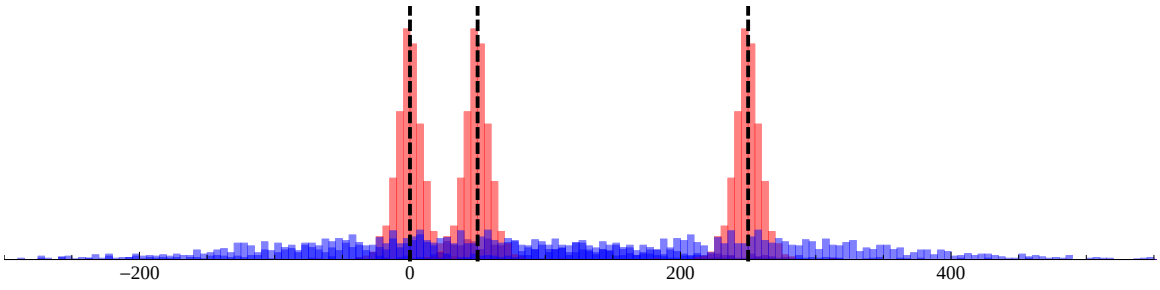
2. Administer treatments and measure responses v_i, which are henceforth fixed.
3. Compute $\delta = \frac{1}{k}\left(v_{i_1} + \cdots + v_{i_{n/2}}\right) - \frac{1}{k}\left(v_{i_{n/2+1}} + \cdots + v_{i_n}\right)$.
4. For $b = 1, \ldots, B$:
   (a) Draw a random sample with replacement $w_{b,1}, \ldots, w_{b,n}$ from $w_1, \ldots, w_n$.
   (b) Find an optimal assignment of these to two groups (permuting randomly):

$$\{i_{b,1}, \ldots, i_{b,n/2}\} \text{ and } \{i_{b,n/2+1}, \ldots, i_{b,n}\}.$$

   (c) Compute $\delta_b = \frac{1}{k}\left(v_{i_{b,1}} + \cdots + v_{i_{b,n/2}}\right) - \frac{1}{k}\left(v_{i_{b,n/2+1}} + \cdots + v_{i_{b,n}}\right)$.

[3] The correctness of modeling using potential outcomes is contingent on the stable unit treatment value assumption. See Rubin (1986).

Figure 6-5: The Distribution of Estimates of Effect Size Under Optimization and Randomization



Note: Optimization is shown in red and randomization in blue. $k = 20$ and effect sizes vary among 0mg, 50mg, and 250mg (dashed lines). The overlap of estimates under randomization of the nonzero effects and of the zero effect elucidate the low statistical power of randomization in detecting the nonzero effects.

5. Compute the p-value $p = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{I}\left[|\delta_b| \geq |\delta|\right]\right)$.

Then, to test our null hypothesis at a significance of α, we only reject it if $p \leq \alpha$. The quantity δ above constitutes our estimate of the difference between the two treatments.

To examine the effect of optimization on making a conclusion about the treatments we consider again the example of a murine tumor study. We consider two groups, each of k mice, with tumor weights initially normally distributed with mean 200mg and standard deviation 300mg (truncated to be nonnegative). Two treatments are considered: a placebo and a proposed treatment. Their effect on the tumor, allowed to grow for a period of a day, is of interest to the study.

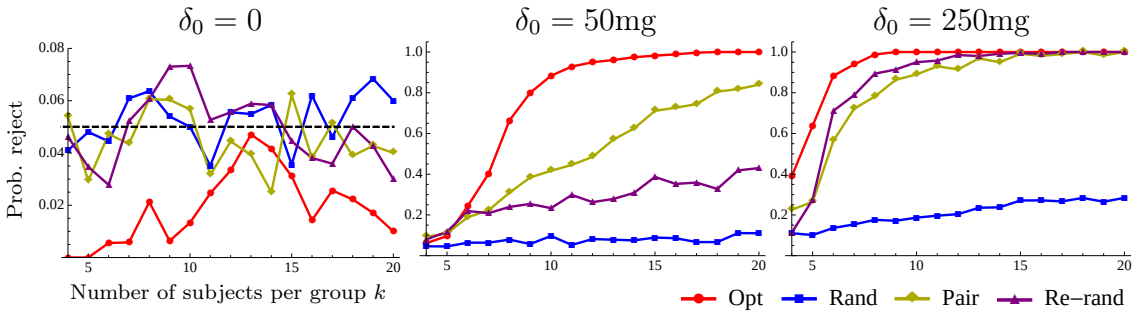
The effects of treatment and placebo are unknown and are to be inferred from the experiment. We consider a hidden reality where the growth of the tumors are dictated by the Gomp-ex model of tumor growth (see Wheldon (1988)). That is, growth is governed by the differential equation:

$$\frac{dw}{dt} = w(t)\left(a + \max\left\{0, b\log\left(w_c/w(t)\right)\right\}\right),$$

where a and b are rate parameters and w_c is the critical weight that marks the change between exponential and logistic growth. We arbitrarily choose $a = 1\frac{1}{\text{day}}$, $b = 5\frac{1}{\text{day}}$, $w_c = 400$mg, and $t = 1$day. We pretend that tumors under either treatment grow according to this equation but subtract δ_0 from the final weights for the proposed treatment. We consider δ_0 being 0mg (no effect), 50mg (small effect), and 250mg (large effect).

For various values of k and for several draws of initial weights, we consider assignments produced by randomization, our optimization approach ($\rho = 0.5$), pairwise matching, and re-randomization. We consider both the post-treatment estimate of the effect and the inference drawn on it at a significance of $\alpha = 0.05$, using our boot-

Figure 6-6: The Probability of Rejecting the Null Hypothesis of No Effect for Various Effect Sizes



strap test for our method and the standard randomization test for the others.[4] In Figure 6-5 we plot the resulting estimates for $k = 20$ and in Figure 6-6 we plot the rates at which the null hypothesis is rejected. When there is no effect, this rate should be no more than the significance $\alpha = 0.05$.[5] When there is an effect, we would want the rate to be as close to 1 as possible. In a sense, the complement of this rate is the fraction of experiments squandered in pursuit of an effective drug. The cost-saving benefits of optimization in this case are clear.

The exact improvements in precision and power depend on the nature of treatment effect. However, comparisons to the existing methods are possible. Morgan et al. (2012) study reduction in variance due to re-randomization only under the additive treatment model, a very restrictive assumption. In this setting, when setting $\rho = 0$, the same analysis as provided in their Theorem 3.2 provides that the reduction in variance provided by en-masse optimization is exponentially better because the reduction in mean mismatch is exponentially better. Nonetheless, treatment effects usually do depend, albeit perhaps to a lesser extent, on higher orders of the covariates and on their interactions. In Tables 6.2 and 6.3 we saw that optimization balances higher and interaction moments no worse than other methods (better for second moments).

6.7 Practical Significance

Here we present evidence that optimization produces groups that are far more similar in mean and variance than those created by randomization, especially in situations in which group size is small, data variability is large, and numerous groups are needed for a single experiment. For each additional subject per group, optimization roughly halves the discrepancy in the covariate, whereas both randomization and subject-pair matchings offer quickly diminishing reductions. Making groups similar before

[4]For non-completely-randomized designs, the randomization test draws random re-assignments according to the method employed at the onset. See Chapter 10 of Edgington and Onghena (2007).

[5]The fact that for our bootstrap test this rate is below 0.05 may be an indication that the test is conservative, i.e., more significant than designed. Nonetheless, despite such conservatism, the test is still more powerful than the other tests.

treatment allows for statistical power beyond what can normally be hoped for with small samples.

We propose that optimization protects against experimental biases at least as well as randomization and that the advantage of optimized groups over randomized groups is substantial. We believe that optimization of experimental group composition, implementable on commonplace software such as Microsoft Excel and on commercial mathematical optimization software, is a practical and desirable alternative to randomization that can improve experimental power in numerous fields, such as cancer research, neurobiology, immunology, investment analysis, market research, behavioral research, proof-of-concept clinical trials, and others.

Chapter 7

Optimal A Priori Balance in the Design of Controlled Experiments

We develop a unified theory of designs for controlled experiments that balance baseline covariates a priori (before treatment and before randomization) using the framework of minimax variance. We establish a "no free lunch" theorem that indicates that, without structural information on the dependence of potential outcomes on baseline covariates, complete randomization is optimal. Restricting the structure of dependence, either parametrically or non-parametrically, leads directly to imbalance metrics and optimal designs. Certain choices of this structure recover known imbalance metrics and designs previously developed ad hoc, including randomized block designs, pairwise-matched designs, and re-randomization. New choices of structure based on reproducing kernel Hilbert spaces lead to new methods, both parametric and non-parametric.

7.1 Introduction

Achieving balance between experimental groups is a corner stone of causal inference, otherwise any observed difference may be attributed to a difference other than the treatment alone. In clinical trials, and more generally controlled experiments, where the experimenter controls the administration of treatment, complete randomization of subjects has been the golden standard for achieving this balance on average.

The expediency of complete randomization, however, has been controversial since the founding of statistical inference in controlled experiments. William Gosset, "Student" of Student's T-test, said of assigning field plots to agricultural interventions that it "would be pedantic to continue with an arrangement of [field] plots known beforehand to be likely to lead to a misleading conclusion," such as arrangements in which one experimental group is on average higher on what he calls the "fertility slope" than the other experimental group Student (1938). Of course, as the opposite is just as likely under complete randomization, this is not an issue of estimation bias in its modern definition, but of estimation variance. Gosset's sentiment is echoed in the common statistical maxim "block what you can, randomize what you cannot"

attributed to George Box and in the words of such individuals as James Heckman ("Randomization is a metaphor and not an ideal or 'gold standard"' Heckman (2008)) and Donald Rubin ("For gold standard answers, complete randomization may not be good enough" Rubin (2008)). In one interpretation, these can be seen as calls for the experimenter to ensure experimental groups are balanced at the onset of the experiment, before applying treatments and before randomization.

There is a variety of designs for controlled experiments that attempt to achieve better balance in terms of measurements made prior to treatment, known as baseline covariates, under the understanding that a predictive relationship possibly holds between baseline covariates and the outcomes of treatment. We term this sort of approach *a priori balancing* as it is done before applying treatments and before randomization (the term a priori is chosen to contrast with post hoc methods such as post stratification, which may be applied after randomization and after treatment McHugh and Matts (1983)). The most notable a priori balancing designs are randomized block designs Fisher (1935), pairwise matching Greevy et al. (2004), and re-randomization Morgan et al. (2012).[1]

Each of these implicitly defines imbalance between experimental groups differently. Blocking attempts to achieve exact matching (when possible): a binary measure of imbalance that is zero only if the experimental groups are identical in their discrete or coarsened baseline covariates. Pairwise matching treats imbalance as the sum of pairwise distances, given some pairwise distance metric such as Mahalanobis. There are both globally optimal and greedy heuristic methods that address this imbalance measure Gu and Rosenbaum (1993). In Morgan et al. (2012), the authors define imbalance as the group-wise Mahalanobis distance and propose re-randomization as a heuristic method for reducing it non-optimally.

It is unclear when each of these different characterizations of imbalance is appropriate and when is deviating from complete randomization justified. The connection between an imbalance metric such as the sum of pairwise distances before treatment and estimation variance after treatment is also unclear. We here argue that, without structural information on the dependence of outcomes on baseline covariates, complete randomization is minimax optimal. Furthermore, when structural knowledge is expressed as membership of conditional expectations in a normed vector space of functions, an alternative minimax-optimal rule arises for the a priori balancing of experimental groups. We show how certain choices of this structure reconstruct each of the aforementioned methods or associated imbalance metrics. We study other choices of structure using reproducing kernel Hilbert spaces (RKHS), which give rise to new methods, both parametric and non-parametric.

We study in generality the characteristics of any such method that arises from our framework, including its estimation variance and consistency, intimately connecting a priori balance to post-treatment estimation. Whenever a parametric model of de-

[1]There are also sequential methods to address the case where allocation must be decided before all subjects are admitted Efron (1971), Pocock and Simon (1975), Kapelner and Krieger (2014). These are beyond the present scope of this chapter. Response-adaptive designs that use outcome data to inform future assignments (see Chow et al. (2008)) lie between a priori and post hoc and are also beyond our scope.

pendence is known to hold, we show that, relative to complete randomization, the variance due to the optimal design converges linearly ($2^{-\Omega(n)}$ for n subjects) to the best theoretically possible – a generalization of the observation on linear convergence made in Bertsimas et al.. We provide algorithms for finding the optimal designs using mixed integer optimization (MIO) and semi-definite optimization (SDO) and hypothesis tests that are appropriate for these designs. We make connections to Bayesian experimental design and shed light on the usefulness of a priori balance in designing experiments plagued by non-compliance.

7.1.1 Structure of this Chapter

In Section 7.2, we consider the effect of structure and the lack thereof. In particular, we set up the problem, argue that complete randomization is optimal in the absence of structural information (Section 7.2.1), define structural information and the resulting imbalance metrics and optimal designs (Section 7.2.2), show how this recovers existing imbalance metrics and designs (Section 7.2.3), study the designs that arise from RKHS structure (Section 7.2.4), and consider a Bayesian interpretation (Section 7.2.4). We end Section 7.2 with simulation studies of fictitious data (Example 7.11) and of clinical data (Example 7.12). In Section 7.3, we characterize the variance (Section 7.3.1), consistency (Section 7.3.2), and rate of convergence (Section 7.3.3) of estimators arising from a priori balancing designs. In Section 7.4, we provide algorithms for finding the optimal designs. In Section 7.5, we provide hypothesis tests for making inferences on treatment effects. We offer some concluding remarks in Section 7.6.

All proofs are given in the supplement. In the supplement, we also consider the benefit of a priori balancing to experiments plagued by non-compliance (Section E.1) and generalizations of structural information (Section E.2).

7.2 The Effect of Structural Information and Lack Thereof

We begin by describing the set up. Let m denote the number of treatments to be investigated (including controls). We index the subjects by $i = 1, \ldots, n$ and assume $n = mp$ is divisible by m. We assume the subjects are independently randomly sampled but we will consider estimating both sample and population effects. We denote assigning subject i to a treatment k by $W_i = k$. We let $w_{ik} = \mathbb{I}[W_i = k]$ and $W = (W_1, \ldots, W_n)$. When $m = 2$, we will use $u_i = w_{i1} - w_{i2}$. As is common for controlled trials, we assume non-interference (see e.g. Rubin (1986), Rosenbaum (2007) and p. 19 of Cox (1958)). I.e., a subject assigned to a certain treatment exhibits the same outcome regardless of others' assignments. Under this assumption we are able to define the potential post-treatment outcome Y_{ik} of subject i were it to be subjected to the treatment k. We let Y denote the matrix of all potential outcomes. We assume throughout Y_{ik} has second moments. Let X_i, taking values in some $\mathcal{X}$, be the baseline covariates of subject i that are recorded before treatment

and let $X = (X_1, \ldots, X_n)$.

We denote by $\mathrm{TE}_{kk'i} = Y_{ik} - Y_{ik'}$ the unobservable causal treatment effect for subject i. There are two unobservable quantities that will be of interest to estimate. One is the *sample average (causal) treatment effect* (SATE):

$$\mathrm{SATE}_{kk'} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{TE}_{kk'i} = \frac{1}{n}\sum_{i=1}^{n} Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} Y_{ik'}.$$

Another is the *population average (causal) treatment effect* (PATE):

$$\mathrm{PATE}_{kk'} = \mathbb{E}\left[\mathrm{TE}_{kk'1}\right] = \mathbb{E}\left[\mathrm{SATE}_{kk'}\right].$$

By construction, SATE is an unbiased and strongly consistent estimate of PATE. Our estimator will always be the simple mean differences estimator

$$\hat{\tau}_{kk'} = \frac{\sum_{i:W_i=k} Y_{ik}}{\sum_{i:W_i=k} 1} - \frac{\sum_{i:W_i=k'} Y_{ik'}}{\sum_{i:W_i=k'} 1}.$$

We drop subscripts when $m = 2$ and set $k = 1$, $k' = 2$.

Throughout we will consider only designs that

$$\begin{aligned}
&\text{do not depend on future information, that is, } W \text{ is independent of } Y, \text{ conditional on } X; && (7.1)\\
&\text{blind (randomize) the identity of treatments, that is, } \mathbb{P}\left(W = (k_1, \ldots, k_n)\,|X\right) = \mathbb{P}\left(W = (\pi(k_1), \ldots, \pi(k_n))\,|X\right) && (7.2)\\
&\text{for any permutation } \pi \text{ of } 1, \ldots, m; \text{ and}\\
&\text{split the sample evenly, that is, surely } \textstyle\sum_{i:W_i=k} 1 = p \quad \forall k. && (7.3)
\end{aligned}$$

We interpret conditions (7.1)-(7.3) as the definition of *a priori balance* as they require that all balancing be done before applying treatments (condition (7.1)) and before randomization (conditions (7.2)-(7.3)). Condition (7.1) is a reflection of the temporal logic of first assigning, then experimenting. Condition (7.2) says that balancing is done before randomization and it ensures that the estimators $\hat{\tau}_{kk'}$ resulting from the design are always unbiased, both conditionally on X, Y (i.e., in estimating SATE) and marginally (i.e., in estimating PATE; more detail given in Theorem 7.13). Condition (7.3) is a way to achieve (7.2) in non-completely-randomized designs. If W is an even assignment then randomly permuting treatment indices will blind their identity. Else, given one fixed uneven assignment, a treatment can be identified by the size of its experimental group.

We denote by $\mathcal{W} \subset \{1, \ldots, m\}^n$ the space of feasible assignments satisfying (7.3) and by $\Delta \subset [0,1]^{|\mathcal{W}|}$ the space of feasible designs (distributions over assignments) satisfying (7.1)-(7.3). For $m = 2$ we also write $\mathcal{W} \cong \mathcal{U} = \{u \in \{-1, +1\}^n : \sum_i u_i = 0\}$ and $\mathcal{P} = \text{covex-hull}(\mathcal{U})$.

7.2.1 No Free Lunch

We will now argue that without structural information on the relationship between X_i and Y_{ik}, complete randomization is minimax optimal. For the rest of this subsection we will restrict to $m = 2$.

Among estimators that are unbiased, the standard way of comparing efficiency is variance. By the law of total variance and by the conditional unbiasedness of any estimator resulting from a design satisfying (7.1)-(7.3),

$$\mathrm{Var}\left(\hat{\tau}\right) = \mathbb{E}\left[\mathrm{Var}\left(\hat{\tau}|X,Y\right)\right] + \mathrm{Var}\left(\mathrm{SATE}\right).$$

The variance of SATE is independent of our choice of a priori balancing design. This choice can only affect the first term. Therefore, an efficient design will seek to minimize $\mathrm{Var}\left(\hat{\tau}|X,Y\right)$ path-by-path, i.e. for the given subjects at hand. Whatever the design does to minimize this term will not affect the second term as long as the design adheres to the above conditions.

Denote by $\hat{\tau}^{\mathrm{CR}}$ the estimator arising from complete randomization, which randomizes uniformly over equal partitions independently of X. Then,

$$\mathrm{Var}\left(\hat{\tau}^{\mathrm{CR}}|X,Y\right) = \frac{4}{n(n-1)}\left|\left|\overline{Y}\right|\right|_2^2$$
$$\text{where } \hat{Y}_i = \frac{Y_{i1}+Y_{i2}}{2},\ \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i,\ \text{and } \overline{Y}_i = \hat{Y}_i - \hat{\mu}.$$

Using this as a benchmark, we compare efficiency based on the normalized unitless ratio $\mathrm{Var}\left(\hat{\tau}|X,Y\right) / \mathrm{Var}\left(\hat{\tau}^{\mathrm{CR}}|X,Y\right)$.

However, we do not know Y, only X (condition (7.1)), and we assume no structural information on their relationship. Therefore, we consider an adversarial Nature that chooses Y so to increase our variance. The following shows that in this situation, complete randomization is optimal.

Theorem 7.1. *Fix $X \in \mathcal{X}^n$. Let $||\cdot||$ be any permutationally invariant seminorm on $\mathbb{R}^n$. Then, among designs satisfying* (7.1)-(7.3) *(i.e., among all $\sigma \in \Delta$), complete randomization minimizes either of*

$$\max_{Y\in\mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|Y_1\|^2+\|Y_2\|^2} = \max_{Y\in\mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|\hat{Y}\|^2} = \max_{Y\in\mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|\overline{Y}\|^2}$$
$$\textit{or, for } ||\cdot|| = ||\cdot||_2, \qquad \max_{Y\in\mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\mathrm{Var}\left(\hat{\tau}^{\mathrm{CR}}|X,Y\right)}.$$

In particular, if one randomly permutes a single *fixed* partition then

$$\max_{Y\in\mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\mathrm{Var}\left(\hat{\tau}^{CR}|X,Y\right)} = n-1. \tag{7.4}$$

Example 7.2. Fix $n = 2^b$ a power of two and $m = 2$. Let

$$X_i = \sum_{t=0}^{b-\max\{2,\log_2 i\}} (-1)^{\lceil i/2^{t-1}\rceil} 2^{-2^{b-1}+2^{b-t-1}+\left(i-1 \bmod 2^{t-1}\right)},$$
$$Y_i = (-1)^i = (-1)^{\log_2(\text{round}(|X_i|))}.$$

This rather complicated construction essentially yields

$$X \approx \text{round}(X) = \left(-1, -2, -4, \ldots, -2^{2^{b-1}-1}, 1, 2, 4, \ldots, 2^{2^{b-1}-1}\right)$$

with some perturbations so that the assignment $W = (1, 2, 1, 2, \ldots, 1, 2)$ uniquely minimizes the group-wise Mahalanobis distance of Morgan et al. (2012). Although X_i completely determines Y_{ik}, we are going to see that complete randomization beats blocking, pairwise matching, and re-randomization in this case. For blocking for $b \geq 4$, let us coarsen the space of baseline covariates into eight consecutive intervals so that each contains the same number of subjects, 2^{b-3}. For pairwise matching, let us use the pairwise Mahalanobis distance. And, for re-randomizaiton of Morgan et al. (2012), we consider both a 1% acceptance probability and an infinitesimal acceptance probability that essentially minimizes the group-wise Mahalanobis metric. We plot the resulting conditional variances $\text{Var}\left(\hat{\tau}\middle|X, Y\right)$ in Figure 7-1. Specifically, we get that complete randomization has a variance of $4/(n-1)$ whereas blocking has $4/(n-8)$, pairwise matching has $8/n$, and re-randomization with infinitesimal acceptance probability has 4, which realizes the worst-case ratio of (7.4) (it can be verified that this construction also realizes the corresponding worst-case ratios for blocking and pairwise matching). The variance of re-randomization with 1% acceptance is similar to infinitesimal acceptance probability for small n and becomes more similar to randomization as n grows. In each case, complete randomization does better, providing a concrete example of the conclusion of Theorem 7.1.

7.2.2 Structural Information and Optimal Designs

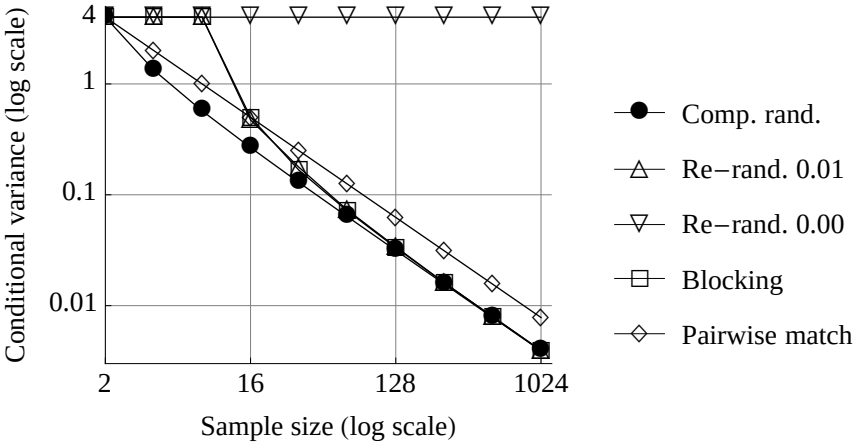
In the above we argued that from a minimax-variance perspective, complete randomization is optimal when no structural information about the dependence between X_i and Y_{ik} is available. We now consider the effect of such information, which we express as structure on the conditional expectations of outcomes.

Let us denote

$$f_k(x) := \mathbb{E}\left[Y_{ik}\middle|X_i = x\right] \quad \text{and} \quad \epsilon_{ik} := Y_{ik} - f_k\left(X_i\right).$$

The non-random function f_k is interchangeably called the *conditional expectation function* or *regression function*. The law of iterated expectation yields that ϵ_{ik} has mean 0, is mean-independent of X_i, and is uncorrelated with any function of X_i.

Figure 7-1: Variance of Estimating Effect Size Under Various Designs in Example 7.2 Conditional on the given X and Y Values



Combined with independence of subjects, this yields[2]

$$\mathrm{Var}\left(\hat{\tau}\right) = \mathbb{E}\left[\mathrm{Var}\left(B(W,\hat{f})\middle|X\right)\right] + \frac{1}{n}\mathrm{Var}\left(\epsilon_{11}+\epsilon_{12}\right) + \mathrm{Var}\left(\mathrm{SATE}\right),$$

$$\text{where } B(W,\hat{f}) = \frac{2}{n}\sum_{i:W_i=1}\hat{f}(X_i) - \frac{2}{n}\sum_{i:W_i=2}\hat{f}(X_i),\ \ \hat{f}(x) = \frac{f_1(x)+f_2(x)}{2}.$$

As before, the marginal variances of SATE and of $(\epsilon_{11}+\epsilon_{12})$ are completely independent of our choice of design and an efficient design will seek to minimize $\mathrm{Var}\left(B(W,\hat{f})\middle|X\right) = \mathbb{E}\left[B(W,\hat{f})^2\middle|X\right]$ path-by-path, i.e. for the given subjects. Now the unknown is $\hat{f}$ and we let Nature choose it adversarially. We will seek to minimize $\mathbb{E}\left[B(W,\hat{f})^2\middle|X\right]$ relative to the magnitude of $\hat{f}$, instead of the magnitude of $\hat{Y}$.

To define a magnitude of $\hat{f}$, we assume that $f_k \in \mathcal{F}\ \forall k$, where $\mathcal{F}$ is a normed vector space with norm $||\cdot|| : \mathcal{F} \to \mathbb{R}_+$. This will represent our structural information about the dependence between X_i and Y_{ik}. This space is a subspace of the vector space $\mathcal{V}$ of all functions $\mathcal{X} \to \mathbb{R}$ under the usual point-wise addition and scaling. For functions f that are not in $\mathcal{F}$ we formally define $||f|| = \infty$. When $\mathcal{F}$ is finite-dimensional, the assumption $\|f_k\| < \infty$ is a parametric one. When $\mathcal{F}$ is infinite-dimensional, it is non-parametric.

Because $B(W,\hat{f})$ is invariant to constant shifts to $\hat{f}$, i.e.,

$$B(W,\hat{f}) = B(W,\hat{f}+c) \ \text{ for } c \in \mathbb{R} \text{ representing a constant function } x \mapsto c,$$

we will want to factor this artifact away. The quotient space $\mathcal{F}/\mathbb{R}$ consists of the classes $[f] = \{f + c : c \in \mathbb{R}\}$ with the norm $||[f]|| = \min_{c\in\mathbb{R}} ||f+c||$. Without loss of generality, we always restrict to this quotient space and write $||f||$ to actually mean

[2]Theorem 7.13 gives an explicit derivation of this decomposition (for general $m \geq 2$).

the norm in this quotient space. Moreover, for worst-case variances to exist, we will restrict our attention to Banach spaces and require that differences in evaluations are continuous (i.e., the map $f \mapsto (f(X_i) - f(X_j))$ is continuous for each i, j). A Banach space is a normed vector space that is a complete metric space (see Ledoux and Talagrand (1991) and Chapter 10 of Royden (1988)).

With all structural information summarized by $\|f_k\| < \infty$, the motivation for the designs we develop next is the bound on the variance that arises:

$$\mathbb{E}\left[B^2(W,\hat{f})|X\right] \leq \|\hat{f}\|^2 \max_{f\in\mathcal{F}} \frac{\mathbb{E}\left[B^2(W,f)|X\right]}{\|f\|^2} = \|\hat{f}\|^2 \max_{\|f\|\leq 1} \mathbb{E}\left[B^2(W,f)|X\right].$$

Minimizing the above bound is independent of the actual value of $\|\hat{f}\|$ as it merely scales the objective. We will study this bound further and in greater generality in Theorems 7.13 and 7.14, leaving this as mere motivation for now.[3]

Borrowing terminology from game theory, we define two type of designs that seek to minimize this bound: the *pure-strategy optimal design* and the *mixed-strategy optimal design*. We now consider general $m \geq 2$ and define

$$B_{kk'}(W,\hat{f}) = \frac{1}{p}\sum_{i:W_i=k} \hat{f}(X_i) - \frac{1}{p}\sum_{i:W_i=k'} \hat{f}(X_i).$$

The pure-strategy optimal design finds single assignments W that on their own minimize these quantities.

Definition 7.3. Given subjects' baseline covariates $X \in \mathcal{X}^n$ and a magnitude function $||\cdot|| : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the *pure-strategy optimal design* chooses W uniformly at random from the set of optimizers

$$W \in \arg\min_{W\in\mathcal{W}} \left\{ M^2_{\text{p}}(W) := \max_{||f||\leq 1} \max_{k\neq k'} B^2_{kk'}(W,f) \right\}.$$

We denote by $M^2_{\text{p-opt}}$ the random variable equal to the optimal value.

The mixed-strategy optimal design directly optimizes the distribution of assignments.

Definition 7.4. Given subjects' baseline covariates $X \in \mathcal{X}^n$ and a magnitude function $||\cdot|| : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the *mixed-strategy optimal design* draws W randomly according to a distribution σ such that

$$\sigma \in \arg\min_{\sigma\in\Delta} \left\{ M^2_{\text{m}}(\sigma) := \max_{||f||\leq 1} \max_{k\neq k'} \sum_{W\in\mathcal{W}} \sigma(W) B^2_{kk'}(W,f) \right\}.$$

We denote by $M^2_{\text{m-opt}}$ the random variable equal to the optimal value.

[3]It can also be noted that this bound is of the same form as the objective in Theorem 7.1 but employing the potentially non-symmetric norm $\|\hat{Y}\| = \min_{f(X_i)=\hat{Y}_i} ||f||$ induced by the quotient of $\mathcal{F}$ over the subspace $\{f \in \mathcal{F} : f(X_i) = 0 \ \forall i\}$.

Both designs satisfy (7.1)-(7.3). The pure-strategy optimal design does due to the symmetry of the objective function (thus, if W is optimal then a treatment-permutation of W is also optimal). The mixed-strategy optimal design does by the construction of Δ. Because the pure-strategy optimal design is feasible in Δ, it is also immediate that $M^2_{\text{m-opt}} \leq M^2_{\text{p-opt}}$.

The objectives $M^2_{\text{p}}(W)$ and $M^2_{\text{m}}(\sigma)$ are the imbalance metrics that the designs seek to minimize. The two are different in nature as one expresses imbalance of a single assignment and the other the imbalance of a whole design. Since evaluation differences are linear and by assumption continuous, both $M^2_{\text{p}}(W)$ and $M^2_{\text{m}}(\sigma)$ are in fact norms taken in the continuous dual Banach space (and this guarantees they are defined). For mixed strategies, $M^2_{\text{m}}(\sigma)$ is actually determined by $n(n-1)/2$ sufficient statistics from σ.

Theorem 7.5. *Let $\sigma \in \Delta$ be given. Then*

$$M^2_m(\sigma) = M^2_m(P(\sigma)) := \max_{||f|| \leq 1} \frac{2}{pn} \sum_{i,j=1}^{n} P_{ij}(\sigma) f(X_i) f(X_j),$$

where $P_{ij}(\sigma) = \sigma(\{W_i = W_j\}) - \frac{1}{m-1}\sigma(\{W_i \neq W_j\})$.

In the case of $m = 2$, $P(\Delta) = \mathcal{P}$ is the space of feasible P matrices, which are always positive semi-definite (i.e., symmetric with nonnegative eigenvalues).

7.2.3 Structural Information and Existing Designs and Imbalance Metrics

We now show how the above framework of optimal design in fact recovers various existing designs that balance baseline covariates a priori. In this section we consider two treatments, $m = 2$.

Blocking and Complete Randomization

Randomized block designs are probably the most common non-completely-randomized designs. In a complete block design the sample is segmented into b disjoint evenly-sized blocks $\{i_{1,1}, \ldots, i_{1,2p_1}\}$, $\ldots$, $\{i_{b,1}, \ldots, i_{b,2p_b}\}$ so that baseline covariates are equal within each block and unequal between blocks, i.e., $X_{i_{\ell,j}} = X_{i_{\ell',j'}}$ if and only if $\ell = \ell'$. (If any coarsening is done, we assume it was done prior and X_i represents the coarsened value.) Then complete randomization is applied to each block separately and independently of the other blocks.

A complete block design is not always feasible, e.g. when there are subjects with a unique value of covariates or there is an otherwise odd number of subjects with a particular equal value of covariates. In an incomplete block design, there are left-over subjects $i_{0,1}, \ldots, i_{0,b'}$. One blocks subjects into evenly-sized blocks so that the number b' is as small as possible, breaking ties randomly as to which subject is left over; complete randomization is then also applied to the left-overs.[4]

[4]Incomplete block designs are much more general than this and cover a much larger scope,

Complete blocking can be thought of as minimizing a binary measure of imbalance: 0 if the sets of baseline covariates in each experimental groups are *exactly* the same, infinity otherwise. Incomplete blocking can be thought of as minimizing a discrete measure of imbalance equal to the complement of the number of exact perfect matches across experimental groups (i.e., b'). If complete blocking is feasible, then incomplete blocking necessarily recovers it. If all values of X_i are distinct, then incomplete blocking is the same as complete randomization. As it is the most general, we will only treat incomplete blocking. It turns out that incomplete blocking's exact matching metric corresponds to the space L^∞, i.e., the space $\mathcal{F}$ of bounded functions endowed with the norm $||f||_\infty = \sup_{w \in W} f(w)$.

Theorem 7.6. *Let $||f|| = ||f||_\infty$. Then the pure-strategy optimal design is equivalent to incomplete blocking.*

As noted before, this also recovers complete blocking (if it is feasible) and complete randomization (if all subjects' baseline covariates are distinct).

Pairwise Matching

In optimal pairwise matching, two treatments are considered, subjects are put into pairs so to minimize the sum of pairwise distances in their covariates, and then each pair is split randomly among the two treatments. Any pairwise distance metric δ on $\mathcal{X}$ can be chosen to define the pairwise distances $\delta(X_i, X_j)$. Usually the pairwise Mahalanobis distance is used for vector-valued covariates. The motivation behind pairwise matching is that subjects with similar covariates should have similar outcomes. This corresponds to the space of Lipschitz functions.

Theorem 7.7. *Let a distance metric δ on $\mathcal{X}$ be given. Let*

$$||f|| = ||f||_{lip} = \sup_{x \neq x'} \frac{f(x) - f(x')}{\delta(x, x')}.$$

Then the pure-strategy optimal design is equivalent to optimal pairwise matching with respect to the pairwise distance metric δ.[5]

Corollary 7.8. *Let $\delta_0 > 0$ and a distance metric δ be given. Define $\delta'(x, x') = \max\{\delta(x, x'), \delta_0\}$ for $x \neq x'$ and $\delta'(x, x) = 0$. Let $||f|| = ||f||_{lip}$ with respect to δ'. Then the pure-strategy optimal design is equivalent to caliper matching if it is feasible, i.e., choose at random from pairwise matchings that have all pairwise distances at most δ_0 after blocking exact matches.*

---

especially when treatments outnumber block size, but in our simple setup they amount to breaking ties randomly while maintaing an even partition.

[5]While $||\cdot||_{\text{lip}}$ is only a seminorm on functions (i.e., $||f||_{\text{lip}} = 0$ doesn't necessarily mean $f = 0$), in the quotient space with respect to constant functions (the kernel of this seminorm) it is a norm and it forms a Banach space. Evaluation differences are well-defined and continuous because they are bounded, $|f(X_i) - f(X_j)| \leq ||f||_{\text{lip}}\, \delta(X_i, X_j)$.

This interpretation of pairwise matching recasts its motivation as structure. Comparing with blocking we see that, whereas blocking treats any two subjects with unequal covariates as potentially having expected outcomes that are as different as any, pairwise matching presumes that unequal but similar covariates should lead to similar expected outcomes. This interpretation of pairwise matching also allows us to generalize it to $m \geq 3$ by using the same space of Lipschitz functions and employing our definition of the optimal designs for general m. We study these new designs in Section 7.4.1.

If we modify the norm and augment it with the sup-norm, we will instead recover an a priori (rather than on-the-fly) version of the method of Kapelner and Krieger (2014).

Theorem 7.9. *Let $\delta_0 > 0$ and a distance metric δ be given and let*

$$||f|| = \max\left\{||f||_{lip}, ||f||_\infty / \delta_0\right\}.$$

Then the pure-strategy optimal design is equivalent to the following: minimizes the sum of pairwise distances with respect to δ with the option of leaving a subject unmatched at a penalty of δ_0 (thus no pairs at a distance greater than $2\delta_0$ will ever be matched); then matched pairs are randomly split between the two groups and unmatched subjects are completely randomized.

Re-Randomization of Morgan et al. (2012)

The method of Morgan et al. (2012) formalizes the common, but arguably often haphazard, practice of re-randomization as a principled, theoretically-grounded a priori balancing method. The authors consider two treatments, vector-valued baseline covariates $\mathcal{X} = \mathbb{R}^d$, and an imbalance metric equal to a group-wise Mahalanobis metric

$$M^2_{\text{Re-rand}}(W) = \left(\frac{2}{n}\sum_{i=1}^{n} u_i X_i\right)^T \hat{\Sigma}^{-1} \left(\frac{2}{n}\sum_{i=1}^{n} u_i X_i\right), \tag{7.5}$$

where $\hat{\Sigma}$ is the sample covariance matrix of X. The authors reinterpret re-randomization as a heuristic algorithm that repeatedly draws random W in order to solve the constraint satisfaction problem $\exists ?W : M^2_{\text{Re-rand}} \leq t$ for a given t (they also propose a normal-approximation method for selecting t to correspond to a particular acceptance probability of a random W).

We can recover (7.5) using our framework. Let $\mathcal{F} = \text{span}\{1, x_1, \ldots, x_d\}$ and define $||f||^2 = \beta^T \hat{\Sigma} \beta + \beta_0^2$ for $f(x) = \beta_0 + \beta^T x$. Using duality of norms,

$$M^2_{\text{p}}(W) = \max_{||f|| \leq 1} B^2(W, f) = \left(\max_{\beta^T \hat{\Sigma} \beta \leq 1} \beta^T \left(\frac{n}{2}\sum_{i=1}^{n} u_i X_i\right)\right)^2 = M^2_{\text{Re-rand}}(W).$$

In Morgan et al. (2012), the authors argue that when a linear model is known to

hold, i.e.,

$$Y_{ik} = \beta_0 + \beta^T X_i + \tau \mathbb{I}[k=1] + \epsilon_i \qquad i = 1, \ldots, n,\ k = 1, 2, \tag{7.6}$$

then fixing t and re-randomizing until $M^2_{\text{Re-rand}}(W) \leq t$ yields a reduction in variance relative to complete randomization that is constant over n:

$$1 - \text{Var}(\hat{\tau}) / \text{Var}(\hat{\tau}^{\text{CR}}) = \eta(1 - \text{Var}(\epsilon_i) / \text{Var}(Y_{i1})),\ \eta \in (0,1) \text{ constant over } n.$$

For us, the imbalance metric is a direct consequence of structure ((7.6) implies $f_k \in \mathcal{F}$) and fully minimizing $M^2_{\text{p}}(W)$ leads to near-best-possible reduction in variance (see Corollary 7.15 and Section 7.3.3):

$$1 - \text{Var}(\hat{\tau}) / \text{Var}(\hat{\tau}^{\text{CR}}) \longrightarrow 1 - \text{Var}(\epsilon_i) / \text{Var}(Y_{i1}) \quad \text{at a linear rate } 2^{-\Omega(n)}.$$

It is important to keep in mind, however, that the assumption that such a finite-dimensional linear model (7.6) is valid is a parametric, and therefore fragile, assumption. Indeed, we saw in Example 7.2 that fully minimizing $M^2_{\text{Re-rand}}$ when the model is misspecified can lead to worse variance.

Other Finite-Dimensional Spaces and the Method of Bertsimas et al.

We can generalize the idea of parametric balancing methods using finite-dimensional spaces with general norms. Consider any finite-dimensional subspace of $\mathcal{V}$, $\mathcal{F} = \text{span}\{\phi_1, \ldots, \phi_r\}$, and any norm on it. Any such space is always a Banach space and evaluations are always continuous (see Theorems 5.33 and 5.35 of Hunter and Nachtergaele (2001)). An important example is the q-norm: $\|\beta_1\phi_1 + \cdots + \beta_r\phi_r\| = ||\beta||_q$ where $||\beta||_q = (\sum_i |\beta_i|^q)^{1/q}$ for $1 \leq q < \infty$ and $||\beta||_\infty = \max_i |\beta_i|$. This yields

$$M^2_{\text{p}}(W) = \left|\left| \left(\frac{n}{2} \sum_{i=1}^n u_i \phi_1(X_i), \ldots, \frac{n}{2} \sum_{i=1}^n u_i \phi_r(X_i) \right) \right|\right|^2_{q^*}$$

for $1/q + 1/q^* = 1$. Hence, the optimal design matches the sample ϕ_j moments between the groups by minimizing a norm in the vector of mismatches.

The covariance-scaled 2-norm on $\mathcal{F} = \text{span}\{1, x_1, \ldots, x_d\}$ was considered in Section 7.2.3 and gave rise to the group-wise Mahalanobis metric. Endowing $\mathcal{F} = \text{span}\{1, x_1, \ldots, x_d, x_1^2/\rho, \ldots, x_d^2/\rho, x_1x_2/(2\rho), \ldots, x_{d-1}x_d/(2\rho)\}$ with the ∞-norm and normalizing the data will recover the method of Bertsimas et al..

7.2.4 New Designs Using RKHS Structure

In our framework, one starts with structural information about the relationship between X_i and Y_{ik} and this leads to measures of imbalance and to optimal designs that minimize them. In the previous section we saw how different structures led to well-known measures of imbalance and designs. We now explore how other choices of

structure lead to new designs. We treat general $m \geq 2$ in this section.

We will express structure using reproducing kernel Hilbert spaces (RKHS). A Hilbert space is an inner-product space such that the norm induced by the inner product, $\|f\|^2 = \langle f, f \rangle$, yields a Banach space. An RKHS $\mathcal{F}$ is a Hilbert space of functions for which evaluation $f \mapsto f(x)$ is continuous for each $x \in \mathcal{X}$ (see Berlinet and Thomas-Agnan (2004)). Continuity and the Riesz representation theorem imply that for each $x \in \mathcal{X}$ there is $\mathcal{K}(x, \cdot) \in \mathcal{F}$ such that $\langle \mathcal{K}(x, \cdot), f(\cdot) \rangle = f(x)$ for every $f \in \mathcal{F}$. The symmetric map $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel of $\mathcal{F}$. The name is motivated by the fact that $\mathcal{F} = \text{closure}_{\mathcal{F}}(\text{span}\{\mathcal{K}(x, \cdot) : x \in \mathcal{X}\})$. Thus $\mathcal{K}$ fully characterizes $\mathcal{F}$. Prominent examples of kernels are:

1. The linear kernel $\mathcal{K}(x, x') = x^T x'$. This spans the finite-dimensional space of linear functions and induces a 2-norm on coefficients.

2. The polynomial kernel $\mathcal{K}_s(x, x') = (1 + x^T x'/s)^s$. It spans the finite-dimensional space of all polynomials of degree up to s.

3. Any kernel $\mathcal{K}(x, x') = \sum_{i=0}^{\infty} a_i (x^T x')^i$ with $a_i \geq 0$ (subject to convergence). This includes the previous two examples. Another case is the exponential kernel $\mathcal{K}(x, x') = e^{x^T x'}$, which can be seen as the infinite-dimensional limit of the polynomial kernel. The corresponding space is infinite-dimensional (non-parametric).

4. The Gaussian kernel $\mathcal{K}(x, x') = e^{-||x - x'||^2}$. The corresponding space is infinite-dimensional (non-parametric) and is studied in Steinwart et al. (2006).

For given $X \in \mathcal{X}^n$ and an RKHS with kernel $\mathcal{K}$, we will often use the Gram matrix $K_{ij} = \mathcal{K}(X_i, X_j)$. The Gram matrix is always positive semi-definite and as such it has a matrix square root $K = \sqrt{K}\sqrt{K}$.

As mentioned above, an RKHS induces a norm. Therefore, in our framework, it also induces imbalance metrics and optimal designs.

Theorem 7.10. *Let $\mathcal{F}$ be an RKHS with kernel $\mathcal{K}$. Then,*

$$M_p^2(W) = \frac{1}{p^2} \max_{k \neq k'} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'}), \quad \textit{and} \tag{7.7}$$

$$M_m^2(P) = \frac{2}{np} \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right). \tag{7.8}$$

Notice that (7.7) corresponds to a discrepancy statistic known as *maximum mean discrepancy* between the experimental groups. Maximum mean discrepancy is used as a test statistic in two-sample testing (see Gretton et al. (2007, 2012), Sejdinovic et al. (2013)).

The problem of minimizing (7.7) or (7.8) can be interpreted as a multi-way multi-criterion number partitioning problem. For $m = 2$, $\mathcal{X} = \mathbb{R}$, and $\mathcal{K}(x, x') = xx'$ ($K = XX^T$), we get the usual balanced number partitioning problem for both (7.7)

and (7.8): recalling our definitions of $\mathcal{U}$ and $\mathcal{P}$,

$$\frac{n}{2}M_{\text{p-opt}} = \sqrt{\min_{u\in\mathcal{U}} u^T(XX^T)u} = \min_{u\in\mathcal{U}}\left|\sum_{i=1}^n u_i X_i\right|,$$
$$\frac{n}{2}M_{\text{m-opt}} = \sqrt{\min_{P\in\mathcal{P}} \text{trace}\,(P(XX^T))} = \min_{u\in\mathcal{U}}\left|\sum_{i=1}^n u_i X_i\right|,$$

where the last equality is due to the facts that $\lambda_{\max}(M) = \text{trace}(M)$ if M is rank-1 positive semi-definite and that a linear objective on a polytope is optimized at a corner point. This reduction also shows that both problems are NP-hard (see problem [SP12] and comment on p. 223 of Garey and Johnson (1979)).

Such partitioning problems generically have unique optima up to permutation so the pure-strategy optimal design usually randomizes among the $m!$ permutations of a single partition of subjects. This is not generally the case for the mixed-strategy optimal design. Consider $m = 2$. Since the affine hull of $\mathcal{U}$ is $(n-1)$-dimensional, the mixed-strategy optimal design mixes at the very least $2(\text{rank}(K) - 1)$ assignments. Moreover, by Carathéodory's theorem any $P \in \mathcal{P}$ can be identified as the convex combination of $n(n-1)$ points in $\{uu^T : u \in \mathcal{U}\}$ (whose affine hull is $(n(n-1)-1)$-dimensional) so that the mixed-strategy objective $M_{\text{m}}^2(\sigma)$ of any a priori balancing design $\sigma \in \Delta$ can also be achieved by mixing no more than $2n(n-1)$ assignments.

In Sections 7.4.1 and 7.4.2 we will study how we solve the pure- and mixed-strategy optimal designs, respectively. For now let us consider two concrete examples with the various designs we have so far studied.

Example 7.11. Consider the following setup: we measure $d \geq 2$ baseline covariates for each subject that are uniformly distributed in the population $X_i \sim \text{Unif}\left([-1,1]^d\right)$, the two treatments $m = 2$ have constant individual effects $Y_{i1} - Y_{i2} = \tau$, and the conditional expectation of outcomes depends on two covariates only $\mathbb{E}\left[Y_{i1}\middle|X = x\right] - \tau/2 = \mathbb{E}\left[Y_{i2}\middle|X = x\right] + \tau/2 = \hat{f}(x_1, x_2)$. We consider a variety of conditional expectation functions:[6]

Linear: $\hat{f}(x_1, x_2) = x_1 - x_2$.

Quadratic: $\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2$.

Cubic: $\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2 + x_1^3 - x_2^3 - 3x_1^2x_2 + 3x_1x_2^3$.

Sinusoidal: $\hat{f}(x_1, x_2) = \sin(\frac{\pi}{3} + \frac{\pi x_1}{3} - \frac{2\pi x_2}{3}) - 6\sin(\frac{\pi x_1}{3} + \frac{\pi x_2}{4}) + 6\sin(\frac{\pi x_1}{3} + \frac{\pi x_2}{6})$.

To simulate the common situation where some covariates matter and some do not and which is which is not known a priori, we consider both the case $d = 2$ (only balance the relevant covariates) and $d = 4$ (also balance some covariates that turn out to be irrelevant).

[6] We do not consider the case of no relationship ($\hat{f}(x_1, x_2) = c$) because Theorem 7.13 proves that in this case any a priori balancing design yields the same estimation variance.

We consider the following designs: (1) complete randomization, i.e., the pure-strategy optimal design for L^∞; (2) blocking on the orthant of X_i (d two-level factors), i.e., the pure-strategy optimal design for L^∞ after coarsening; (3) re-randomization with 1% acceptance probability and Mahalanobis objective; (4) pairwise matching with Mahalanobis distance, i.e., the pure-strategy optimal design for the Lipschitz norm; (5) the pure-strategy optimal design with respect to the linear kernel; (6) the pure-strategy optimal design with respect to the quadratic kernel (polynomial kernel with $s = 2$); (7) the mixed-strategy optimal design with respect to the Gaussian kernel; and (8) the mixed-strategy optimal design with respect to the exponential kernel.[7] All of these designs result in an unbiased estimate of SATE = PATE = τ and can therefore be compared on their variance. In Figure 7-2 we plot the variances of the resulting estimators relative to $V_n = \mathrm{Var}\,(\mathrm{SATE}) + \mathrm{Var}\,(\epsilon_{11} + \epsilon_{12})\,/n$ (see Theorem 7.13).

There are several features to note. One is that when a parametric model is correctly specified and specifically optimized for, the variance (relative to V_n) shrinks linearly (inverse exponentially) – we argue this is a general phenomenon in Section 7.3.3. This phenomenon is clearest in the case of linear conditional expectation and the pure-strategy optimal design with respect to the linear kernel, but the same design does not do so well when the linear model is misspecified. The pure-strategy optimal design with respect to the quadratic kernel also has a linear, but slower, convergence for the linear conditional expectation, but it performs better in the other cases, both when a quadratic model is correctly specified and when it is not. The mixed-strategy optimal designs with respect to the Gaussian and exponential kernels seem to have uniformly good performance in all cases and in particular still exhibit what would seem to be linear convergence for the linear and quadratic cases.[8] It would seem that these non-parametric methods strike a good compromise between efficiency and robustness. Finally, we note that compared to balancing only those covariates that matter most ($d = 2$), balancing also other covariates ($d = 4$) leads to loss of efficiency, as would be expected, but the order of convergence (linear) is the same.

Example 7.12. We now consider the effect of a priori balance on a real dataset. We use the diabetes study dataset from Efron et al. (2004) described therein as follows: "Ten [$d = 10$] baseline variables [X_i], age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of [442] diabetes patients, as well as the response of interest [Y_i'], a quantitative measure of disease progression one year after baseline." We consider a hypothetical experiment where the prognostic features X_i are measured at the onset, a control or treatment is applied, and the response after one year is measured. In our hypothetical setup, the treatment reduces disease progression by exactly τ so that $Y_{i1} = Y_i'$ and $Y_{i2} = Y_i' - \tau$. Fixing n, we draw n subjects with replacement from the population of 442, normalize the covariate data so that the sample of n has zero sample mean and

[7] For the mixed-strategy designs we use the heuristic solution given by Algorithm 7.4.3.

[8] The argument in Section 7.3.3 concerns only finite-dimensional spaces and does not support this observation as a general phenomenon.

Figure 7-2: The Estimation Variance $\mathrm{Var}(\hat{\tau}) - V_n$ in Example 7.11

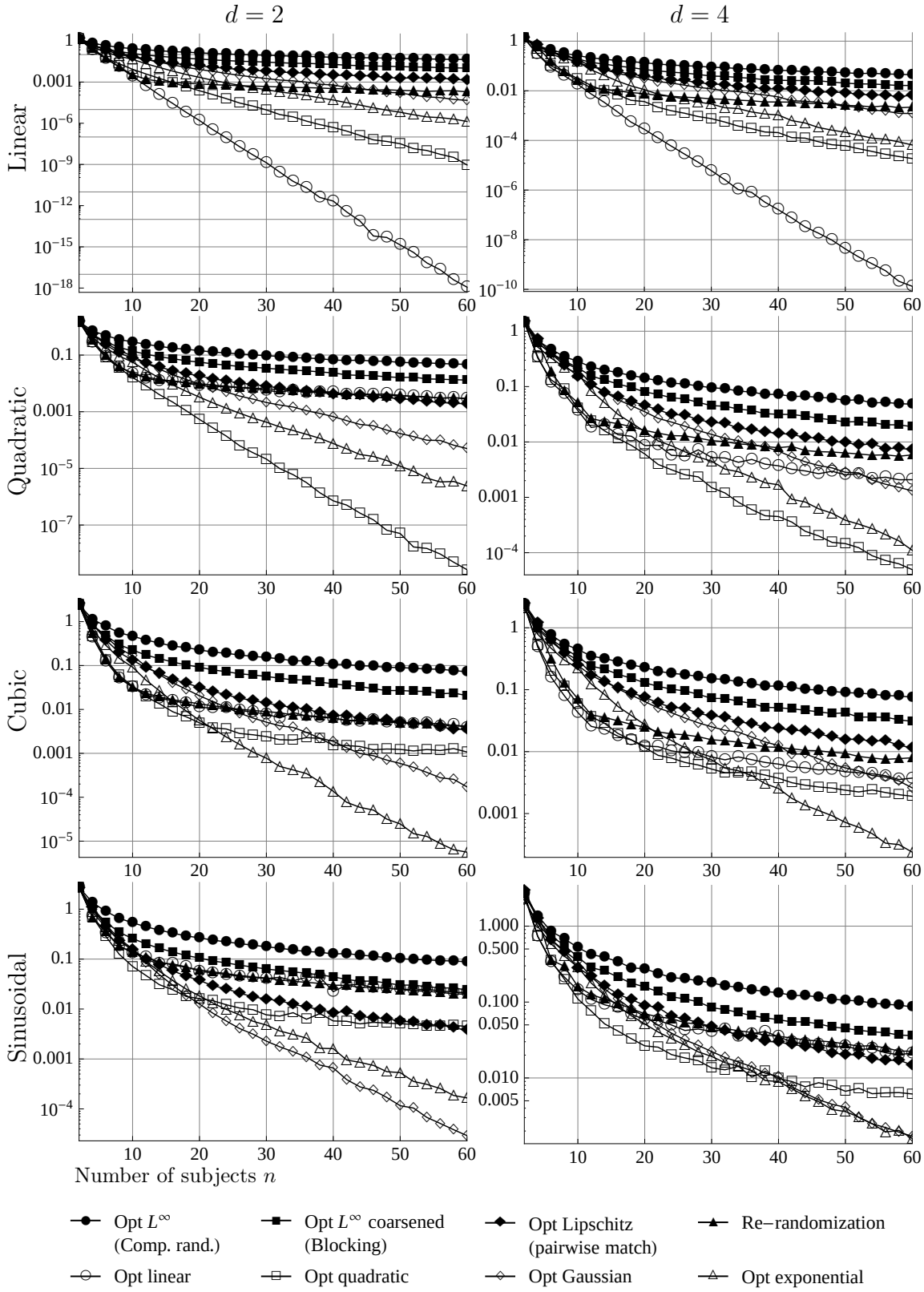
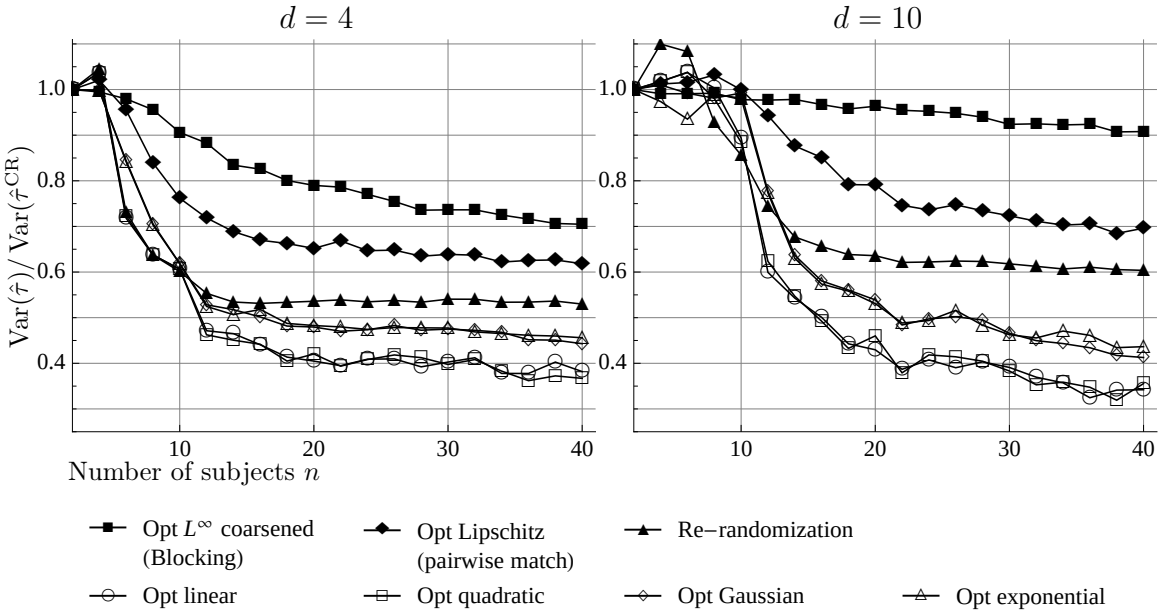


Figure 7-3: Relative Estimation Variance $\mathrm{Var}(\hat{\tau})/\mathrm{Var}(\hat{\tau}^{\mathrm{CR}})$ for the Diabetes Dataset in Example 7.12



identity sample covariance and divide by $d = 10$, apply each of the a priori balancing designs considered in Example 7.11 to the normalized covariates, and finally apply the treatments and measure the responses and the mean differences $\hat{\tau}$. Again, we consider either balancing all $d = 10$ covariates or only the $d = 4$ covariates that are ranked first by Efron et al. (2004) (these are $\{3, 9, 4, 7\}$). We plot estimation variances relative to complete randomization in Figure 7-3.

For larger n, the relative variance of each method stabilizes around a particular ratio. Each of blocking, pairwise matching, and re-randomization result in a higher ratio when attempting to balance all covariates compared to balancing only the four most important. For example, re-randomization on all 10 covariates gives $\sim$60% of complete randomization's variance whereas restricting to the important covariates yields $\sim$53%. On the other hand, the RKHS-based optimal designs yield lower relative variances for both $d = 10$ and $d = 4$, converging slower for $d = 10$ but using the small additional prognostic content of the extra covariates to reduce variance further. For example, the pure-strategy optimal designs with respect to the linear and quadratic kernels both yield $\sim$40% of complete randomization's variance for $d = 4$ and $\sim$35% for $d = 10$, taking only slightly longer to get below $\sim$40% when $d = 10$. This can be attributed to the linear rate at which the optimal designs eliminate imbalances (see Section 7.3.3). Thus, even if there are some less relevant variables, all are immediately near-perfectly balanced for modest n; the only limiting factors are the residuals (ϵ_{ik}), which, by definition, cannot be controlled for using the covariates X alone (see Corollary 7.15).

Aside: a Bayesian Interpretation

The pure-strategy optimal design can also be interpreted in a Bayesian perspective as an optimal design. The interpretation is very similar to the standard Bayesian interpretation of regularized regression using Gaussian processes (see e.g. Kimeldorf and Wahba (1970) and §6.2 of Rasmussen and Williams (2006)). Let $m = 2$ and let $\mathcal{F}$ be a given RKHS with kernel $\mathcal{K}$. Let us assume a Gaussian prior on $\hat{f}$ with covariance operator $\mathcal{K}$, i.e. $\hat{f}(x)$ is Gaussian for every $x \in \mathcal{X}$ and the covariance of $\hat{f}(x)$ and $\hat{f}(x')$ is equal to $\mathcal{K}(x, x')$. Then we have that the Bayes variance risk of a design W is

$$\begin{aligned}\mathbb{E}\left[B^2(W,f)|X,Y,W\right] &= \frac{4}{n^2}\sum_{i,j=1}^{n} u_i u_j \mathbb{E}\left[f(X_i)f(X_j)|X,Y\right] \\ &= \frac{4}{n^2}\sum_{i,j=1}^{n} u_i u_j \mathcal{K}(X_i,X_j) = \frac{4}{n^2}u^T K u = M_{\mathrm{p}}^2(W).\end{aligned}$$

Note however that randomization is not necessary from a standard Bayesian perspective (for futher discussion see Kadane and Seidenfeld (1990), Savage (1961)) and therefore a Bayesian design may not satisfy (7.1)-(7.2). In contrast, the pure- and mixed-strategy optimal designs both randomize by construction. Moreover, for the mixed-strategy optimal design, it is generally optimal to randomize beyond just random permutations of one partition.

7.3 Characterizations of A Priori Balancing Designs

We now try to characterize the estimators that arise from pure- and mixed-strategy optimal designs as well as a priori balancing designs in general. We argue the estimator is unbiased and then bound its variance in terms of a priori imbalance – a result that intimately connects imbalance prior to treatment to variance of estimation after treatment. We also discuss consistency and the convergence rate of imbalance (and hence variance).

7.3.1 Variance

We begin by decomposing the variance of any estimator arising from an a priori balancing design, that is, one satisfying (7.1)-(7.3).

Theorem 7.13. *Suppose* (7.1)-(7.3) *are satisfied. Then, for all* $k \neq k'$*,*

(a) $\hat{\tau}_{kk'}$ *is conditionally and marginally unbiased, i.e.,*

$$\mathbb{E}\left[\hat{\tau}_{kk'}\middle|X,Y\right] = \mathrm{SATE}_{kk'}, \quad \mathbb{E}\left[\hat{\tau}_{kk'}\right] = \mathrm{PATE}_{kk'}\,.$$

(b) $\hat{\tau}_{kk'} = \text{SATE}_{kk'} + D_{kk'} + E_{kk'}$,

$$\textit{where} \quad D_{kk'} := \frac{1}{m}\sum_{l\neq k} B_{kl}\left(f_k\right) - \frac{1}{m}\sum_{l\neq k'} B_{k'l}\left(f_{k'}\right),$$

$$E_{kk'} := \frac{1}{n}\sum_{i=1}^{n}\left((mw_{ik}-1)\epsilon_{ik} - (mw_{ik'}-1)\epsilon_{ik'}\right).$$

(c) $\text{SATE}_{kk'}$, $D_{kk'}$, *and* $E_{kk'}$ *are all uncorrelated so that*

$$\begin{aligned}\text{Var}\left(\hat{\tau}_{kk'}\right) = \frac{1}{n}\text{Var}\left(Y_{1k} - Y_{1k'}\right) + \text{Var}\left(D_{kk'}\right) \\ + \frac{1}{n}\text{Var}\left(\epsilon_{1k} + \epsilon_{1k'}\right) + \frac{m-2}{n}\left(\text{Var}\left(\epsilon_{1k}\right) + \text{Var}\left(\epsilon_{1k'}\right)\right).\end{aligned}$$

(Note that the last term drops when only two treatments are considered.)

Note that in part (c), every term except for $\text{Var}\left(D_{kk'}\right)$ is completely unaffected by any a priori balancing. Below we provide a bound on it based on the expected minimal imbalance produced by an optimal design.

Theorem 7.14. *If the pure- or mixed-strategy optimal design is used,*

$$\text{Var}\left(D_{kk'}\right) \leq \frac{\left(||f_k|| + ||f_{k'}||\right)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M^2_{opt}\right], \tag{7.9}$$

where $M^2_{opt} = M^2_{p\text{-}opt}$ *or* $M^2_{opt} = M^2_{m\text{-}opt}$*, respectively.*

In (7.9), $\left(||f_k|| + ||f_{k'}||\right)^2$ is unknown but constant, merely scaling the bound.

Combining the two theorems we get that when the pure- or mixed-strategy optimal design is used, the variance of our estimator is bounded as follows:

$$\begin{aligned}\text{Var}\left(\hat{\tau}_{kk'}\right) \leq \frac{1}{n}\text{Var}\left(Y_{1k} - Y_{1k'}\right) + \frac{\left(||f_k|| + ||f_{k'}||\right)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M^2_{\text{opt}}\right] \\ + \frac{1}{n}\text{Var}\left(\epsilon_{1k} + \epsilon_{1k'}\right) + \frac{m-2}{n}\left(\text{Var}\left(\epsilon_{1k}\right) + \text{Var}\left(\epsilon_{1k'}\right)\right).\end{aligned}$$

This intimately connects balance prior to treatment and randomization to estimation variance afterward. For example, for pairwise matching this explicitly connects the sum of pair differences before treatment to estimation variance after via the Lipschitz constant of the unknown regression function.

Basic arithmetic with this bound yields the following simplification.

Corollary 7.15. *Suppose* $m = 2$ *and that individual effects are constant* $Y_{i1} - Y_{i2} = \tau$*. Denote* $\sigma^2 = \text{Var}\left(Y_{i1}\right) = \text{Var}\left(Y_{i2}\right)$*,* $\xi^2 = \text{Var}\left(\epsilon_{i1}\right) = \text{Var}\left(\epsilon_{i2}\right)$*, and* $R^2 = 1 - \xi^2/\sigma^2$ *(explained variance fraction). Then, the variance due to the optimal design relative to complete randomization is bounded as follows:*

$$1 - R^2 \leq \frac{\text{Var}\left(\hat{\tau}\right)}{\text{Var}\left(\hat{\tau}^{CR}\right)} \leq 1 - R^2 - \frac{n}{16\sigma^2}\left(||f_k|| + ||f_{k'}||\right)^2\mathbb{E}\left[M^2_{opt}\right].$$

Alternatively, the relative reduction in variance is simply one minus the above. Despite the constant effect assumption, this bound provides important insights. On the one hand, it says that any a priori balancing effort can never do better than $(1 - R^2)$ relative to complete randomization. This makes sense: balancing based on X alone can only help to the extent that it is predictive of outcomes. On the other hand, it says that if $\mathbb{E}\left[M_{\text{opt}}^2\right]$ decays super-logarithmically, i.e. $o(1/n)$, then the relative variance converges to the best possible, which is $(1 - R^2)$. In Section 7.3.3 we study a case where the convergence is linear, i.e. $2^{-\Omega(n)}$, much faster than logarithmic.

When $f_k \notin \mathcal{F}$ we have $||f_k|| = \infty$ and the bound (7.9) is trivial. Accounting for the distance between f_k and $\mathcal{F}$, an alternative bound is possible.

Theorem 7.16. *If the pure- or mixed-strategy optimal design is used,*

$$\text{Var}\left(D_{kk'}\right) \leq \left(1 - \frac{1}{m}\right) \inf_{g_k, g_{k'} \in \mathcal{F}} \Big(\left(||g_k|| + ||g_{k'}||\right)^2 \mathbb{E}\left[M_{opt}^2\right] + \frac{2}{m}\left(||f_k - g_k||_2 + ||f_{k'} - g_{k'}||_2\right)^2 \Big),$$

where $M_{opt}^2 = M_{p\text{-}opt}^2$ or $M_{opt}^2 = M_{m\text{-}opt}^2$, respectively, and $||g||_2^2 = \mathbb{E}\left[g(X_1)^2\right]$ is the L^2 norm with respect to the measure of X_1. (By the assumption that potential outcomes have second moments, we have $||f_k||_2 < \infty$.)

7.3.2 Consistency

An estimator is said to be strongly consistent if it converges almost surely to the estimand, the quantity it tries to estimate. In light of Theorem 7.13(b), an a priori balancing design results in a strongly consistent estimator if and only if $D_{kk'}$ converges to 0 almost surely (since $\text{SATE}_{kk'} + E_{kk'}$ is already strongly consistent). Employing laws of large numbers in Banach spaces, we can express sufficient conditions for strong consistency in terms of a functional analytical property of $\mathcal{F}$ known as B-convexity.

Definition 7.17. A Banach space is said to be *B-convex* if there exists $N \in \mathbb{N}$ and $\eta < N$ such that for every $g_1, \ldots, g_N$ with $||g_i|| \leq 1 \ \forall i$ there exists a choice of signs so that $||\pm g_1 \pm \cdots \pm g_N|| \leq \eta$.

It is easy to verify that all the Banach spaces so far considered are B-convex with the exception of L^∞. In particular, every Hilbert space or finite-dimensional Banach space is B-convex. We use this condition to characterize consistency in the following.

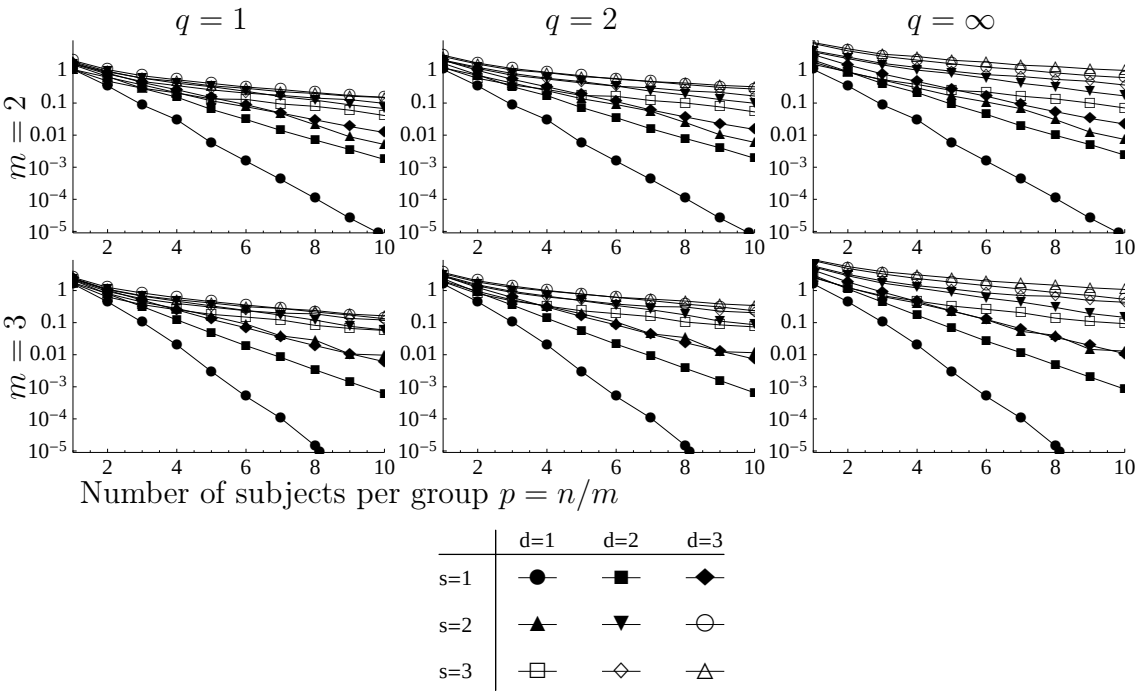
Theorem 7.18. *Suppose $f_k, f_{k'} \in \mathcal{F}$. If either*

(a) $\mathcal{F}$ is B-convex and $\mathbb{E}\left(\max_{||f|| \leq 1} \left(f(X_1) - f(X_2)\right)\right)^2 < \infty$ or

(b) $\mathcal{F}$ is a Hilbert space and $\mathbb{E}\left|\max_{||f|| \leq 1} \left(f(X_1) - f(X_2)\right)\right| < \infty$

then the estimator $\hat{\tau}_{kk'}$ arising from either the pure- or mixed-strategy optimal design is strongly consistent.

Figure 7-4: The Convergence of $\mathbb{E}M^2_{\text{p-opt}}$ as the Number of Subjects Per Group p Increases for Banach Spaces of Finite Dimension $\binom{d+s}{s}$



7.3.3 Linear Rate of Convergence for Parametric Designs

In Theorem 7.18, we argued that the estimator converges, i.e., it is consistent, but we did not discuss its rate of convergence. In this section, we study the rate of convergence of $\mathbb{E}M^2_{\text{opt}}$ for the pure- and mixed-strategy designs and hence the convergence of the corresponding estimator's variance as per Theorem 7.14. In particular, we now argue that $\mathbb{E}M^2_{\text{opt}} = 2^{-\Omega(n)}$ for the case $m = 2$ and $\mathcal{F}$ finite dimensional (i.e., parametric). We will also study $m \geq 3$ empirically and observe similar convergence.

Let $\phi_1, \ldots, \phi_r$ be a basis for the finite-dimensional $\mathcal{F}$ and $\Phi_{ij} = \phi_j(X_i)$. Because all norms in finite dimensions are equivalent, i.e., $c\,||\cdot||' \leq ||\cdot|| \leq C\,||\cdot||'$ (see Theorem 5.36 of Hunter and Nachtergaele (2001)), it follows that any rate of convergence that applies when $\mathcal{F}$ is endowed with the 2-norm ($||\beta_1\phi_1 + \cdots + \beta_r\phi_r|| = ||\beta||_2$) also applies when $\mathcal{F}$ has any given norm. Next note that since $M^2_{\text{m-opt}} \leq M^2_{\text{p-opt}}$, any rate of convergence for $M^2_{\text{p-opt}}$ applies also to $M^2_{\text{m-opt}}$. So, we restrict our attention to pure-strategy optimal designs under the 2-norm.

Our argument is a heuristic one (not a precise proof) and will follow the asymptotic approximation of the configurations W with energies $M^2_{\text{p}}(W)$ as a spin glass following the random energy model (REM) where energies are assumed independent. This approximation is commonly used to study the distributions of the optima of combinatorial optimization problems with random inputs and has been found to be valid asymptotically for partition problems similar to the one we are considering (see Mertens (2001), Borgs et al. (2009a,b)).

Let $\Sigma_{ij} = \text{Cov}\left(\phi_i(X_1), \phi_j(X_1)\right)$ and let $\lambda_1, \ldots, \lambda_{r'} > 0$ be its positive eigenvalues

where $r' = \operatorname{rank}(\Sigma)$. The distribution of $M_{\text{p}}^2(W)$ is the same for any one fixed W. Fix $W_i = (i \mod 2) + 1$ ($u_i = (-1)^{i+1}$). By the multivariate central limit theorem we have the following convergence in distribution,

$$\frac{2}{n}\Phi^T u = \left(\frac{2}{n}\sum_{i=1}^{n/2}\left(\phi_j(X_{2i-1}) - \phi_j(X_{2i})\right)\right)_{j=1}^{r} \xrightarrow{d} \mathcal{N}(0, 2\Sigma).$$

By continuous transformation, we also have

$$M_{\text{p}}^2(W) = \sup_{||\beta||_2 \le 1}\left(\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{r} u_i\beta_j\phi_j(X_i)\right)^2 = \left|\left|\frac{2}{n}\Phi^T u\right|\right|_2^2 \xrightarrow{d} \sum_{i=1}^{r'} 2\lambda_i\chi_1^2,$$

the weighted sum of independent chi-squared random variables with one degree of freedom. Denote the corresponding CDF by H and PDF by h, which are given in series representation in Kotz et al. (1967). In following with the REM approximation we assume independent energies so that $M_{\text{p-opt}}^2$ is distributed as the smallest order statistic among $\binom{n}{n/2}$-many independent draws from H. By Theorem 11.3 of Ahsanullah et al. (2013) and $\lim_{t\to 0^+} th(t)/H(t) = r'/2$, we have that

$$\mathbb{P}\left(M_{\text{p-opt}}^2/\beta_n \le t\right) \longrightarrow 1 - \exp\left(-t^{r'/2}\right)$$

for β_n satisfying $H(\beta_n)\cdot\binom{n}{n/2} \to 1$. By formula (40) of Kotz et al. (1967) this is true for

$$\beta_n = 4\left(\Gamma(r'/2+1)\Big/\binom{n}{n/2}\right)^{2/r'}\prod_{i=1}^{r'}\lambda_i^{1/r'}.$$

Thus, $\mathbb{E}M_{\text{p-opt}}^2 \approx \beta_n\Gamma(2/r'+1)$ asymptotically. By Stirling's formula,

$$\mathbb{E}M_{\text{m-opt}}^2 \le \mathbb{E}M_{\text{p-opt}}^2 = O\left(2^{-2n/r'}n^{1/r'}\right) = 2^{-\Omega(n)}.$$

We plot the convergence of $\mathbb{E}M_{\text{p-opt}}^2$ for a range of cases in Figure 7-4. We consider $m = 2, 3$, $X_i \sim \mathcal{N}(0, I_d)$, $\phi_\theta(x) = s^{1-\sum_i\theta_i}\prod_{i=1}^d x_i^{\theta_i}$, $d = 1, 2, 3$, $r = \binom{d+s}{s}$ (all monomials up to degree s) for $s = 1, 2, 3$, and q-norms 1, 2, and ∞. All exhibit linear convergence (note log scale).

7.4 Algorithms for Optimal Design

We now address how to actually realize the optimal designs, i.e., solve the optimization problems in the definitions of the pure- and mixed-strategy optimal designs. For complete randomization, blocking, and pairwise matching (with two treatments), how to do so is already clear; here we address the other designs that arouse from our framework. For the pure-strategy optimal designs, the optimization problems will

be linear, quadratic, and second-order cone optimization problems subject to integer constraints on some of the variables. Therefore, for these we can use integer optimization software to find the optimal design. In all numerical results in this chapter, we use Gurobi v5.6 Gurobi Optimization Inc. (2013). For the mixed-strategy optimal design, the problem is too hard to solve exactly and we provide heuristics based on semi-definite optimization.

7.4.1 Optimizing Pure Strategies

The pure-strategy optimization problem can be written as

$$\begin{aligned}\sqrt{\min_{W\in\mathcal{W}} M_{\mathrm{p}}^2(W)} = \min_{\lambda\in\mathbb{R},\, w\in\{0,1\}^n} \lambda \\ \text{s.t. } \lambda \geq \max_{||f||\leq 1} \frac{1}{p}\sum_{i=1}^{n}\left(w_{ik}-w_{ik'}\right) f(X_i) \quad \forall k<k' \qquad (7.10)\\ \sum_{k=1}^{m} w_{ik} = 1 \qquad \forall i=1,\ldots,n \\ \sum_{i=1}^{n} w_{ik} = p \qquad \forall k=1,\ldots,m,\end{aligned}$$

where we have used the fact that optimizing the square is the same as optimizing the absolute value and then used the symmetry of the norm to remove the absolute value and rid of excess constraints ($k > k'$). What remains is to write the constraints (7.10) in a way fitting for a linear, quadratic, or second-order cone optimization problem. We assume the solver software will arbitrarily return any one optimal solution at random. In case this is not so, we still randomly permute the result to ensure condition (7.2) holds.

Finite-Dimensional q-space

For the setup as in Section 7.2.3,

$$\begin{aligned}&\max_{||f||\leq 1} \frac{1}{p}\sum_{i=1}^{n}\left(w_{ik}-w_{ik'}\right) f(X_i) \\ &\quad = \left|\left|\left(\frac{n}{2}\sum_{i=1}^{n}\left(w_{ik}-w_{ik'}\right)\phi_1\left(X_i\right),\ldots,\frac{n}{2}\sum_{i=1}^{n}\left(w_{ik}-w_{ik'}\right)\phi_r\left(X_i\right)\right)\right|\right|_{q^*}\end{aligned}$$

for $1/q + 1/q^* = 1$. It follows that for $q = 1, \infty$, the pure-strategy optimization problem is a linear optimization problem with integer variables. For $q = 2$, the problem for $m = 2$ is a quadratic optimization problem with integer variables and for $m \geq 3$ it is a second-order cone optimization problem with integer variables (the difference being whether the quadratic term is in the objective or constraints). Rational q can also be dealt with using second-order cone optimization via the results

of Lobo et al. (1998). For example, $m = 2$, $q = 2$, and $\Phi_{ij} = \phi_j(X_i)$, leads to a binary quadratic optimization problem:

$$M^2(W) = \frac{4}{n^2} \min_{u \in \mathcal{U}} u^T \Phi \Phi^T u.$$

Lipschitz Functions

Given a pairwise distance metric δ, we define the norm $||f|| = ||f||_{\text{lip}}$. When $m = 2$, Theorem 7.7 shows that the pure-strategy optimal design is equivalent to pairwise matching. The corresponding optimization problem is weighted non-bipartite matching, which can be solved in polynomial time using Edmond's algorithm Edmonds (1965). For $m \geq 3$, we let $D_{ij} = \delta(X_i, X_j)$ and use linear optimization duality Bertsimas and Tsitsiklis (1997) to write

$$\begin{aligned}
&\lambda \geq \max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) = \max_{ve^T - ev^T \leq D} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) v_i \\
&\iff \exists S \in \mathbb{R}_+^{n \times n} \text{ s.t. } \begin{array}{l} \lambda \geq \text{trace}(DS)/p, \\ \sum_{j=1}^{n} (S_{ij} - S_{ji}) = w_{ik} - w_{ik'} \quad \forall i = 1, \ldots, n, \end{array}
\end{aligned}$$

yielding a linear optimization problem with integer variables.

For the modification $||f|| = \max\left\{||f||_{\text{lip}}, ||f||_\infty / \delta_0\right\}$ considered in Theorem 7.9, we can instead write

$$\begin{aligned}
&\lambda \geq \max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) = \max_{ve^T - ev^T \leq D, \, ||v||_\infty \leq \delta_0} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) v_i \\
&\iff \exists \begin{array}{l} S \in \mathbb{R}_+^{n \times n} \\ t \in \mathbb{R}^n \end{array} \text{ s.t. } \begin{array}{l} \lambda \geq (\text{trace}(DS) + \delta_0 ||t||_1)/p, \\ \sum_{j=1}^{n} (S_{ij} - S_{ji}) + t_i = w_{ik} - w_{ik'} \quad \forall i = 1, \ldots, n. \end{array}
\end{aligned}$$

This also leads to a linear optimization problem with integer variables.

RKHS

As in Theorem 7.10 we have

$$\left(\max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) \right)^2 = \frac{1}{p} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'}).$$

Therefore, for $m = 2$ the pure-strategy optimization problem is a quadratic optimization problem with integer variables and for $m \geq 3$ it is a second-order cone optimization problem with integer variables. Namely, for $m = 2$, we get the binary quadratic optimization problem:

$$M^2(W) = \frac{4}{n^2} \min_{u \in \mathcal{U}} u^T K u.$$

7.4.2 Optimizing Mixed Strategies

For the case of mixed strategies we only consider the case of $m = 2$ and $\mathcal{F}$ being an RKHS. As per Theorems 7.5 and 7.10, the corresponding optimization problem is

$$\frac{4}{n^2} \min_{P \in \mathcal{P}} \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right).$$

From the proof of Theorem 7.1 it can be gathered that if $\sigma \in \Delta$ then,

$$\max_{Y \in \mathbb{R}^{n \times 2}} \frac{\text{Var}\left(\hat{\tau}|X, Y\right)}{\text{Var}\left(\hat{\tau}^{CR}|X, Y\right)} = \left(1 - \frac{1}{n}\right) \lambda_{\max}\left(P(\sigma)\right).$$

Therefore, if we wish, we may ensure that we do not stray too far from complete randomization in the worst realization of outcomes by instead solving

$$\begin{aligned} \frac{4}{n^2} \min_{P \in \mathcal{P}} \quad & \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right) \\ \text{s.t.} \quad & \left(1 - \frac{1}{n}\right) \lambda_{\max}(P) \leq \rho. \end{aligned} \tag{7.11}$$

Since setting $\rho = \infty$ eliminates the constraint, we will only treat (7.11) as it is most general. Setting $\rho = 1$ forces (7.11) to choose complete randomization.

While the problem (7.11) has a convex objective and convex feasible region, we have already observed in Section 7.2.4 that the problem is NP-hard. When $\rho = \infty$, the feasible region is $\mathcal{P}$, which is a polytope. But what makes (7.11) with $\rho = \infty$ more difficult than the problem encountered in Section 7.4.1 is that, at the same time as being NP-hard, it is not amenable to the branch-and-bound techniques employed by integer optimization software because its optimum generally does not occur at a corner point of the polytope, as we observed in Section 7.2.4. The polytope $\mathcal{P}$ is known as the equipartition polytope of the complete graph on n vertices Conforti et al. (1990a,b).

Therefore, we propose only heuristic solutions to the problem. These heuristics are based on semi-definite optimization (SDO), i.e., optimization over the cone S_+^n of $n \times n$ positive semi-definite matrices (see Boyd and Vandenberghe (2004) for more information on SDO). In particular, the heuristics run in polynomial time. We use Mosek (Mosek, APS (2009)) to solve all SDO problems in our numerical experiments.

The first heuristic is based on a semidefinite outer approximation $\mathcal{P} \subset \left\{ P \in S_+^n : \text{diag}(P) = e,\, Pe = 0 \right\}$ and is motivated by Goemans and Williamson (1995) and Bertsimas and Ye (1999).

Algorithm 7.4.1. Let $\hat{P}$ be a solution to the SDO

$$\begin{aligned} \min_{\lambda\in\mathbb{R},\, P\in S_+^n} \quad & \lambda \\ \text{s.t.} \quad & \lambda I - \sqrt{K}P\sqrt{K} \in S_+^n \\ & \rho I - \left(1 - \frac{1}{n}\right) P \in S_+^n \\ & \operatorname{diag}(P) = e,\ Pe = 0. \end{aligned}$$

Let $\hat{\sigma}$ be the distribution of $u_i = \operatorname{sign}(v_i - \operatorname{median}(v))$ where $v \sim \mathcal{N}(0, \hat{P})$. (This provides a sampling mechanism without needing to fully specify $\hat{\sigma}$).

The second heuristic is based on an inner approximation of $\mathcal{P}$.

Algorithm 7.4.2. Given $u_1, \ldots, u_T \in \mathcal{U}$, let $\hat{\theta}$ be the solution to the SDO

$$\begin{aligned} \min_{\lambda\in\mathbb{R},\ \theta\in\mathbb{R}^T} \quad & \lambda \\ \text{s.t.} \quad & \lambda I - \sum_{t=1}^{T} \theta_t \sqrt{K} u_t u_t^T \sqrt{K} \in S_+^n \\ & \rho I - \left(1 - \frac{1}{n}\right) \sum_{t=1}^{T} \theta_t u_t u_t^T \in S_+^n \\ & \theta \geq 0,\ \sum_{t=1}^{T} \theta_t = 1. \end{aligned}$$

Let $\hat{\sigma}$ be the distribution of $u = \pm u'$ equiprobably where u' is drawn randomly from $\{u_t\}$ according to weights $\hat{\theta}$.

The inputs to Algorithm 7.4.2 can be generated in two ways. One way is to run Algorithm 7.4.1 and use the solution to draw u_t (filtering non-unique values up to negation). Another way is to use as inputs the top T solutions to the pure-strategy problem. As this is the method we use in our numerical experiments we describe it explicitly below.

Algorithm 7.4.3. Let $\mathcal{U}_1 = \mathcal{U} \cap \{u_1 = 1\}$. For $t = 1, \ldots, T$ do:

1: Solve $u_t \in \arg\min_{u\in\mathcal{U}_t} u^T K u$.

2: Set $\mathcal{U}_{t+1} = \mathcal{U}_t \cap \left\{u_t^T u \leq n - 4\right\}$.

Run Algorithm 7.4.2 using $u_1, \ldots, u_T$.

The definition of $\mathcal{U}_1$ simply eliminates the symmetry of negation. Each further refinement in step 2 cuts away the last optimal solution.

7.5 Algorithms for Inference

A priori balance has the potential to significantly reduce estimation variance. One would expect therefore that inferences on the treatment effect can also have higher

statistical power. In this section, we will consider $m = 2$ and the sharp null hypothesis

$$H_0 : (\text{TE}_i = 0 \; \forall i = 1, \ldots, n) \, .$$

Under H_0 all post-treatment responses are exchangeable regardless of treatment given ($Y_{i1} = Y_{i2}$). We can therefore simulate what would happen under another assignment and compare. This is the idea behind Fisher's randomization test, where new simulated assignments are drawn from the same design as used at the onset of the experiment. However, the pure-strategy optimal design when $\mathcal{F}$ is an RKHS generally only randomizes over treatment-permutations of a single partition, which does not provide enough comparison (applying Fisher's randomization test will always yield $p = 1$). Therefore, we develop an alternative test based on the bootstrap Efron and Tibshirani (1993):

Algorithm 7.5.1. For a confidence level $0 < 1 - \alpha < 1$:

1: Draw W^0 from the pure-strategy optimal design for the baseline covariates $X_1, \ldots, X_n$, assign subjects, apply treatments, measure outcomes $Y_{iW_i^0}$, and compute $\hat{\tau}$.

2: For $t = 1, \ldots, T$ do:

2.1: Sample $i_j^t \sim \text{Unif}\{1, \ldots, n\}$ independently for $j = 1, \ldots, n$.

2.2: Draw W^t from the pure-strategy optimal design for the baseline covariates $X_{i_1^t}, \ldots, X_{i_n^t}$.

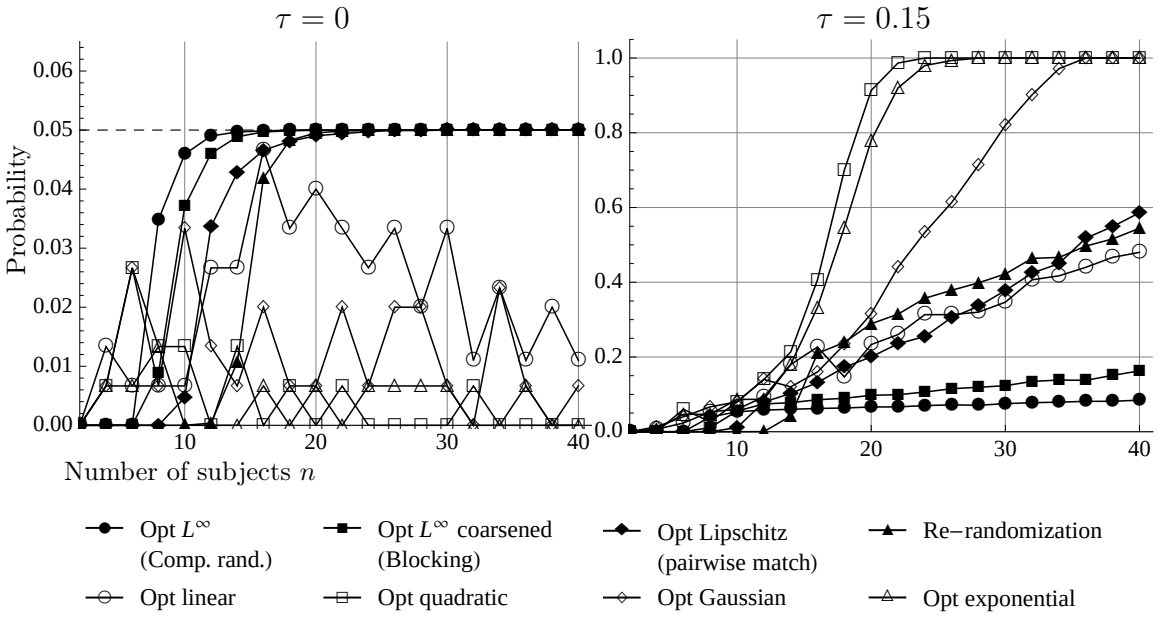
2.3: Compute $\tilde{\tau}^t = \frac{1}{p} \sum_{i:W_i^t=1} Y_{iW_i^0} - \frac{1}{p} \sum_{i:W_i^t=2} Y_{iW_i^0}$.
(Notice we only use the outcomes we chose to observe in step 1.)

3: The p-value of H_0 is $p = (1 + |\{t \; : \; |\tilde{\tau}^t| \geq |\hat{\tau}|\}|) \, / \, (1 + T)$.
If $p \leq \alpha$, then reject H_0.

Algorithm 7.5.1 can also be used to answer inferential questions for mixed-strategy designs, letting W^t be drawn from the corresponding mixed-strategy optimal design σ^t in step 2.2. However, the additional randomization of mixed-strategy optimal designs (and of complete randomization, blocking, pairwise matching, and re-randomization for that matter) allows one to use the standard randomization and exact permutation tests instead (where new assignments are drawn from the same design as used at the onset of the experiment). As these tests are standard we defer further discussion to supplemental Section E.3. We next consider an example using Algorithm 7.5.1.

Example 7.19. Consider the setup as in Example 7.11 with $d = 2$, quadratic $\hat{f}$, and $\epsilon_{i1} = \epsilon_{i2} = 0$. For various values of τ, we test H_0 at significance $\alpha = 0.05$ for each of the designs in Example 7.11 (replacing the mixed-strategy optimal designs with corresponding pure-strategy optimal designs) using Algorithm 7.5.1 for all RKHS-based optimal designs and the standard randomization test for all other designs (see Algorithm E.3.2 in supplemental Section E.3). We plot in Figure 7-5 the probability of rejecting H_0 as n grows.

Figure 7-5: Probability of Rejecting H_0 Under No Effect $\tau = 0$ and a Positive Effect $\tau = 0.15$ at $\alpha = 5\%$ as in Example 7.19



When τ is positive, the quadratic and exponential RKHS-based designs detect the difference in treatments almost immediately, the Gaussian a bit later. The linear RKHS-based design parametrically misspecifies the regression function in this particular case but does not do much worse than the other designs nonetheless. Interestingly, as imbalance disappears, Algorithm 7.5.1 has much lower type I error than the significance $\alpha = 0.05$.

7.6 Conclusions

Designs that provide balance in controlled experiments before treatments are applied and before randomization provide one answer to the criticism that complete randomization may lead to assignments that the experimenter knows will lead to misleading conclusions. In this chapter we unified these designs under the umbrella of a priori balance. We argued that structural information on the dependence of outcomes on baseline covariates was the key to any a priori balance beyond complete randomization and developed a framework of optimal designs based on structure expressed on the conditional expectation function. We have shown how existing a priori balancing designs, including blocking, pairwise matching, and other designs, are optimal for certain structures and how existing imbalance metrics, such as the group-wise Mahalanobis metric of Morgan et al. (2012), arise from other choices of structure. That this theoretical framework fit so well into existing practice, led us to endeavor to discover what other designs may arise from it. We considered a wide range of designs that follow from structure expressed using RKHS, encompassing both parametric and

non-parametric methods. We argued and shown numerically that parametric models (when correctly specified) coupled with optimization lead to estimation variance that converges very fast to the best theoretically possible.

It has not escaped my notice that this unified perspective on a priori balance suggests a possible rephrasing of Box's maxim: "*balance* what you can, randomize what you cannot."

Part IV

Appendices

Appendix A

Appendix to Chapter 2

A.1 Asymptotic Optimality for Mixing Processes and Proofs

In this supplemental section, we generalize the asymptotic results to mixing process and provide the omitted proofs.

A.1.1 Mixing Processes

We begin by defining stationary and mixing processes.

Definition A.1. A sequence of random variables $V_1, V_2, \ldots$ is called *stationary* if joint distributions of finitely many consecutive variables are invariant to shifting. That is,

$$\mu_{V_t,\ldots,V_{t+k}} = \mu_{V_s,\ldots,V_{s+k}} \quad \forall s, t \in \mathbb{N},\, k \geq 0.$$

In particular, if a sequence is stationary then the variables have identical marginal distributions, but they may not be independent and the sequence may not be exchangeable. Instead of independence, mixing is the property that if standing at particular point in the sequence we look far enough ahead, the head and the tail look nearly independent, where "nearly" is defined by different metrics for different definitions of mixing.

Definition A.2. Given a stationary sequence $\{V_t\}_{t\in\mathbb{N}}$, denote by $\mathcal{A}^t = \sigma(V_1, \ldots, V_t)$ the sigma-algebra generated by the first t variables and by $\mathcal{A}_t = \sigma(V_t, V_{t+1}, \ldots)$ the sigma-algebra generated by the subsequence starting at t. Define the *mixing coefficients at lag k*

$$\begin{aligned}
\alpha(k) &= \sup_{t\in\mathbb{N},\, A\in\mathcal{A}^t,\, B\in\mathcal{A}_{t+k}} |\mu(A\cap B) - \mu(A)\mu(B)| \\
\beta(k) &= \sup_{t\in\mathbb{N}} \left|\left|\mu_{\{V_s\}_{s\leq t}} \otimes \mu_{\{V_s\}_{s\geq t+k}} - \mu_{\{V_s\}_{s\leq t \vee s\geq t+k}}\right|\right|_{\mathrm{TV}} \\
\rho(k) &= \sup_{t\in\mathbb{N},\, Q\in L_2(\mathcal{A}^t),\, R\in L_2(\mathcal{A}_{t+k})} |\mathrm{Corr}(Q,R)|
\end{aligned}$$

where $||\cdot||_{\text{TV}}$ is the total variance and $L_2(\mathcal{A})$ is the set of $\mathcal{A}$-measurable square-integrable real-valued random variables.

$\{V_t\}$ is said to be α-mixing if $\alpha(k) \overset{k\to\infty}{\longrightarrow} 0$, β-mixing if $\beta(k) \overset{k\to\infty}{\longrightarrow} 0$, and ρ-mixing if $\rho(k) \overset{k\to\infty}{\longrightarrow} 0$.

Notice that an iid sequence has $\alpha(k) = \beta(k) = \rho(k) = 0$. Bradley (1986) establishes that $2\alpha(k) \leq \beta(k)$ and $4\alpha(k) \leq \rho(k)$ so that either β- or ρ-mixing implies α-mixing.

Many processes satisfy mixing conditions under mild assumptions: auto-egressive moving-average (ARMA) processes (cf. Mokkadem (1988)), generalized autoregressive conditional heteroskedasticity (GARCH) processes (cf. Carrasco and Chen (2002)), and certain Markov chains. For a thorough discussion and more examples see Doukhan (1994) and Bradley (2005). Mixing rates are often given explicitly by model parameters but they can also be estimated from data (cf. Mcdonald et al. (2011)). Sampling from such processes models many real-life sampling situations where observations are taken from an evolving system such as, for example, the stock market, inter-dependent product demands, or aggregates of doubly stochastic arrival processes as in the posts on social media.

A.1.2 Asymptotic Optimality

Let us now restate the results of Section 2.4.2 in more general terms, encompassing both iid and mixing conditions on S_N. We will also establish that our cost estimates converge, i.e.,

$$\min_{z\in\mathcal{Z}} \sum_{i=1}^{N} w_{N,i}(x)c(z;y^i) \to v^*(x), \tag{A.1}$$

for μ_X-almost-everywhere $x \in X$ (henceforth, μ_X-a.e.x) almost surely (henceforth, a.s.).

Theorem A.3 (kNN)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold and that S_N is generated by iid sampling. Let $w_{N,i}(x)$ be as in* (2.12) *with $k = \min\left\{\lceil CN^\delta\rceil, N-1\right\}$ for some $C > 0$, $0 < \delta < 1$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal and* (A.1) *holds.*

Theorem A.4 (Kernel Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold and that $\mathbb{E}\left[|c(z;Y)|\max\left\{\log|c(z;Y)|, 0\right\}\right] < \infty$ for each z. Let $w_{N,i}(x)$ be as in* (2.13) *with K being any of the kernels in* (2.14) *and $h = CN^{-\delta}$ for $C, \delta > 0$. Let $\hat{z}_N(x)$ be as in* (2.3)*. If S_N comes from*

1. *an iid process and $\delta < 1/d_x$, or*
2. *a ρ-mixing process with $\rho(k) = O(k^{-\gamma})$ ($\gamma > 0$) and $\delta < 2\gamma/(d_x + 2d_x\gamma)$, or*
3. *an α-mixing process with $\alpha(k) = O(k^{-\gamma})$ ($\gamma > 1$) and $\delta < 2(\gamma-1)/(3d_x+2d_x\gamma)$,*

then $\hat{z}_N(x)$ is asymptotically optimal and (A.1) *holds.*

Theorem A.5 (Recursive Kernel Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold and that S_N comes from a ρ-mixing process with $\sum_{k=1}^{\infty} \rho(k) < \infty$ (or iid). Let $w_{N,i}(x)$ be as in* (2.15) *with K being the naïve kernel and with $h_i = Ci^{-\delta}$ for some $C > 0$, $0 < \delta < 1/(2d_x)$. Let $\hat{z}_N(x)$ be as in* (2.3)*. Then $\hat{z}_N(x)$ is asymptotically optimal and* (A.1) *holds.*

Theorem A.6 (Local Linear Methods)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold, that μ_X is absolutely continuous and has density bounded away from 0 and ∞ on the support of X, and that costs are bounded over y for each z (i.e., $|c(z;y)| \leq g(z)$). Let $w_{N,i}(x)$ be as in* (2.16) *with K being any of the kernels in* (2.14) *and with $h_N = CN^{-\delta}$ for some $C, \delta > 0$. Let $\hat{z}_N(x)$ be as in* (2.3)*. If S_N comes from*

1. *an iid process and $\delta < 1/d_x$, or*
2. *an α-mixing process with $\alpha(k) = O(k^{-\gamma})$, $\gamma > d_x + 3$, and $\delta < (\gamma - d_x - 3)/(d_x(\gamma - d_x + 3))$,*

then $\hat{z}_N(x)$ is asymptotically optimal and (A.1) *holds.*

A.1.3 Proofs of Asymptotic Results for Local Predictive Prescriptions

First, we establish some preliminary results. In what follows, let

$$\begin{aligned} C(z|x) &= \mathbb{E}\left[c(z;Y)\middle|X = x\right], \\ \widehat{C}_N(z|x) &= \sum_{i=1}^{N} w_{N,i}(x)c(z;y^i), \\ \mu_{Y|x}(A) &= \mathbb{E}\left[\mathbb{I}\left[Y \in A\right]\middle|X = x\right], \\ \hat{\mu}_{Y|x,N}(A) &= \sum_{i=1}^{N} w_{N,i}(x)\mathbb{I}\left[x_i \in A\right]. \end{aligned}$$

Lemma A.7. *If $\{(x^i, y^i)\}_{i\in\mathbb{N}}$ is stationary and $f : \mathbb{R}^{m_Y} \to \mathbb{R}$ is measurable then $\{(x^i, f(y^i))\}_{i\in\mathbb{N}}$ is also stationary and has mixing coefficients no larger than those of $\{(x^i, y^i)\}_{i\in\mathbb{N}}$.*

Proof. This is simply because a transform can only make the generated sigma-algebra coarser. For a single time point, if f is measurable and $B \in \mathcal{B}(\mathbb{R})$ then by definition $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ and, therefore, $\{Y^{-1}(f^{-1}(B)) : B \in \mathcal{B}(\mathbb{R})\} \subset \{Y^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^{m_Y})\}$. Here the transform is applied independently across time so the result holds ($f \times \cdots \times f$ remains measurable). □

Lemma A.8. *Suppose Assumptions 2.3 and 2.4 hold. Fix $x \in \mathcal{X}$ and a sample path of data such that, for every $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to C(z|x)$. Then $\widehat{C}_N(z|x) \to C(z|x)$ uniformly in z over any compact subset of $\mathcal{Z}$.*

Proof. Let any convergent sequence $z_N \to z$ and $\epsilon > 0$ be given. By equicontinuity and $z_N \to z$, $\exists N_1$ such that $|c(z_N; y) - c(z; y)| \leq \epsilon/2 \ \forall N \geq N_1$. Then $\left|\widehat{C}_N(z_N|x) - \widehat{C}_N(z|x)\right| \leq \mathbb{E}_{\hat{\mu}_{Y|x,N}} |c(z_N; y) - c(z; y)| \leq \epsilon/2 \ \forall N \geq N_1$. By assumption $\widehat{C}_N(z|x) \to C(z|x)$ and hence $\exists N_2$ such that $\left|\widehat{C}_N(z|x) - C(z|x)\right| \leq \epsilon/2$. Therefore, for $N \geq \max\{N_1, N_2\}$,

$$\left|\widehat{C}_N(z_N|x) - C(z|x)\right| \leq \left|\widehat{C}_N(z_N|x) - \widehat{C}_N(z|x)\right| + \left|\widehat{C}_N(z|x) - C(z|x)\right| \leq \epsilon.$$

Hence $\widehat{C}_N(z_N|x) \to C(z|x)$ for any convergent sequence $z_N \to z$.

Now fix $E \subset \mathcal{Z}$ compact and suppose that $\sup_{z \in E} \left|\widehat{C}_N(z|x) - C(z|x)\right| \not\to 0$ for contradiction. Then $\exists \epsilon > 0$ and $z_N \in E$ such that $\left|\widehat{C}_N(z_N|x) - C(z_N|x)\right| \geq \epsilon$ infinitely often. Restricting first to a subsequence where this always happens and then using the compactness of E, there exists a convergent subsequence $z_{N_k} \to z \in E$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x) - C(z_{N_k}|x)\right| \geq \epsilon$ for every k. Then,

$$0 < \epsilon \leq \left|\widehat{C}_{N_k}(z_{N_k}|x) - C(z_{N_k}|x)\right| \leq \left|\widehat{C}_{N_k}(z_{N_k}|x) - C(z|x)\right| + \left|C(z|x) - C(z_{N_k}|x)\right|.$$

Since $z_{N_k} \to z$, we have shown before that $\exists k_1$ such that $\left|\widehat{C}_{N_k}(z_{N_k}|x) - C(z|x)\right| \leq \epsilon/2$ $\forall k \geq k_1$. By equicontinuity and $z_{N_k} \to z$, $\exists k_2$ such that $|c(z_{N_k}; y) - c(z; y)| \leq \epsilon/4$ $\forall k \geq k_2$. Hence, also $|C(z|x) - C(z_{N_k}|x)| \leq \mathbb{E}\left[|c(z_{N_k}; y) - c(z; y)| \,\middle|\, X = x\right] \leq \epsilon/4$ $\forall k \geq k_2$. Considering $k = \max\{k_1, k_2\}$ we get the contradiction that $0 < \epsilon \leq \epsilon/2$. □

Lemma A.9. *Suppose Assumptions 2.3, 2.4, and 2.5 hold. Fix $x \in \mathcal{X}$ and a sample path of data such that $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly and, for every $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to C(z|x)$. Then $\lim_{N\to\infty} \left(\min_{z \in \mathcal{Z}} \widehat{C}_N(z|x)\right) = v^*(x)$ and every sequence $z_N \in \arg\min_{z \in \mathcal{Z}} \widehat{C}_N(z|x)$ satisfies $\lim_{N\to\infty} C(z_N|x) = v^*(x)$ and all of its limit points are contained in $\arg\min_{z \in \mathcal{Z}} C(z|x)$.*

Proof. Suppose that case 1 or 3 of Assumption 2.5 holds (i.e. boundedness or infinite limit). First, we show $\widehat{C}_N(z|x)$ and $C(z|x)$ are continuous and eventually coercive. Let $\epsilon > 0$ be given. By equicontinuity, $\exists \delta > 0$ such that $|c(z; y) - c(z'; y)| \ \forall y \in \mathcal{Y}$ whenever $||z - z'|| \leq \delta$. Hence, whenever $||z - z'|| \leq \delta$, we have $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \leq \mathbb{E}_{\hat{\mu}_{Y|x,N}} |c(z; y) - c(z'; y)| \leq \epsilon$ and $|C(z|x) - C(z'|x)| \leq \mathbb{E}\left[|c(z; y) - c(z'; y)| \,\middle|\, X = x\right] \leq \epsilon$. This gives continuity. Coerciveness is trivial if $\mathcal{Z}$ is bounded. Suppose it is not. Without loss of generality D_x is compact, otherwise we can take any compact subset of it that has positive probability on it. Then by assumption of weak convergence $\exists N_0$ such that $\hat{\mu}_{Y|x,N}(D_x) \geq \mu_{Y|x}(D_x)/2 > 0$ for all $N \geq N_0$. Now let $z_k \in \mathcal{Z}$ be any sequence such that $||z_k|| \to \infty$. Let $M > 0$ be given. Let $\lambda' = \liminf_{||k|| \to \infty} \inf_{y \notin D_x} c(z_k; y)$ and $\lambda = \max\{\lambda', 0\}$. By assumption $\lambda' > -\infty$. Hence $\exists k_0$ such that $\inf_{y \notin D_x} c(z_k; y) \geq \lambda' \ \forall k \geq k_0$. By D_x-uniform coerciveness and $||z_k|| \to \infty$, $\exists k_1 \geq k_0$ such that $c(z_k; y) \geq (2M - 2\lambda)/\mu_{Y|x}(D_x) \ \forall k \geq k_1$ and $y \in D_x$.

Hence, $\forall k \geq k_1$ and $N \geq N_0$,

$$C(z|x) \geq \mu_{Y|x}(D) \times (2M - 2\lambda)/\mu_{Y|x}(D_x) + (1 - \mu_{Y|x}(D))\lambda' \geq 2M - 2\lambda + \lambda \geq M,$$
$$\widehat{C}_N(z|x) \geq \hat{\mu}_{Y|x,N}(D) \times (2M - 2\lambda)/\hat{\mu}_{Y|x,N}(D_x) + (1 - \hat{\mu}_{Y|x,N}(D))\lambda' \geq M,$$

since $\alpha\lambda' \geq \lambda$ if $\alpha \geq 0$. This gives coerciveness eventually. By the usual extreme value theorem (c.f. Bertsekas (1999), pg. 669), $\widehat{\mathcal{Z}}_N(x) = \arg\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x)$ and $\mathcal{Z}^*(x) = \arg\min_{z\in\mathcal{Z}} C(z|x)$ exist, are nonempty, and are compact.

Now we show there exists $\mathcal{Z}^*_\infty(x)$ compact such that $\mathcal{Z}^*(x) \subset \mathcal{Z}^*_\infty(x)$ and $\widehat{\mathcal{Z}}_N(x) \subset \mathcal{Z}^*_\infty(x)$ eventually. If $\mathcal{Z}$ is bounded this is trivial. So suppose otherwise (and again, without loss of generality D_x is compact). Fix any $z^* \in \mathcal{Z}^*(x)$. Then by Lemma A.8 we have $\widehat{C}_N(z^*|x) \to C(z^*|x)$. Since $\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \leq \widehat{C}_N(z^*|x)$, we have $\limsup_{N\to\infty} \min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) \leq C(z^*|x) = \min_{z\in\mathcal{Z}} C(z|x) = v^*$. Now suppose for contradiction no such $\mathcal{Z}^*_\infty(x)$ exists. Then there must be a subsequence $z_{N_k} \in \widehat{\mathcal{Z}}_{N_k}$ such that $||z_{N_k}|| \to \infty$. By D_x-uniform coerciveness and $||z_{N_k}|| \to \infty$, $\exists k_1 \geq k_0$ such that $c(z_{N_k}; y) \geq 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x)$ $\forall k \geq k_1$ and $y \in D_x$. Hence, $\forall k \geq k_1$ and $N \geq N_0$,

$$\widehat{C}_N(z_{N_k}|x) \geq \hat{\mu}_{Y|x,N}(D) \times 2\left(v^* + 1 - \lambda\right)/\mu_{Y|x}(D_x) + (1 - \hat{\mu}_{Y|x,N}(D)) \geq v^* + 1.$$

This yields a contradiction $v^* + 1 \leq v^*$. So $\mathcal{Z}^*_\infty(x)$ exists.

Applying Lemma A.8,

$$\tau_N = \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| \to 0.$$

The first result follows from

$$\delta_N = \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \leq \sup_{z\in\mathcal{Z}^*_\infty(x)} \left|\widehat{C}_N(z|x) - C(z|x)\right| = \tau_N \to 0.$$

Now consider any sequence $z_N \in \widehat{\mathcal{Z}}_N(x)$. The second result follows from

$$\begin{aligned}\left|C\left(\hat{z}_N|x\right) - \min_{z\in\mathcal{Z}} C(z|x)\right| &\leq \left|\widehat{C}_N(\hat{z}_N(x)|x) - C\left(\hat{z}_N|x\right)\right| + \left|\min_{z\in\mathcal{Z}} \widehat{C}_N(z|x) - \min_{z\in\mathcal{Z}} C(z|x)\right| \\ &\leq \tau_N + \delta_N \to 0.\end{aligned}$$

Since $\widehat{\mathcal{Z}}_N(x) \subset \mathcal{Z}^*_\infty(x)$ and $\mathcal{Z}^*_\infty(x)$ is compact, z_N has at least one convergence subsequence. Let $z_{N_k} \to z$ be any convergent subsequence. Suppose for contradiction that $z \notin \mathcal{Z}^*(x)$, i.e., $\epsilon = C(z|x) - v^* > 0$. Since $z_{N_k} \to z$ and by equicontinuinty, $\exists k_2$ such that $|c(z_{N_k}; y) - c(z; y)| \leq \epsilon/2$ $\forall y \in \mathcal{Y}$ $\forall k \geq k_2$. Then, $|C(z_{N_k}|x) - C(z|x)| \leq \mathbb{E}\left[|c(z_{N_k}; y) - c(z; y)| \,\middle|\, X = x\right] \leq \epsilon/4$ $\forall k \geq k_2$. In addition, there exists k_3 such that $\delta_{N_k} \leq \epsilon/4$ $\forall k \geq k_3$. Then, $\forall k \geq \max\{k_2, k_3\}$, we have

$$\min_{z\in\mathcal{Z}} \widehat{C}_{N_k}(z|x) = \widehat{C}_{N_k}(z_{N_k}|x) \geq C(z_{N_k}|x) - \epsilon/4 \geq C(z|x) - \epsilon/2 \geq v^* + \epsilon/2.$$

Taking limits, we derive a contradiction, yielding the third result.

Now suppose that case 2 of Assumption 2.5 holds (i.e. convexity). By Lemma A.8, $\widehat{C}_N(z|x) \to C(z|x)$ uniformly in z over any compact subset of $\mathcal{Z}$. By Theorem 6.2 of Rockafellar (1997), closed convex $\mathcal{Z}$ has a non-empty relative interior. Let us restrict to its affine hull where it has a non-empty interior. We have already shown that Assumption 2.4 implies that $\widehat{C}_N(z|x)$ and $C(z|x)$ are continuous. Hence they are lower semi-continuous. Therefore, by Theorem 7.17 of Rockafellar and Wets (1998), $\widehat{C}_N(z|x)$ epi-converges to $C(z|x)$ on $\mathcal{Z}$. Consider any $z^* \in \mathcal{Z}^*(x) \neq \varnothing$. Then clearly $\min_{z \in \{z^*\}} \widehat{C}_N(z|x) = \widehat{C}_N(z^*|x) \to C(z^*|x) = \min_{z \in \mathcal{Z}} C(z|x)$ and $\{z^*\}$ is compact. By Theorem 7.31 of Rockafellar and Wets (1998) we have precisely the results desired. □

Lemma A.10. *Suppose $c(z;y)$ is equicontinuous in z. Suppose moreover that for each fixed $z \in \mathcal{Z} \subset \mathbb{R}^d$ we have that $\widehat{C}_N(z|x) \to C(z|x)$ a.s. for μ_X-a.e.x and that for each fixed measurable $D \subset \mathcal{Y}$ we have that $\hat{\mu}_{Y|x,N}(D) \to \mu_{Y|x}(D)$ a.s. for μ_X-a.e.x. Then, a.s. for μ_X-a.e.x, $\widehat{C}_N(z|x) \to C(z|x)$ for all $z \in \mathcal{Z}$ and $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly.*

Proof. Since Euclidean space is separable, $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ weakly a.s. for μ_X-a.e.x (c.f. Theorem 11.4.1 of Dudley (2002)).

Consider the set $\mathcal{Z}' = \mathcal{Z} \cap \mathbb{Q}^d \cup \{\text{the isolated points of } \mathcal{Z}\}$. Then $\mathcal{Z}'$ is countable and dense in $\mathcal{Z}$. Since $\mathcal{Z}'$ is countable, by continuity of measure, a.s. for μ_X-a.e.x, $\widehat{C}_N(z'|x) \to C(z'|x)$ for all $z' \in \mathcal{Z}'$. Restrict to a sample path and x where this event occurs. Consider any $z \in \mathcal{Z}$ and $\epsilon > 0$. By equicontinuity $\exists \delta > 0$ such that $|c(z;y) - c(z';y)| \leq \epsilon/2$ whenever $||z - z'|| \leq \delta$. By density there exists such $z' \in \mathcal{Z}'$. Then, $\left|\widehat{C}_N(z|x) - \widehat{C}_N(z'|x)\right| \leq \mathbb{E}_{\hat{\mu}_{Y|x,N}}\left[|c(z;y) - c(z';y)|\right] \leq \epsilon/2$ and $|C(z|x) - C(z'|x)| \leq \mathbb{E}\left[|c(z;y) - c(z';y)| \,\middle|\, X = x\right] \leq \epsilon/2$. Therefore, $0 \leq \left|\widehat{C}_N(z|x) - C(z|x)\right| \leq \left|\widehat{C}_N(z'|x) - C(z'|x)\right| + \epsilon \to \epsilon$. Since true for each ϵ, the result follows for all $z \in \mathcal{Z}$. The choice of particular sample path and x constitute a measure-1 event by assumption. □

Now, we prove the general form of the asymptotic results from Section A.1.2.

Proof of Theorem A.3. Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 2.3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 5 of Walk (2010) to Y'. By iid sampling and choice of k, we have that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0,1]$. Therefore applying Theorem 5 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma A.10 we obtain that assumptions for Lemma A.9 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem A.4. Fix $z \in \mathcal{Z}$. Set $Y' = c(z;y)$. By Assumption 2.3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 3 of Walk (2010) to Y'. By assumption in theorem statement, we also have that $\mathbb{E}\left\{|Y'| \max\left\{\log|Y'|, 0\right\}\right\} < \infty$. Moreover each of the kernels in (2.14) can be rewritten $K(x) = H(||x||)$ such that $H(0) > 0$ and $\lim_{t \to \infty} t^{d_X} H(t) \to 0$.

Consider the case of iid sampling. Then our data on (X, Y') is ρ-mixing with $\rho(k) = 0$. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Consider the case of ρ-mixing or α-mixing. By Lemma A.7, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choices of kernel and h_N, Theorem 3 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and $\mathbb{E}\{|Y'| \max\{\log|Y'|, 0\}\} \leq 1 < \infty$. Therefore applying Theorem 3 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma A.10 we obtain that assumptions for Lemma A.9 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem A.5. Fix $z \in \mathcal{Z}$. Set $Y' = c(z; y)$. By Assumption 2.3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 4 of Walk (2010) to Y'. Note that the naïve kernel satisfies the necessary conditions.

Since our data on (X, Y) is ρ-mixing by assumption, we have that by Lemma A.7, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Using these conditions and our choice of the naïve kernel and h_N, Theorem 4 of Walk (2010) gives that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ for μ_X-a.e.x, a.s.

Now fix D measurable. Set $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability. Therefore applying Theorem 4 of Walk (2010) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma A.10 we obtain that assumptions for Lemma A.9 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem A.6. Fix $z \in \mathcal{Z}$ and $x \in \mathcal{X}$. Set $Y' = c(z; Y)$. By Assumption 2.3, $\mathbb{E}[|Y'|] < \infty$. Let us apply Theorem 11 of Hansen (2008) to Y' and use the notation thereof. Fix the neighborhood of consideration to the point x (i.e., set $c_N = 0$) since uniformity in x is not of interest. All of the kernels in (2.14) are bounded above and square integrable and therefore satisfy Assumption 1 of Hansen (2008). Let f be the density of X. By assumption $0 < \delta \leq f(x) \leq B_0 < \infty$ for all $x \in \mathcal{X}$. Moreover, our choice of h_N satisfies $h_N \to 0$.

Consider first the iid case. Then we have $\alpha(k) = 0 = O(k^{-\gamma})$ for $\gamma = \infty$ (β in Hansen (2008)). Combined with boundedness conditions of Y' and f ($|Y'| \leq g(z) < \infty$ and $\delta < f < B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $\gamma = \infty$, $s = \infty$ in (17) of Hansen (2008) we get $\theta = 1$. Therefore, since $h = O(N^{-1/d_x})$ we have

$$\frac{(\log\log N)^4 (\log N)^2}{N^\theta h_N^{d_x}} \to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ a.s.

Now consider the α-mixing case. If the mixing conditions hold for X, Y then by Lemma A.7, equal or lower mixing coefficients hold for X, Y'. By letting $s = \infty$ we have $\gamma > d_x + 3 > 2$. Combined with boundedness conditions of Y' and f

($|Y'| \leq g(z) < \infty$ and $\delta < f < B_0$), we satisfy Assumption 2 of Hansen (2008). Setting $q = \infty$, $s = \infty$ in (16) and (17) of Hansen (2008) we get $\theta = \frac{\gamma - d_x - 3}{\gamma - d_x + 3}$. Therefore, since $h_N = O(N^{-\theta/d_x})$ we have

$$\frac{(\log\log N)^4 (\log N)^2}{N^\theta h_N^{d_x}} \to 0.$$

Having satisfied all the conditions of Theorem 11 of Hansen (2008), we have again that $\widehat{C}_N(z|x) \to \mathbb{E}[Y'|X = x]$ a.s.

Since $x \in \mathcal{X}$ was arbitrary we have convergence for μ_X-a.e.x a.s.

Now fix D measurable. Consider a response variable $Y' = \mathbb{I}[y \in D]$. Then $\mathbb{E}[Y']$ exists by measurability and Y' is bounded in $[0, 1]$. In addition, by Lemma A.7, equal or lower mixing coefficients hold for X, Y' as hold for X, Y. Therefore applying Theorem 11 of Hansen (2008) in the same manner again, $\hat{\mu}_{Y|x,N}(D)$ converges to $\mu_{Y|x}(D)$ for μ_X-a.e.x a.s.

Applying Lemma A.10 we obtain that assumptions for Lemma A.9 hold for μ_X-a.e.x, a.s., which in turn yields the result desired. □

Proof of Theorem 2.10. By assumption of Y and V sharing no atoms, $\delta \overset{a.s.}{=} \tilde{\delta} = \mathbb{I}[Y \leq V]$ is observable so let us replace δ^i by $\tilde{\delta}^i$ in (2.21). Let

$$\begin{aligned}
F(y|x) &= \mathbb{E}\left[\mathbb{I}[Y > y] \,\middle|\, X = x\right] \\
\hat{F}_N(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y\right] w_{N,i}^{\text{Kaplan-Meier}}(x), \\
H_1(y|x) &= \mathbb{E}\left[\mathbb{I}\left[U > y, \tilde{\delta} = 1\right] \,\middle|\, X = x\right] \\
\hat{H}_{1,N}(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y,\, \tilde{\delta}^i = 1\right] w_{N,i}(x), \\
H_2(y|x) &= \mathbb{E}\left[\mathbb{I}[U > y] \,\middle|\, X = x\right] \\
\hat{H}_{2,N}(y|x) &= \sum_{i=1}^N \mathbb{I}\left[u^i > y\right] w_{N,i}(x).
\end{aligned}$$

By assumption on conditional supports of Y and V, $\sup\{y : F(y : x) > 0\} \leq \sup\{y : H_2(y : x) > 0\}$. By the same arguments as in Theorem 2.6, 2.7, 2.8, or 2.9, we have that, for all y, $\hat{H}_{1,N}(y|x) \to H_1(y|x)$, $\hat{H}_{2,N}(y|x) \to H_2(y|x)$ a.s. for μ_X-a.e.x. By assumption of conditional independence and by the main result of Beran (1981), we have that, for all y, $F_N(y|x) \to F(y|x)$ a.s. for μ_X-a.e.x. Since $\mathcal{Y}$ is a separable space we can bring the "for all y" inside the statement, i.e., we have weak convergence (c.f. Theorem 11.4.1 of Dudley (2002)): $\hat{\mu}_{Y|x,N} \to \mu_{Y|x}$ a.s. for μ_X-a.e.x where $\hat{\mu}_{Y|x,N}$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Since costs are bounded, the portmanteau lemma (see Theorem 2.1 of Billingsley (1999)) gives that for each $z \in \mathcal{Z}$, $\widehat{C}_N(z|x) \to \mathbb{E}[c(z; Y)|X = x]$ where $\widehat{C}_N(z|x)$ is based on weights $w_{N,i}^{\text{Kaplan-Meier}}(x)$. Applying Lemma A.10 we obtain that assumptions for Lemma A.9 hold for μ_X-a.e.x,

a.s., which in turn yields the result desired. □

A.2 Out-of-Sample Guarantees for Mixing Processes and Proofs

First we establish a comparison lemma that is an extension of Theorem 4.12 of Ledoux and Talagrand (1991) to our multivariate case.

Lemma A.11. *Suppose that c is L-Lipschitz uniformly over y with respect to ∞-norm:*

$$\sup_{z \neq z' \in \mathcal{Z},\, y \in \mathcal{Y}} \frac{c(z;y) - c(z';y)}{\max_{k=1,\ldots,d} |z_k - z'_k|} \leq L < \infty.$$

Let $\mathcal{G} = \{(x,y) \mapsto c\left(f(x);y\right) \,:\, f \in \mathcal{F}\}$. Then we have that $\widehat{\mathfrak{R}}_n(\mathcal{G}; S_N) \leq L\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N^x)$ and therefore also that $\mathfrak{R}_n(\mathcal{G}) \leq L\mathfrak{R}_n(\mathcal{F})$. (Notice that one is a univariate complexity and one multivariate and that the complexity of $\mathcal{F}$ involves only the sampling of x.)

Proof. Write $\phi_i(z) = c(z; y^i)/L$. Then by Lipschitz assumption and by part 2 of Proposition 2.2.1 from Bertsekas et al. (2003), for each i, ϕ_i is 1-Lipchitz. We now would like to show the inequality in

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\mathcal{G}; S_N) &= L\mathbb{E}\left[\frac{2}{n}\sup_{z\in\mathcal{F}}\sum_{i=1}^{n}\sigma_{i0}\phi_i(z(x^i)) \;\middle|\; S_N\right] \\
&\leq L\mathbb{E}\left[\frac{2}{n}\sup_{z\in\mathcal{F}}\sum_{i=1}^{n}\sum_{k=1}^{d}\sigma_{ik}z_k(x^i) \;\middle|\; S_N^x\right] \\
&= L\widehat{\mathfrak{R}}_n(\mathcal{F}; S_N^x).
\end{aligned}$$

By conditioning and iterating, it suffices to show that for any $T \subset \mathbb{R} \times \mathcal{Z}$ and 1-Lipchitz ϕ,

$$\mathbb{E}\left[\sup_{t,z\in T}\left(t + \sigma_0\phi(z)\right)\right] \leq \mathbb{E}\left[\sup_{t,z\in T}\left(t + \sum_{k=1}^{d}\sigma_k z_k\right)\right]. \tag{A.2}$$

The expectation on the left-hand-side is over two values ($\sigma_0 = \pm 1$) so there are two choices of (t,z), one for each scenario. Let any $(t^{(+1)}, z^{(+1)}), (t^{(-1)}, z^{(-1)}) \in T$ be given. Let k^* and $s^* = \pm 1$ be such that

$$\max_{k=1,\ldots,d}\left|z_k^{(+1)} - z_k^{(-1)}\right| = s^*\left(z_{k^*}^{(+1)} - z_{k^*}^{(-1)}\right).$$

Fix $(\tilde{t}^{(\pm 1)}, \tilde{z}^{(\pm 1)}) = (t^{(\pm s^*)}, z^{(\pm s^*)})$. Then, since these are feasible choices in the inner supremum, choosing $(t,z)(\sigma) = (\tilde{t}^{(\sigma_{k^*})}, \tilde{z}^{(\sigma_{k^*})})$, we see that the right-hand-side of (A.2)

has

$$\begin{aligned}\text{RHS (A.2)} &\geq \frac{1}{2}\mathbb{E}\left[\tilde{t}^{(+1)} + \tilde{z}_{k^*}^{(+1)} + \sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(+1)}\right] \\ &\quad + \frac{1}{2}\mathbb{E}\left[\tilde{t}^{(-1)} - \tilde{z}_{k^*}^{(-1)} + \sum_{k\neq k^*}\sigma_k\tilde{z}_k^{(-1)}\right] \\ &= \frac{1}{2}\left(t^{(+1)} + t^{(-1)} + \max_{k=1,\ldots,d}\left|z_k^{(+1)} - z_k^{(-1)}\right|\right) \\ &\geq \frac{1}{2}\left(t^{(+1)} + \phi\left(z^{(+1)}\right)\right) + \frac{1}{2}\left(t^{(-1)} - \phi\left(z^{(-1)}\right)\right)\end{aligned}$$

where the last inequality is due to the Lipschitz condition. Since true for any $(t^{(\pm 1)}, z^{(\pm 1)})$ given, taking suprema over the left-hand-side completes the proof. □

Next, the theorem below is a combination and restatement of the main results of Bartlett and Mendelson (2003) (for iid) and Mohri and Rostamizadeh (2008) (for mixing) about univariate Rademacher complexities. These are mostly direct result of McDiarmid's inequality.

Theorem A.12. *Consider a class $\mathcal{G}$ of functions $\mathcal{U} \to \mathbb{R}$ that are bounded: $|g(u)| \leq \overline{g}$ $\forall g \in \mathcal{G}$, $u \in \mathcal{U}$. Consider a sample $S_n = (u^1, \ldots, u^N)$ of some random variable $T \in \mathcal{T}$. Fix $\delta > 0$. If S_N is generated by IID sampling, let $\delta' = \delta'' = \delta$ and $\nu = N$. If S_N comes from a β-mixing process, fix some t, ν such that $2t\nu = N$, let $\delta' = \delta/2 - (\nu - 1)\beta(t)$ and $\delta'' = \delta/2 - 2(\nu - 1)\beta(t)$. Then (only for $\delta' > 0$ or $\delta'' > 0$ where they appear), we have that with probability $1 - \delta$,*

$$\mathbb{E}\left[g(T)\right] \leq \frac{1}{N}\sum_{i=1}^{N} g(u^i) + \overline{g}\sqrt{\log(1/\delta')/2\nu} + \mathfrak{R}_\nu(\mathcal{G}) \qquad \forall g \in \mathcal{G}, \tag{A.3}$$

and that, again, with probability $1 - \delta$,

$$\mathbb{E}\left[g(T)\right] \leq \frac{1}{N}\sum_{i=1}^{N} g(u^i) + 3\overline{g}\sqrt{\log(2/\delta')/2\nu} + \widehat{\mathfrak{R}}_\nu(\mathcal{G}) \qquad \forall g \in \mathcal{G}. \tag{A.4}$$

Now, we can prove Theorem 2.13 and extend it to the case of data generated by a mixing process.

Proof of Theorem 2.13. Apply Theorem A.12 to the random variable $U = (X, Y)$ and function class $\mathcal{G} = \{(x, y) \mapsto c\left(f(x); y\right) : f \in \mathcal{F}\}$. Note that by assumption we have boundedness of functions in $\mathcal{G}$ by the constant $\overline{c}$. Bound the complexity of $\mathcal{G}$ by that of $\mathcal{F}$ using Lemma A.11 and the assumption of $c(z; y)$ being L-Lipschitz. Equations (A.3) and (A.4) hold for every $g \in \mathcal{G}$ and hence for every $f \in \mathcal{F}$ and $g(x, y) = c\left(f(x); y\right)$, of which the expectation is the expected costs of the decision rule f. □

Next, we prove our bounds on the complexities of the function classes we consider.

Proof of Lemma 2.14. Consider $\mathcal{F}_k = \{z_k(\cdot) : z \in \mathcal{F}\} = \{z_k(x) = w^T x : ||w||_p \leq \frac{R}{\gamma_k}\}$, the projection of $\mathcal{F}$ onto the k^{th} coordinate. Then $\mathcal{F} \subset \mathcal{F}_1 \times \cdots \times \mathcal{F}_{d_z}$ and $\mathfrak{R}_N(\mathcal{F}) \leq \sum_{k=1}^{d_z} \mathfrak{R}_N(\mathcal{F}_k)$. The latter right-hand-side complexities are the common univariate Rademacher complexities. Applying Theorem 1 of Kakade et al. (2008) to each component we get $\mathfrak{R}_N(\mathcal{F}_k) \leq 2M\sqrt{\frac{p-1}{N}}\frac{R}{\gamma_k}$. □

Proof of Lemma 2.15. Let q be p's conjugate exponent ($1/p + 1/q = 1$). In terms of vector norms on $v \in \mathbb{R}^d$, if $q \geq 2$ then $||v||_p \leq ||v||_2$ and if $q \leq 2$ then $||b||_p \leq d^{1/2-1/p}\,||v||_2$. Let F be the matrix $F_{ji} = x_j^i$. Note that $F\sigma \in \mathbb{R}^{d_x \times d_z}$. By Jensen's inequality and since Schatten norms are vector norms on singular values,

$$\begin{aligned}
\widehat{\mathfrak{R}}_N^2(\mathcal{F}; S_N^x) &\leq \frac{4}{N^2}\mathbb{E}\left[\sup_{||W||_p \leq R} \text{Trace}\,(WF\sigma)^2 \,\middle|\, S_N^x\right] \\
&= \frac{4R^2}{N^2}\mathbb{E}\left[||F\sigma||_q^2 \middle| S_N^x\right] \\
&\leq \frac{4R^2}{N^2}\max\left\{\min\{d_z, d_x\}^{1-2/p}, 1\right\}\mathbb{E}\left[||F\sigma||_2^2 \middle| S_N^x\right] \\
&\leq \frac{4R^2}{N^2}\max\left\{d_z^{1-2/p}, 1\right\}\mathbb{E}\left[||F\sigma||_2^2 \middle| S_N^x\right].
\end{aligned}$$

The first result follows because

$$\begin{aligned}
\frac{1}{N}\mathbb{E}\left[||F\sigma||_2^2 \,\middle|\, S_N^x\right] &= \frac{1}{n}\sum_{k=1}^{d_z}\sum_{j=1}^{d_x}\sum_{i,i'=1}^{N} x_j^i x_j^{i'}\mathbb{E}\left[\sigma_{ik}\sigma_{i'k}\right] \\
&= \frac{d_z}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_x}(x_j^i)^2 = d_z\,\widehat{\mathbb{E}}_N\,||x||_2^2
\end{aligned}$$

The second result follows by applying Jensen's inequality again to pass the expectation over S_n into the square. □

A.3 Proofs of Tractability Results

Proof of Theorem 2.1. Let $I = \{i : w_{N,i}(x) > 0\}$, $w = (w_{N,i}(x))_{i \in I}$. Rewrite (2.3) as $\min w^T\theta$ over $(z, \theta) \in \mathbb{R}^{d \times n_0}$ subject to $z \in \mathcal{Z}$ and $\theta_i \geq c(z; y^i)\ \forall i \in I$. Weak optimization of a linear objective over a closed convex body is reducible to weak separation via the ellipsoid algorithm (see Grotschel et al. (1993)). A weak separation oracle for $\mathcal{Z}$ is assumed given. To separate over the i^{th} cost constraint at fixed z', θ_i' call the evaluation oracle to check violation and if violated call the subgradient oracle to get $s \in \partial_z c(z'; y^i)$ with $||s||_\infty \leq 1$ and produce the cut $\theta_i \geq c(z'; y^i) + s^T(z - z')$. □

Proof of Theorem 2.11. In the case of (2.22), $z(x^i) = Wx^i$. By computing the norm of W we have a trivial weak membership algorithm for the norm constraint and hence by Theorems 4.3.2 and 4.4.4 of Grotschel et al. (1993) we have a weak separation algorithm. By adding affine constraints $\zeta_{ij} = z_j(x^i)$, all that is left is to separate over constraints of the form $\theta_i \geq c(\zeta_i; y^i)$, which can be done as in the proof of Theorem 2.1. □

A.4 Omitted Details from Section 2.1.1

A.4.1 Portfolio Allocation Example

In our portfolio allocation example, we consider constructing a portfolio with $d_y = d_z = 12$ securities. We simulate the observation of $d_x = 3$ market factors X that, instead of iid, evolve as a 3-dimensioanl ARMA(2,2) process:

$$X(t) - \Phi_1 X(t-1) - \Phi_2 X(t-2) = U(t) + \Theta_1 U(t-1) + \Theta_2 U(t-2)$$

where $U \sim \mathcal{N}(0, \Sigma_U)$ are innovations and

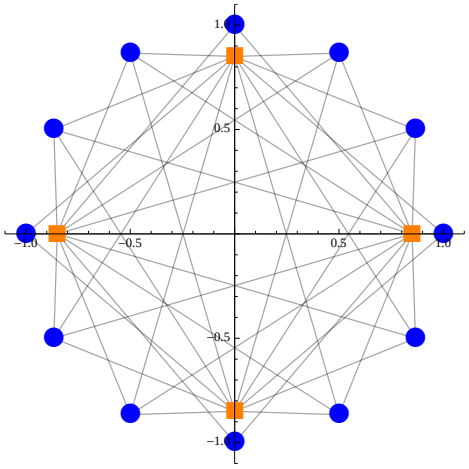
$$(\Sigma_U)_{ij} = \left(\mathbb{I}\left[i = j\right] \frac{8}{7} - (-1)^{i+j} \frac{1}{7} \right) 0.05,$$

$$\Phi_1 = \begin{pmatrix} 0.5 & -0.9 & 0 \\ 1.1 & -0.7 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} 0. & -0.5 & 0 \\ -0.5 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\Theta_1 = \begin{pmatrix} 0.4 & 0.8 & 0 \\ -1.1 & -0.3 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Theta_2 = \begin{pmatrix} 0 & -0.8 & 0 \\ -1.1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

We suppose the returns are generated according to a factor model $Y_i = A_i^T \left(X + \delta_i/4\right) + \left(B_i^T X\right) \epsilon_i$, where A_i is the mean-dependence of the i^{th} security on these factors with some idiosyncratic noise, B_i the variance-dependence, and ϵ_i and δ_i are independent

Figure A-1: Network Data for the Shipment Planning Example



(a) The network of warehouses (orange) and locations (blue)

$$D^T = \begin{pmatrix} 0.15 & 1.3124 & 1.85 & 1.3124 \\ 0.50026 & 0.93408 & 1.7874 & 1.6039 \\ 0.93408 & 0.50026 & 1.6039 & 1.7874 \\ 1.3124 & 0.15 & 1.3124 & 1.85 \\ 1.6039 & 0.50026 & 0.93408 & 1.7874 \\ 1.7874 & 0.93408 & 0.50026 & 1.6039 \\ 1.85 & 1.3124 & 0.15 & 1.3124 \\ 1.7874 & 1.6039 & 0.50026 & 0.93408 \\ 1.6039 & 1.7874 & 0.93408 & 0.50026 \\ 1.3124 & 1.85 & 1.3124 & 0.15 \\ 0.93408 & 1.7874 & 1.6039 & 0.50026 \\ 0.50026 & 1.6039 & 1.7874 & 0.93408 \end{pmatrix}$$

(b) The distance matrix

standard Gaussian idiosyncratic contributions. For A and B we use

$$A = 2.5\% \times \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \quad B = 7.5\% \times \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Marginally, all of the returns have mean 0% and standard deviations 20∼30%.

In our objective, we use $\lambda = 0$ and $\epsilon = 0.15$, i.e., we minimize the conditional value at risk at level 15%.

A.4.2 Shipment Planning Example

In our shipment planning example, we consider stocking $d_z = 4$ warehouses to serve $d_y = 12$ locations. We take locations spaced evenly on the 2-dimensional unit circle and warehouses spaced evenly on the circle of radius 0.85. The resulting network and its associated distance matrix are shown in Figure A-1. We suppose shipping costs from warehouse i to location j are $c_{ij} = \$10 D_{ij}$ and that production costs are \$5 per

unit when done in advance and \$100 per unit when done last minute.

We consider observing $d_x = 3$ demand-predictive features X. We simulate X in the same manner as in the portfolio allocation example. We simulate demands as

$$Y_i = 100 \max\{0,\ A_i^T \left(X + \delta_i/4\right) + \left(B_i^T X\right) \epsilon_i\}$$

with A, B, δ, ϵ as in the portfolio allocation example.

Appendix B

Appendix to Chapter 3

B.1 Omitted proofs

Proof of Theorem 3.16. Assumption 3.8 gives $R(p) = \mathbb{E}\left[\mathbb{E}\left[r(P)D\middle|P=p,X\right]\right]$, i.e. profit is given by taking a partial mean with $P=p$ fixed of the regression of $r(P)D$ on P and X. Partial means of kernel estimators is studied in Newey (1994). Assumptions 3.13, 3.14, and 3.15 imply the assumptions of Theorem 4.1 of Newey (1994), with a trivial constant trimming function. Applying the Theorem for each fixed $p \in \mathcal{P}$, we get

$$\sqrt{nh_n}(R(p) - \overline{R}_n(p)) \xrightarrow{d} \mathcal{N}\left(0,\, \eta_p\right) \;\forall p \in \mathcal{P},$$

where η_p is a paraphrasing of the asymptotic variance derived therein.

Optimizing partial means of kernel estimators is studied in Flores (2005). Assumptions 3.13, 3.14, and 3.15 imply the assumptions of Theorem 3 of Flores (2005). Applying the Theorem, we get

$$\sqrt{nh_n^3}(p^* - \overline{p}_n) \xrightarrow{d} \mathcal{N}\left(0,\, \frac{\eta'}{R''(p^*)^2}\right), \tag{B.1}$$

paraphrasing the asymptotic variance.

By Assumption 3.14, $R(p)$ is twice continuously differentiable. Using Taylor's theorem to expand $R(p)$ around $p = p^*$, there exists $p_n \in [\min(p^*, \overline{p}_n),\, \max(p^*, \overline{p}_n)]$ such that

$$R(\overline{p}_n) = R(p^*) + R'(p^*)(\overline{p}_n - p^*) + \frac{1}{2}R''(p_n)(\overline{p}_n - p^*)^2.$$

By first order optimality conditions, $R'(p^*) = 0$. Hence, rearranging, we have

$$R(p^*) - R(\overline{p}_n) = -\frac{1}{2}R''(p_n)(\overline{p}_n - p^*)^2. \tag{B.2}$$

By continuous transformation of eq. (B.1), we have

$$\left(nh_n^3\right)(\overline{p}_n - p^*)^2 \xrightarrow{d} \frac{\eta'}{R''(p^*)^2}\chi_1^2. \tag{B.3}$$

Eq. (B.1) also implies $\overline{p}_n \xrightarrow{\mathbb{P}} p^*$, which also implies $p_n \xrightarrow{\mathbb{P}} p^*$ since p_n is sandwiched between $\overline{p}_n$ and p^*. Since $R''(p)$ is continuous, we also get by continuous transformation that

$$R''(p_n) \xrightarrow{\mathbb{P}} R''(p^*). \tag{B.4}$$

Combining eqs. (B.2)-(B.4), we get the desired result,

$$\left(nh_n^3\right)\left(R(p^*) - R(\overline{p}_n)\right) \xrightarrow{d} \frac{-\eta'}{2R''(p^*)}\chi_1^2.$$

If $nh_n^{2s+1} \to 0$, then Assumptions 3.13, 3.14, and 3.15 also imply the assumptions of Theorem 4 of Flores (2005) (with equal bandwidths). Applying the Theorem, we get

$$\sqrt{nh_n}(R(p^*) - \overline{R}_n(\overline{p}_n)) \xrightarrow{d} \mathcal{N}\left(0,\, \eta\right),$$

paraphrasing the asymptotic variance. □

Proof of Theorem 3.17. We begin by showing that $D(p) \perp\!\!\!\perp P\big|\phi(p, X)$. On the one hand we have

$$\begin{aligned}
f_{P|\phi(p,X)}(p|q) &= \mathbb{E}\left[\delta(P-p)\middle|\phi(p,X) = q\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta(P-p)\middle|\phi(p,X) = q, X\right]\middle|\phi(p,X) = q\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta(P-p)\middle|X\right]\middle|\phi(p,X) = q\right] \\
&= \mathbb{E}\left[\phi(p,X)\middle|\phi(p,X) = q\right] \\
&= q.
\end{aligned}$$

On the other hand, using weak ignorability, we have

$$\begin{aligned}
f_{P|\phi(p,X),D(p)}(p|q,d) &= \mathbb{E}\left[\delta(P-p)\middle|\phi(p,X) = q, D(p) = d\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta(P-p)\middle|\phi(p,X) = q, D(p) = d, X\right]\middle|\phi(p,X) = q, D(p) = d\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta(P-p)\middle|D(p) = d, X\right]\middle|\phi(p,X) = q, D(p) = d\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\delta(P-p)\middle|X\right]\middle|\phi(p,X) = q, D(p) = d\right] \\
&= \mathbb{E}\left[\phi(p,X)\middle|\phi(p,X) = q, D(p) = d\right] \\
&= q.
\end{aligned}$$

Equality between the two conditional probabilities implies the desired independence.

Using this independence and then plugging in $P = p$, we have

$$\mathbb{E}\left[D(p)\middle|\phi(p,X) = q\right] = \mathbb{E}\left[D(p)\middle|P = p, \phi(p,X)\right] = \mathbb{E}\left[D\middle|P = p, Q = q\right] = d(p,q).$$

By iterated expectations, we get

$$\mathbb{E}\left[D(p)\right] = \mathbb{E}\left[\mathbb{E}\left[D(p)\middle|\phi(p,X)\right]\right] = \mathbb{E}\left[d(p,\phi(p,X))\right]$$

as desired. □

Proof of Theorem 3.21. The proof follows the rough outline of the proof of Theorem 2 of Besbes et al. (2010), but applied to our testing case and non-experimental estimators.

Decompose the test statistic ρ_n into three terms:

$$\rho_n = \overline{R}_n(\overline{p}_n) - \overline{R}_n(\hat{p}_n) = A_n + B_n + C_n,$$

where

$$\begin{aligned} A_n &= \overline{R}_n(\overline{p}_n) - \overline{R}_n(p^*), \\ B_n &= \overline{R}_n(p^*) - \overline{R}_n(\hat{p}), \\ C_n &= \overline{R}_n(\hat{p}) - \overline{R}_n(\hat{p}_n). \end{aligned}$$

We begin by showing that $(nh_n^3)\, A_n \xrightarrow{d} \Gamma\chi_1^2$. By Assumption 3.13, we have that $\overline{R}_n(p)$ is twice continuously differentiable. Thus, using Taylor's theorem to expand $\overline{R}_n(p)$ around $p = \overline{p}_n$, we get that there exists $p_n \in [\min(p^*, \overline{p}_n), \max(p^*, \overline{p}_n)]$ such that

$$\overline{R}_n(p^*) = \overline{R}_n(\overline{p}_n) + \overline{R}_n'(\overline{p}_n)(p^* - \overline{p}_n) + \frac{1}{2}\overline{R}_n''(p_n)(p^* - \overline{p}_n)^2.$$

By first order optimality conditions, $\overline{R}_n'(\overline{p}_n) = 0$. Hence, rearranging, we have

$$A_n = -\frac{1}{2}\overline{R}_n''(p_n)(p^* - \overline{p}_n)^2. \tag{B.5}$$

Next we show that $\overline{R}_n''(p_n) \xrightarrow{\mathbb{P}} R''(p^*)$. Note that

$$\left|\overline{R}_n''(p_n) - R''(p^*)\right| \leq \left|\overline{R}_n''(p_n) - R''(p_n)\right| + \left|R''(p_n) - R''(p^*)\right|. \tag{B.6}$$

Assumptions 3.13, 3.14, and 3.15 imply the assumptions of Lemma 5.1 of Newey (1994) applied to $R''(p)$, which in turn yields the uniform convergence in probability of $\overline{R}_n''(p)$ over $\mathcal{P}$ since it is compact by Assumption 3.14. Hence,

$$\left|\overline{R}_n''(p_n) - R''(p_n)\right| \leq \sup_{p\in\mathcal{P}} \left|\overline{R}_n''(p) - R''(p)\right| \xrightarrow{\mathbb{P}} 0. \tag{B.7}$$

By Theorem 3.16, $\overline{p}_n \xrightarrow{\mathbb{P}} p^*$. Because p_n is sandwiched between $\overline{p}_n$ and p^*, we also get $p_n \xrightarrow{\mathbb{P}} p^*$. Since $R''(p)$ is continuous by Assumption 3.13, we have

$$\left|R''(p_n) - R''(p^*)\right| \xrightarrow{\mathbb{P}} 0 \tag{B.8}$$

by continuous transformation of the former. Combining eqs. (B.6)-(B.8), we get

$$\overline{R}_n''(p_n) \xrightarrow{\mathbb{P}} R''(p^*). \tag{B.9}$$

By continuous transformation of the result of Theorem 3.16 (eq. (B.1)), we have

$$\left(nh_n^3\right)(\overline{p}_n - p^*)^2 \xrightarrow{d} \frac{\eta'}{R''(p^*)^2}\chi_1^2. \tag{B.10}$$

Combining eqs. (B.5)-(B.10), we get

$$\left(nh_n^3\right) A_n \xrightarrow{d} \frac{-\eta'}{2R''(p^*)}\chi_1^2 = \Gamma\chi_1^2.$$

Next, we show that $(nh_n^3)\, C_n \xrightarrow{\mathbb{P}} 0$. By Assumption 3.13, we have that $\overline{R}_n(p)$ is twice continuously differentiable. Thus, using Taylor's theorem to expand $\overline{R}_n(p)$ around $p = \hat{p}$, we get that there exists $p'_n \in [\min(\hat{p}, \hat{p}_n),\, \max(\hat{p}, \hat{p}_n)]$ such that

$$\overline{R}_n(\hat{p}_n) = \overline{R}_n(\hat{p}) + \overline{R}'_n(\hat{p})(\hat{p}_n - \hat{p}) + \frac{1}{2}\overline{R}''_n(p'_n)(\hat{p}_n - \hat{p})^2.$$

Rearranging, we have

$$\left(nh_n^3\right) C_n = -h_n^{3/2}\left(\sqrt{nh_n^3}\overline{R}'_n(\hat{p})\right)\left(\sqrt{n}(\hat{p}_n - \hat{p})\right) - \frac{1}{2}h_n^3\overline{R}''_n(p'_n)\left(\sqrt{n}(\hat{p}_n - \hat{p})\right)^2. \tag{B.11}$$

By Assumption 3.20, we have that

$$\sqrt{n}(\hat{p}_n - \hat{p}) = O_p(1), \text{ and hence also } \left(\sqrt{n}(\hat{p}_n - \hat{p})\right)^2 = O_p(1). \tag{B.12}$$

Applying Theorem 4 of Newey (1994) we get the convergence in distribution of $\sqrt{nh_n^3}\left(\overline{R}'_n(p) - R'(p)\right)$ for any fixed p, including $\hat{p}$ and hence we have

$$\sqrt{nh_n^3}\overline{R}'_n(\hat{p}) = O_p(1). \tag{B.13}$$

Next we show that $\overline{R}''_n(p'_n) = O_p(1)$. Note that

$$\left|\overline{R}''_n(p'_n) - R''(\hat{p})\right| \leq \left|\overline{R}''_n(p'_n) - R''(p'_n)\right| + \left|R''(p'_n) - R''(\hat{p})\right|. \tag{B.14}$$

As before, $\overline{R}''_n(p)$ converges uniformly to $R''(p)$ in probability over $\mathcal{P}$ and so

$$\left|\overline{R}''_n(p'_n) - R''(p'_n)\right| \leq \sup_{p\in\mathcal{P}}\left|\overline{R}''_n(p) - R''(p)\right| \xrightarrow{\mathbb{P}} 0. \tag{B.15}$$

By Assumption 3.20, $\hat{p}_n \xrightarrow{\mathbb{P}} \hat{p}$. Because p'_n is sandwiched between $\hat{p}_n$ and $\hat{p}$, we also get $p'_n \xrightarrow{\mathbb{P}} \hat{p}$. Since $R''(p)$ is continuous by Assumption 3.13, we have

$$\left|R''(p'_n) - R''(\hat{p})\right| \xrightarrow{\mathbb{P}} 0 \tag{B.16}$$

by continuous transformation of the former. Combining eqs. (B.14)-(B.16), we get

$$\overline{R}_n''(p_n') \xrightarrow{\mathbb{P}} R''(\hat{p}). \tag{B.17}$$

Combining eqs. (B.11)-(B.17) gives $(nh_n^3)\, C_n = -h_n^{3/2} O_p(1) - h_n^3 O_p(1)$. Hence, because $h_n \to 0$, we get $(nh_n^3)\, C_n \xrightarrow{\mathbb{P}} 0$.

Finally, we treat B_n. Under H_0, $B_n = 0$ because unique optimizer (Assumption 3.14) and $R(p^*) = R(\hat{p})$ (H_0) imply $p^* = \hat{p}$. Next, we show that under H_1, $(nh_n^3)\, B_n \xrightarrow{\mathbb{P}} \infty$. By applying the first results of Theorem 3.16 twice, we have that $B_n \xrightarrow{\mathbb{P}} R(p^*) - R(\hat{R})$. Since $k \geq 0$, Assumption 3.13 implies $nh_n^5 / \log(n) \to \infty$, which, since we also assume $h_n \to 0$, implies $nh_n^3 \to \infty$. Hence, since $R(p^*) - R(\hat{R}) > 0$ under H_1, we have that $(nh_n^3)\, B_n \xrightarrow{\mathbb{P}} \infty$. □

Appendix C

Appendix to Chapter 4

C.1 Computing a threshold $Q_{C_N}(\alpha)$

We provide two ways to compute $Q_{C_N}(\alpha)$ for use with the LCX-based GoF test. One is an exact, closed form formula, but which may be loose. Another uses the bootstrap to compute a tighter, but approximate threshold.

The theorem below employs a bound on $\mathbb{E}_F\left[||\xi||_2^2\right]$ to provide a valid threshold. This bound could either stem from known support bounds or from changing (4.14) to a two-sided hypothesis with two-sided confidence interval, using the lower bound as in (4.16) and the upper bound in (C.2) given below.

Theorem C.1. *Let $N \geq 2$. Suppose that with probability at least $1-\alpha_2$, $\mathbb{E}_F\left[||\xi||_2^2\right] \leq \overline{Q_{R_N}}(\alpha_2)$. Let $\alpha_1 \in (0,1)$ be given and suppose $F_0 \preceq_{LCX} F$. Then, with probability at least $1-\alpha_1-\alpha_2$,*

$$\mathbb{E}_F\left[||\xi||_2^2\right] \leq \overline{Q_{R_N}}(\alpha_2) \quad \textit{and}$$

$$C_N(F_0) \leq \left(1+\overline{Q_{R_N}}(\alpha_2)\right)\left(1+\frac{p}{2-p}\right)\frac{2^{\frac{1}{2}+\frac{1}{p}}}{N^{1-\frac{1}{p}}} \tag{C.1}$$

$$\times\sqrt{d+1+(d+1)\log\left(\frac{N}{d+1}\right)+\log\left(\frac{4}{\alpha_1}\right)}, \tag{C.2}$$

where

$$p=\frac{1}{2}\left(\sqrt{\log(256)+8\log(N)+(\log(2N))^2}-\log(2N)\right)\in(1,2). \tag{C.3}$$

Hence, defining $Q_{C_N}(\alpha_1)$ equal to the right-hand side of (C.2)*, we get a valid threshold for C_N in testing $F_0 \preceq_{LCX} F$ at level α_1.*

Proof. Fix any $p \in (1,2)$. Since

$$\mathcal{S}=\left\{\left\{\xi\in\Xi:\max\left\{a^T\xi-b,0\right\}\leq t\right\}:||a||_1+|b|\leq 1,\, t\in\mathbb{R}\right\}$$

is contained in the class of the empty set and all halfspaces, it has Vapnik-Chervonenkis

dimension at most $d+1$. Notice that for any $||a||_1 + |b| \leq 1$, $0 \leq \max\left\{a^T\xi - b, 0\right\} \leq \max\left\{1, ||\xi||_\infty\right\} \leq \max\left\{1, ||\xi||_2\right\}$. Therefore $\mathbb{E}_F\left[\max\left\{a^T\xi - b, 0\right\}^2\right] \leq 1 + \mathbb{E}_F\left[||\xi||_2^2\right]$ and

$$\begin{aligned}\int_0^\infty \left(\mathbb{P}_F\left(\max\left\{a^T\xi - b, 0\right\} > t\right)\right)^{1/p} dt &\leq 1 + \int_1^\infty \frac{\left(\mathbb{E}_F\left[\max\left\{a^T\xi - b, 0\right\}^2\right]\right)^{1/p}}{t^{2/p}} dt \\ &\leq \left(1 + \mathbb{E}_F\left[||\xi||_2^2\right]\right)^{1/p}\left(1 + \frac{p}{2-p}\right) \leq \left(1 + \mathbb{E}_F\left[||\xi||_2^2\right]\right)\left(1 + \frac{p}{2-p}\right)\end{aligned}$$

by Markov's inequality and $1 < p < 2$. Observe

$$\begin{aligned}C_N(F_0) &\leq \sup_{||a||_1+|b|\leq 1}\left(\mathbb{E}_{F_0}[\max\left\{a^T\xi - b, 0\right\}] - \mathbb{E}_F[\max\left\{a^T\xi - b, 0\right\}]\right) \\ &\quad + \sup_{||a||_1+|b|\leq 1}\left(\mathbb{E}_F[\max\left\{a^T\xi - b, 0\right\}] - \frac{1}{N}\sum_{i=1}^N \max\left\{a^T\xi^i - b, 0\right\}\right) \\ &\leq \sup_{||a||_1+|b|\leq 1}\left(\mathbb{E}_F[\max\left\{a^T\xi - b, 0\right\}] - \frac{1}{N}\sum_{i=1}^N \max\left\{a^T\xi^i - b, 0\right\}\right),\end{aligned} \tag{C.4}$$

where the second inequality follows because $F_0 \preceq_{LCX} F$. By applying Theorem 5.2 of Vapnik (1998) to the bottom-rightmost end of (C.4), we conclude that (C.2) holds for any $p \in (1, 2)$. The p given in (C.3) optimizes the bound for $N \geq 2$. □

Next we show how to bootstrap an approximate threshold $Q_{C_N}(\alpha)$. Recall that we seek a threshold $Q_{C_N}(\alpha)$ such that $\mathbb{P}\left(C_N(F_0) > Q_{C_N}(\alpha)\right) \leq \alpha$ whenever $F_0 \preceq_{\text{LCX}} F$. Employing (C.4), we see that a sufficient threshold is the $(1-\alpha)^{\text{th}}$ quantile of

$$\sup_{||a||_1+|b|\leq 1}\left(\mathbb{E}_F[\max\left\{a^T\xi - b, 0\right\}] - \frac{1}{N}\sum_{i=1}^N \max\left\{a^T\xi^i - b, 0\right\}\right),$$

where ξ^i are drawn IID from F. The bootstrap Efron and Tibshirani (1993) approximates this by replacing F with the empirical distribution $\hat{F}_N$. In particular, given an iteration count B, for $t = 1, \ldots, B$ it sets

$$Q^t = \sup_{||a||_1+|b|\leq 1}\left(\frac{1}{N}\sum_{i=1}^N \max\left\{a^T\xi^i - b, 0\right\} - \frac{1}{N}\sum_{i=1}^N \max\left\{a^T\tilde{\xi}^{t,i} - b, 0\right\}\right) \tag{C.5}$$

where $\tilde{\xi}^{t,i}$ are drawn IID from $\hat{F}_N$, i.e., IID random choices from $\{\xi^1, \ldots, \xi^N\}$. Then the bootstrap approximates $Q_{C_N}(\alpha)$ by the $(1-\alpha)^{\text{th}}$ quantile of $\{Q^1, \ldots, Q^B\}$. However, it may be difficult to compute (C.5) as the problem is non-convex. Fortunately (C.5) can be solved with a standard MILP formulation or by discretizing the space and enumerating (the objective is Lipschitz).

In particular, our bootstrap algorithm for computing $Q_{C_N}(\alpha)$ is as follows:

Input: $\xi^1, \ldots, \xi^N$ drawn from F, significance $0 < \alpha < 1$, precision $\delta > 0$, iteration count B
Output: Threshold $Q_{C_N}(\alpha)$ such that $\mathbb{P}\left(C_N(F_0) > Q_{C_N}(\alpha)\right) \lesssim \alpha$ whenever $F_0 \preceq_{\text{LCX}} F$.

For $t = 1, \ldots, B$:

1. Draw $\tilde{\xi}^{t,1}, \ldots, \tilde{\xi}^{t,N}$ IID from $\hat{F}_N$.
2. Solve $Q^t = \sup_{||a||_1 + |b| \leq 1} \left(\frac{1}{N} \sum_{i=1}^{N} \max\left\{a^T \xi^i - b, 0\right\} - \frac{1}{N} \sum_{i=1}^{N} \max\left\{a^T \tilde{\xi}^{t,i} - b, 0\right\} \right)$ to precision δ.

Sort $Q^{(1)} \leq \cdots \leq Q^{(B)}$ and return $Q^{(\lceil (1-\alpha)B \rceil)} + \delta$.

C.2 Omitted Proofs

Proof of Theorem 4.2. Fix any $x \in X$. Let $\epsilon > 0$ be given. By equicontinuity of the cost at x there is a $\delta > 0$ such that any $y \in X$ with $||x - y|| \leq \delta$ has $|c(x;\xi) - c(y;\xi)| \leq \epsilon$ for all $\xi \in \Xi$. Fix any such y. Then

$$\mathcal{C}(y;\mathcal{F}) = \sup_{F_0 \in \mathcal{F}} \mathbb{E}_{F_0}[c(y;\xi)] \leq \sup_{F_0 \in \mathcal{F}} \mathbb{E}_{F_0}[c(x;\xi)] + \epsilon = \mathcal{C}(x;\mathcal{F}) + \epsilon, \tag{C.6}$$

$$\mathcal{C}(x;\mathcal{F}) = \sup_{F_0 \in \mathcal{F}} \mathbb{E}_{F_0}[c(x;\xi)] \leq \sup_{F_0 \in \mathcal{F}} \mathbb{E}_{F_0}[c(y;\xi)] + \epsilon = \mathcal{C}(y;\mathcal{F}) + \epsilon. \tag{C.7}$$

Let $S = \{x \in X : \mathcal{C}(x;\mathcal{F}) < \infty\}$. By assumption $x_0 \in S$, so $S \neq \varnothing$. (C.7) implies that S is closed relative to X, which is closed, and therefore closed relative to $\mathbb{R}^{d_x}$. Since the objective is only finite on S we restrict our attention to S. (C.6) and (C.7) imply that $\mathcal{C}(x;\mathcal{F})$ is continuous in x on S.

If X is compact then S is compact. Suppose S is not compact and let $x_i \in S$ be any sequence such that $\lim_{i\to\infty} ||x_0 - x_i|| = \infty$. Then by coerciveness, $c_i(\xi) = c(x_i;\xi)$ diverges pointwise to infinity. Fix any $F_0 \in \mathcal{F}$. Let $c_i'(\xi) = \inf_{j \geq i} c_j(\xi)$, which is then pointwise monotone nondecreasing and pointwise divergent to infinity. Then, by Lebesgue's monotone convergence theorem, $\lim_{i\to\infty} \mathbb{E}_{F_0}[c_i'(\xi)] = \infty$. Since $c_i' \leq c_j$ pointwise for any $j \geq i$, we have $\mathbb{E}_{F_0}[c_i'(\xi)] \leq \inf_{j \geq i} \mathbb{E}_{F_0}[c_j(\xi)]$ and therefore

$$\infty = \lim_{i\to\infty} \mathbb{E}_{F_0}[c_i'(\xi)] \leq \lim_{i\to\infty} \inf_{j \geq i} \mathbb{E}_{F_0}[c_i(\xi)] = \liminf_{i\to\infty} \mathbb{E}_{F_0}[c_i(\xi)].$$

Thus $\mathcal{C}(x;\mathcal{F}) \geq \mathbb{E}_{F_0}[c(x;\xi)]$ is also coercive in x over S.

By the usual extreme value theorem, with either compactness or coerciveness, the continuous $\mathcal{C}(x;\mathcal{F})$ attains its minimal (finite) value at an $x \in S \subseteq X$. □

Proof of Proposition 4.4. Suppose that $c(x;\xi) \to \infty$ as $\xi \to \infty$. The case of unboundedness in the negative direction is similar. Let M be given. Choose $\rho > 0$ small so that $\xi^{(i)} - \xi^{(i-1)} > 2\rho$ for all i. For $\delta > 0$ and $\xi' \geq \xi^{(N)} + \rho$, let $F_{\delta,\xi'}$ be the

measure with density function

$$f(\xi;\delta,\xi') = \begin{cases} 1/(2N\rho) & \xi^{(i)} - \rho \le \xi \le \xi^{(i)} + \rho \text{ for } 1 \le i \le N-1, \\ 1/(2N\rho) & \xi^{(N)} - \rho + \delta\rho \le \xi \le \xi^{(N)} + \rho - \delta\rho, \\ 1/(2N\rho) & \xi' \le \xi \le \xi' + 2\delta\rho, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that for any ξ', $F_{0,\xi'}$ (i.e., take $\delta = 0$) minimizes $S_N(F_0)$ over distributions F_0. Since $\alpha > 0$, $Q_{S_N}(\alpha)$ is strictly greater than this minimum. Since $S_N(F_{\delta,\xi'})$ increases continuously with δ independently of ξ', there must exist $\delta > 0$ small enough so that $F_{\delta,\xi'} \in \mathcal{F}^{\alpha}_{S_N}$ for any $\xi' > \xi^{(N)} + \rho$. By infinite limit of the cost function, there exists $\xi' > \xi^{(N)} + \rho$ sufficiently large such that $c(x;\xi) \ge MN/\delta$ for all $\xi \ge \xi'$. Then, we have $\mathcal{C}\left(x;\mathcal{F}^{\alpha}_{S_N}\right) \ge \mathbb{E}_{F_{\delta,\xi'}}\left[c(x;\xi)\right] \ge \mathbb{P}\left(\xi \ge \xi'\right) MN/\delta = M$.

Since we have shown this for every $M > 0$, we have $\mathcal{C}\left(x;\mathcal{F}^{\alpha}_{S_N}\right) = \infty$. □

Proof of Proposition 4.8. We first prove that a uniformly consistent test is consistent. Let $G_0 \neq F$ be given. Denote by d be the Lévy-Prokhorov metric, which metrizes weak convergence Billingsley (1999), and observe that $d(G_0, F) > 0$.

Next, define $R_N = \sup_{F_0 \in \mathcal{F}_N} d(F_0, F)$. We claim that if the test is uniformly consistent, then $\mathbb{P}\left(R_N \to 0\right) = 1$. Indeed, suppose for some sample path, $R_N \not\to 0$. By the definition of the supremum, there must exist $\delta > 0$ and a sequence $F_N \in \mathcal{F}_N$ such that $d(F_N, F) \ge \delta$ i.o. Since d metrizes weak convergence, F_N does not converge to F. However, $F_N \in \mathcal{F}_N$ for all N, i.e. it is never rejected, which contradicts what must hold a.s. under uniform consistency.

Finally, since $\mathbb{P}\left(R_N \to 0\right) = 1$ and a.s. convergence implies convergence in probability, we have that $\mathbb{P}\left(R_N < \epsilon\right) \to 1$ for every $\epsilon > 0$, and, in particular, for $\epsilon = d(G_0, F)$. Then, $\mathbb{P}\left(G_0 \text{ rejected}\right) = \mathbb{P}\left(G_0 \notin \mathcal{F}_N\right) \ge \mathbb{P}\left(R_N < d(G_0, F)\right) \to 1$. This proves the first part of the proposition.

For the second part, we describe a test which is consistent but not uniformly consistent. Consider testing a continuous distribution F with the following univariate GoF test:

Given data $\xi^1, \ldots, \xi^N$ drawn from F and a hypothetical continuous distribution F_0:

Let $j = \lfloor \log_2 N \rfloor$, $i = N - 2^j$.

If $\dfrac{i}{2^j} \le F_0(\xi^1) \le \dfrac{i+1}{2^j}$ then F_0 is not rejected.

Otherwise, reject F_0 if it is rejected by the KS test at level $\dfrac{\alpha}{1 - 2^{-j}}$

applied to the data $\xi^2, \ldots, \xi^N$.

Notice that under the null-hypothesis, the probability of rejection is

$$\begin{aligned}\mathbb{P}(F_0 \text{ rejected }) &= \mathbb{P}\left(F_0(\xi^1) \notin \left[\frac{i}{2^j}, \frac{i+1}{2^j}\right]\right) \mathbb{P}(F_0 \text{ is rejected by the KS test }) \\ &= (1-2^{-j})\frac{\alpha}{1-2^{-j}} = \alpha,\end{aligned}$$

where we've used that ξ^1 is independent of the rest of the sample, and $F_0(\xi^1)$ is uniformly distributed for F_0 continuous. Consequently, the test is a valid GoF test and it has significance α.

We claim this test is also consistent. Specifically, consider any $F_0 \neq F$. By continuity of F_0 and consistency of the KS test,

$$\mathbb{P}\left(F_0 \text{ is rejected}\right) = \mathbb{P}\left(F_0(\xi^1) \notin \left[\frac{i}{2^j}, \frac{i+1}{2^j}\right]\right) \mathbb{P}\left(F_0 \text{ is rejected by the KS test}\right) \longrightarrow 1.$$

However, the test is not uniformly consistent. Fix any continuous $F_0 \neq F$ and let

$$F_N = \begin{cases} F_0 & \text{if } \frac{i}{2^j} \leq F_0(\xi^1) \leq \frac{i+1}{2^j}, \\ \hat{F}_N & \text{otherwise.} \end{cases}$$

Observe that $0 \leq F_0(\xi^1) \leq 1$ and $[0,1] = \bigcup_{i=0}^{2^j-1} \left[\frac{i}{2^j}, \frac{i+1}{2^j}\right]$. That is, for every j, $F_N = F_0$ at least once for $N \in \{2^j, \ldots, 2^{j+1}-1\}$. Therefore $F_N = F_0$ i.o, so it does not converge weakly to F. However, as constructed, F_N is never rejected by the above test. This is done for every sample path so the test cannot be uniformly consistent. □

To prove Theorems 4.12 and 4.15 we first establish two useful results.

Proposition C.2. *Suppose $\mathcal{F}_N$ is the confidence region of a uniformly consistent test and that Assumptions* (4.9) *and* (4.10) *hold. Then, almost surely, $\mathbb{E}_{F_N}[c(x;\xi)] \to E_F[c(x;\xi)]$ for any $x \in X$, $F_N \in \mathcal{F}_N$.*

Proof. Restrict to the a.s. event that ($F_N \not\to F \implies F_N \notin \mathcal{F}_N$ i.o). Fix $F_N \in \mathcal{F}_N$. Then the contrapositive gives $F_N \to F$. Fix x. If Ξ is bounded (Assumption 4.10a) then the result follows from the portmanteau lemma (see for example Theorem 2.1 of Billingsley (1999)). Suppose otherwise (Assumption (4.10)b). Then $\mathbb{E}_{F_N}[\phi(\xi)] \to \mathbb{E}_F[\phi(\xi)]$. By Theorem 3.6 of Billingsley (1999), $\phi(\xi)$ is uniformly integrable over $\{F_1, F_2, \ldots\}$. Since $c(x;\xi) = O(\phi(\xi))$, it is also uniformly integrable over these. Then the result follows by Theorem 3.5 of Billingsley (1999). □

Proposition C.3. *Suppose Assumption 4.9 holds and $\mathcal{C}(x_N;\mathcal{F}_N) \to \mathbb{E}_F[c(x;\xi)]$ for any convergent sequence $x_N \to x$. Then* (4.17) *holds.*

Proof. Let $E \subseteq X$ compact be given and suppose $\sup_{x\in E} |\mathcal{C}(x;\mathcal{F}_N) - \mathbb{E}_F[c(x;\xi)]| \not\to 0$ for contradiction. Then $\exists \epsilon > 0$ and $x_N \in E$ such that $|\mathcal{C}(x_N;\mathcal{F}_N) - \mathbb{E}_F[c(x_N;\xi)]| \geq \epsilon$ i.o. This, combined with compactness, means that there exists a subsequence $N_1 <$

$N_2 < \cdots < N_k \to \infty$ such that $x_{N_k} \to x \in E$ and $|\mathcal{C}(x_{N_k}; \mathcal{F}_{N_k}) - \mathbb{E}_F[c(x_{N_k}; \xi)]| \geq \epsilon$ $\forall k$. Then,

$$\begin{aligned} 0 < \epsilon &\leq |\mathcal{C}(x_{N_k}; \mathcal{F}_{N_k}) - \mathbb{E}_F[c(x_{N_k}; \xi)]| \\ &\leq |\mathcal{C}(x_{N_k}; \mathcal{F}_{N_k}) - \mathbb{E}_F[c(x; \xi)]| + |\mathbb{E}_F[c(x; \xi)] - \mathbb{E}_F[c(x_{N_k}; \xi)]| \,. \end{aligned}$$

By assumption, $\exists k_1$ such that $|\mathcal{C}(x_{N_k}; \mathcal{F}_{N_k}) - \mathbb{E}_F[c(x; \xi)]| \leq \epsilon/4$ $\forall k \geq k_1$. By equicontinuity and $x_{N_k} \to x$, $\exists k_2$ such that $|c(x; \xi) - c(x_{N_k}; \xi)| \leq \epsilon/4$ $\forall \xi$, $k \geq k_2$. Then,

$$|\mathbb{E}_F[c(x; \xi)] - \mathbb{E}_F[c(x_{N_k}; \xi)]| \leq \mathbb{E}_F[|c(x; \xi) - c(x_{N_k}; \xi)|] \leq \epsilon/4 \quad \forall \xi,\, k \geq k_2.$$

Combining and considering $k = \max\{k_1, k_2\}$, we get the contradiction $\epsilon \leq \epsilon/2$ for strictly positive ϵ. □

We prove the "if" and "only if" sides of Theorem 4.12 separately.

Proofs of Theorem 4.15 and the "if" side of Theorem 4.12. For either theorem restrict to the a.s. event that

$$\mathbb{E}_{F_N}[c(x; \xi)] \to \mathbb{E}_F[c(x; \xi)] \text{ for every } x \in X,\, F_N \in \mathcal{F}_N \tag{C.8}$$

(using Proposition C.2 for Theorem 4.12 or by assumption of c-consistency for Theorem 4.15).

Let any convergent sequence $x_N \to x$ and $\epsilon > 0$ be given. By equicontinuity and $x_N \to x$, $\exists N_1$ such that $|c(x_N; \xi) - c(x; \xi)| \leq \epsilon/2$ $\forall \xi$, $N \geq N_1$. Then, $|\mathcal{C}(x_N; \mathcal{F}_N) - \mathcal{C}(x; \mathcal{F}_N)| \leq \sup_{F_0 \in \mathcal{F}_N} \mathbb{E}_{F_0}[|c(x_N; \xi) - c(x; \xi)|] \leq \epsilon/2$ $\forall N \geq N_1$. By definition of supremum, $\exists F_N \in \mathcal{F}_N$ such that $\mathcal{C}(x; \mathcal{F}_N) \leq \mathbb{E}_{F_N}[c(x; \xi)] + \epsilon/4$. By (C.8), $\mathbb{E}_{F_N}[c(x; \xi)] \to \mathbb{E}_F[c(x; \xi)]$. Hence, $\exists N_2$ such that $|\mathbb{E}_{F_N}[c(x; \xi)] - \mathbb{E}_F[c(x; \xi)]| \leq \epsilon/4$ $\forall N \geq N_2$. Combining these with

$$|\mathcal{C}(x_N; \mathcal{F}_N) - \mathbb{E}_F[c(x; \xi)]| \leq |\mathcal{C}(x_N; \mathcal{F}_N) - \mathcal{C}(x; \mathcal{F}_N)| + |\mathcal{C}(x; \mathcal{F}_N) - \mathbb{E}_F[c(x; \xi)]| \,,$$

we get

$$|\mathcal{C}(x_N; \mathcal{F}_N) - \mathbb{E}_F[c(x; \xi)]| \leq \epsilon \quad \forall N \geq \max\{N_1, N_2\} \,.$$

Thus, by Proposition C.3, we get that (4.17) holds.

Let $A_N = \arg\min_{x \in X} \mathcal{C}(x; \mathcal{F}_N)$. We now show that $\bigcup_N A_N$ is bounded. If X is compact (Assumption 4.11a) then this is trivial. Suppose X is not compact (Assumption 4.11b). Using the same arguments as in the proof of Theorem 4.2, we have in particular that $\lim_{||x|| \to \infty} \mathbb{E}_F[c(x; \xi)] = \infty$, $z_{\text{stoch}} = \min_{x \in X} \mathbb{E}_F[c(x; \xi)] < \infty$, that $A = \arg\min_{x \in X} \mathbb{E}_F[c(x; \xi)]$ is compact, and each A_N is compact. Let $x^* \in A$. Fix $\epsilon > 0$. By definition of supremum $\exists F_N \in \mathcal{F}_N$ such that $\mathcal{C}(x^*; \mathcal{F}_N) \leq \mathbb{E}_{F_N}[c(x^*; \xi)] + \epsilon$. By (C.8), $\mathbb{E}_{F_N}[c(x^*; \xi)] \to \mathbb{E}_F[c(x^*; \xi)] = z_{\text{stoch}}$. Since true for any ϵ and since $\min_{x \in X} \mathcal{C}(x; \mathcal{F}_N) \leq \mathcal{C}(x^*; \mathcal{F}_N)$, we have $\limsup_{N \to \infty} \min_{x \in X} \mathcal{C}(x; \mathcal{F}_N) \leq z_{\text{stoch}}$. Now, suppose for contradiction that $\bigcup_N A_N$ is unbounded, i.e. there is a subsequence $N_1 < N_2 < \cdots < N_k \to \infty$ and $x_{N_k} \in A_{N_k}$ such that $||x_{N_k}|| \to \infty$. Let $\delta' = \limsup_{k \to \infty} \inf_{\xi \notin D} c(x_{N_k}; \xi) \geq \liminf_{N \to \infty} \inf_{\xi \notin D} c(x_N; \xi) > -\infty$ and $\delta =$

$\min\{0, \delta'\}$. By D-uniform coerciveness, $\exists k_0$ such that $c(x_{N_k}; \xi) \geq (z_{\text{stoch}}+1-\delta)/F(D)$ $\forall \xi \in D$, $k \geq k_0$. In the case of Theorem 4.12, let F_N be any $F_N \in \mathcal{F}_N$. In the case of Theorem 4.15, let F_N be the empirical distribution $F_N = \hat{F}_N \in \mathcal{F}_N$. In either case, we get $F_N \to F$ weakly. In particular, $F_N(D) \to F(D)$. Then $\mathbb{E}_{F_N}[c(x_{N_k}; \xi)] \geq F_N(D) \times (z_{\text{stoch}} + 1 - \delta)/F(D) + \min\{0, \inf_{\xi \notin D} c(x_{N_k}; \xi)\}$ $\forall k \geq k_0$. Thus $\limsup_{N\to\infty} \min_{x\in X} \mathcal{C}(x; \mathcal{F}_N) \geq \limsup_{k\to\infty} \min_{x\in X} \mathcal{C}(x; \mathcal{F}_{N_k}) \geq z_{\text{stoch}} + 1 - \delta + \delta = z_{\text{stoch}} + 1$, yielding the contradiction $z_{\text{stoch}} + 1 \leq z_{\text{stoch}}$.

Thus $\exists A_\infty$ compact such that $A \subseteq A_\infty$, $A_N \subseteq A_\infty$. Then, by (4.17),

$$\delta_N = \left| \min_{x\in X} \mathcal{C}(x; \mathcal{F}_N) - \min_{x\in X} \mathbb{E}_F[c(x; \xi)] \right| = \left| \min_{x\in A_\infty} \mathcal{C}(x; \mathcal{F}_N) - \min_{x\in A_\infty} \mathbb{E}_F[c(x; \xi)] \right|$$
$$\leq \sup_{x\in A_\infty} |\mathcal{C}(x; \mathcal{F}_N) - \mathbb{E}_F[c(x; \xi)]| \to 0,$$

yielding (4.18). Let $x_N \in A_N$. Since A_∞ is compact, x_N has at least one convergent subsequence. Let $x_{N_k} \to x$ be any convergent subsequence. Suppose for contradiction $x \notin A$, i.e., $\epsilon = \mathbb{E}_F[c(x; \xi)] - z_{\text{stoch}} > 0$. Since $x_{N_k} \to x$ and by equicontinuity, $\exists k_1$ such that $|c(x_{N_k}; \xi) - c(x; \xi)| \leq \epsilon/4$ $\forall \xi$, $k \geq k_1$. Then, $|\mathbb{E}_F[c(x_{N_k}; \xi)] - \mathbb{E}_F[c(x; \xi)]| \leq \mathbb{E}_F[|c(x_{N_k}; \xi) - c(x; \xi)|] \leq \epsilon/4$ $\forall k \geq k_1$. Also $\exists k_2$ such that $\delta_{N_k} \leq \epsilon/4$ $\forall k \geq k_2$. Then, for $k \geq \max\{k_1, k_2\}$,

$$\min_{x\in X} \mathcal{C}(x; \mathcal{F}_{N_k}) = \mathcal{C}(x_{N_k}; \mathcal{F}_N) \geq \mathbb{E}_F[c(x_{N_k}; \xi)] - \delta_N \geq \mathbb{E}_F[c(x; \xi)] - \epsilon/2 \geq z_{\text{stoch}} + \epsilon/2.$$

Taking limits, we contradict (4.18). □

Proof of the "only if" side of Theorem 4.12. Consider any Ξ bounded, $R = \sup_{\xi\in\Xi} ||\xi|| < \infty$. Let $X = \mathbb{R}^d$, and

$$c_1(x; \xi) = ||x|| \left(2 + \text{Re}\left(e^{ix^T\xi}\right)\right), \quad c_2(x; \xi) = ||x|| \left(2 - \text{Re}\left(e^{ix^T\xi}\right) + 2\right),$$
$$c_3(x; \xi) = ||x|| \left(2 + \text{Im}\left(e^{ix^T\xi}\right)\right), \quad c_4(x; \xi) = ||x|| \left(2 - \text{Im}\left(e^{ix^T\xi}\right) + 2\right).$$

Since $|c_i((x, y), \xi)| \leq 3\,||x||$, expectations exist. The gradient of each c_i at x has magnitude bounded by $R\,||x|| + 3$ uniformly over ξ, so equicontinuity is satisfied. Also, $\lim_{||x||\to\infty} c_i(x, y; \xi) \geq \lim_{||x||\to\infty} ||x|| = \infty$ uniformly over all $\xi \in \Xi$ and $c_i(x; \xi) \geq 0$, so Assumption 4.11 is satisfied. Restrict to the a.s. event that (4.17) applies

simultaneously for c_1, c_2, c_3, c_4. Then we have that, for every $x \in \mathbb{R}^d$,

$$2\,||x|| + ||x|| \sup_{F_0 \in \mathcal{F}_N} \operatorname{Re}\left(\mathbb{E}_{F_0}\left[e^{ix^T\xi}\right]\right) \longrightarrow 2\,||x|| + ||x||\operatorname{Re}\left(\mathbb{E}_{F}\left[e^{ix^T\xi}\right]\right)$$
$$2\,||x|| - ||x|| \inf_{F_0 \in \mathcal{F}_N} \operatorname{Re}\left(\mathbb{E}_{F_0}\left[e^{ix^T\xi}\right]\right) \longrightarrow 2\,||x|| - ||x||\operatorname{Re}\left(\mathbb{E}_{F}\left[e^{ix^T\xi}\right]\right)$$
$$2\,||x|| + ||x|| \sup_{F_0 \in \mathcal{F}_N} \operatorname{Im}\left(\mathbb{E}_{F_0}\left[e^{ix^T\xi}\right]\right) \longrightarrow 2\,||x|| + ||x||\operatorname{Im}\left(\mathbb{E}_{F}\left[e^{ix^T\xi}\right]\right)$$
$$2\,||x|| - ||x|| \inf_{F_0 \in \mathcal{F}_N} \operatorname{Im}\left(\mathbb{E}_{F_0}\left[e^{ix^T\xi}\right]\right) \longrightarrow 2\,||x|| - ||x||\operatorname{Im}\left(\mathbb{E}_{F}\left[e^{ix^T\xi}\right]\right)$$

This implies that $\sup_{F_0 \in \mathcal{F}_N} \left|\mathbb{E}_{F_0}\left[e^{ix^T\xi}\right] - \mathbb{E}_{F}\left[e^{ix^T\xi}\right]\right| \to 0$ for every x. Fix F_N such that $F_N \in \mathcal{F}_N$ eventually. Then $\mathbb{E}_{F_N}\left[e^{ix^T\xi}\right] \to \mathbb{E}_{F}\left[e^{ix^T\xi}\right]$ for every x. By the Lévy continuity theorem, F_N converge weakly to F. This is the contrapositive of the uniform consistency condition. □

Proof of Theorem 4.16. In the case of finite support $\Xi = \{\hat{\xi}^1, \ldots, \hat{\xi}^n\}$, total variation metrizes weak convergence:

$$d_{\mathrm{TV}}(q, q') = \frac{1}{2}\sum_{j=1}^{n} |q(j) - q'(j)|\,.$$

Restrict to the almost sure event $d_{\mathrm{TV}}(\hat{p}_N, p) \to 0$ (see Theorem 11.4.1 of Dudley (2002)). We need only show that now $\sup_{p_0 \in \mathcal{F}_N} d_{\mathrm{TV}}(\hat{p}_N, p_0) \to 0$, yielding the contrapositive of the uniform consistency condition.

By an application of the Cauchy-Schwartz inequality,

$$\begin{aligned} d_{\mathrm{TV}}(\hat{p}_N, p_0) = \frac{1}{2}\sum_{j=1}^{n} |\hat{p}_N(j) - p_0(j)| &\leq \frac{1}{2}\sum_{j=1}^{n} \frac{|\hat{p}_N(j) - p_0(j)|}{\sqrt{p_0(j)}} \\ &\leq \frac{1}{2}\left(\sum_{j=1}^{n} \frac{(\hat{p}_N(j) - p_0(j))^2}{p_0(j)}\right)^{1/2} = \frac{X_N(p_0)}{2}. \end{aligned}$$

By Kullback (1967),

$$d_{\mathrm{TV}}(\hat{p}_N, p_0) \leq \frac{1}{\sqrt{2}}\left(\sum_{j=1}^{n}\sum_{j=1}^{n} \hat{p}_N(j) \log\left(\hat{p}_N(j)/p_0(j)\right)\right)^{1/2} = \frac{G_N(p_0)}{2}.$$

Since both the χ^2 and G-tests use a rejection threshold equal to $\sqrt{Q/N}$ where Q is the $(1-\alpha)^{\text{th}}$ quantile of a χ^2 distribution with $n-1$ degrees of freedom (Q is independent of N), we have that $d_{\mathrm{TV}}(\hat{p}_N, p_0)$ is uniformly bounded over $p_0 \in \mathcal{F}_N$ by a quantity diminishing with N. □

Proof of Theorem 4.17. In the case of univariate support, the Lévy metric metrizes

weak convergence:

$$d_{\text{Lévy}}(G, G') = \inf\{\epsilon > 0 : G(\xi - \epsilon) - \epsilon \leq G'(\xi) \leq F_0(\xi + \epsilon) + \epsilon \ \forall \xi \in \mathbb{R}\}.$$

Restrict to the almost sure event $d_{\text{Lévy}}(\hat{F}_n, F) \to 0$ (see Theorem 11.4.1 of Dudley (2002)). We need only show that now $\sup_{F_0 \in \mathcal{F}_N} d_{\text{Lévy}}(\hat{F}_N, F_0) \to 0$, yielding the contrapositive of the uniform consistency condition.

Fix F_0 and let $0 \leq \epsilon < d_{\text{Lévy}}(\hat{F}_N, F_0)$. Then $\exists \xi_0$ such that either (1) $\hat{F}_N(\xi_0 - \epsilon) - \epsilon > F_0(\xi_0)$ or (2) $\hat{F}_N(\xi_0 + \epsilon) + \epsilon < F_0(\xi_0)$. Since F_0 is monotonically non-decreasing, (1) implies $D_N(F_0) \geq \hat{F}_N(\xi_0 - \epsilon) - F_0(\xi_0 - \epsilon) > \epsilon$ and (2) implies $D_N(F_0) \geq F_0(\xi_0 + \epsilon) - \hat{F}_N(\xi_0 + \epsilon) > \epsilon$. Hence $d_{\text{Lévy}}(\hat{F}_N, F_0) \leq D_N(F_0)$. Moreover, $D_N \leq V_N$ by definition. Since $\sup_{F_0 \in \mathcal{F}^{\alpha}_{S_N}} S_N(F_0) = Q_{S_N}(\alpha) = O(N^{-1/2})$ for either statistic, both the KS and Kuiper tests are uniformly consistent.

Consider $D'_N(F_0) = \max_{i=1,\ldots,N} \left| F_0(\xi^{(i)}) - \frac{2i-1}{2N} \right| = \sigma \left(F_0(\xi^{(j)}) - \frac{2j-1}{2N} \right)$, where j and σ are the maximizing index and sign, respectively. Suppose $D'_N(F_0) \geq 1/\sqrt{N} + 1/N$. If $\sigma = +1$, this necessarily means that $1 - \frac{2j-1}{2N} \geq 1/\sqrt{N} + 1/N$ and therefore $N - j \geq \lceil \sqrt{N} \rceil + 1$. By monotonicity of F_0 we have for $0 \leq k \leq \lceil \sqrt{N} \rceil$ that $j + k \leq N$ and

$$F_0(\xi^{(j+k)}) - \frac{2(j+k)-1}{2N} \geq F_0(\xi^{(j)}) - \frac{2j-1}{2N} - \frac{k}{N} = D'_N(F_0) - \frac{k}{N} \geq 0.$$

If instead $\sigma = -1$, this necessarily means that $\frac{2j-1}{2N} \geq 1/\sqrt{N} + 1/N$ and therefore $j \geq \lceil \sqrt{N} \rceil + 1$. By monotonicity of F_0 we have for $0 \leq k \leq \lceil \sqrt{N} \rceil$ that $j - k \geq 1$ and

$$\frac{2(j-k)-1}{2N} - F_0(\xi^{(j-k)}) \geq \frac{2j-1}{2N} - F_0(\xi^{(j)}) - \frac{k}{N} = D'_N(F_0) - \frac{k}{N} \geq 0.$$

In either case we have that

$$\begin{aligned} W^2_N &= \frac{1}{12N^2} + \frac{1}{N} \sum_{i=1}^{N} \left(F_0(\xi^{(i)}) - \frac{2i-1}{2N} \right)^2 \\ &\geq \frac{1}{12N^2} + \frac{1}{N} \sum_{k=0}^{\lceil \sqrt{N} \rceil} \left(D'_N - \frac{k}{N} \right)^2 \geq \frac{D^2_N}{\sqrt{N}} - \frac{2}{N} \end{aligned}$$

using $D'_N(F_0) \geq 1/\sqrt{N} + 1/N$ and $|D'_N(F_0) - D_N(F_0)| \leq 1/(2N)$ in the last inequality. Therefore,

$$D^2_N(F_0) \leq \max \left\{ \frac{1}{\sqrt{N}} + \frac{3}{2N}, \sqrt{N} W^2_N(F_0) + \frac{2}{\sqrt{N}} \right\}.$$

Since $F_0(\xi)(1 - F_0(\xi)) \leq 1$ and by using the integral formulation of CvM and AD (see Thas (2009)) the same is true replacing W^2_N by A^2_N. Since $Q_{S_N}(\alpha) = O(N^{-1/2})$ so that $\sup_{F_0 \in \mathcal{F}^{\alpha}_{S_N}} S^2_N(F_0) = O(N^{-1})$ for either statistic, both the CvM and AD tests

are uniformly consistent.

$$\begin{aligned} W_N^2 - U_N^2 = & \left(\frac{1}{N} \sum_{i=1}^{N} F_0(\xi^{(i)}) - \frac{1}{2} \right)^2 \\ & \leq \max \left\{ \left(\frac{1}{N} \sum_{i=1}^{N} \min \left\{ 1, \frac{2i-1}{2N} + D'_N(F_0) \right\} - \frac{1}{2} \right)^2, \right. \\ & \qquad \left. \left(\frac{1}{N} \sum_{i=1}^{N} \max \left\{ 0, \frac{2i-1}{2N} - D'_N(F_0) \right\} - \frac{1}{2} \right)^2 \right\}. \end{aligned}$$

Letting $M = \lfloor \frac{1}{2} + N(1 - D'_N(F_0)) \rfloor$ we have

$$\sum_{i=1}^{N} \min \left\{ 1, \frac{2i-1}{2N} + D'_N(F_0) \right\} = \frac{M^2}{2N} + M D'_N(F_0) + N - M$$

so that in the case of $D'_N(F_0) \geq 1/\sqrt{N} + 1/N$,

$$\left(\frac{1}{N} \sum_{i=1}^{N} \min \left\{ 1, \frac{2i-1}{2N} + D'_N(F_0) \right\} - \frac{1}{2} \right)^2 = O(1/N).$$

Thus, the Watson test is also uniformly consistent. □

Proof of Proposition 4.18. Apply Theorem 4.12 to each i and restrict to the almost sure event that (4.17) holds for all i. Fix F_N such that $F_N \in \mathcal{F}_N$ eventually. Then, (4.17) yields $\mathbb{E}_{F_N}[c_i(x;\xi_i)] \to \mathbb{E}_F[c_i(x;\xi_i)]$ for every $x \in X$. Summing over i yields the contrapositive of the c-consistency condition. □

Proof of Proposition 4.19. Restrict to a sample path in the a.s. event $\mathbb{E}_{\hat{F}_N}[\xi_i] \to \mathbb{E}_{\hat{F}}[\xi_i]$, $\mathbb{E}_{\hat{F}_N}[\xi_i \xi_j] \to \mathbb{E}_{\hat{F}}[\xi_i \xi_j]$ for all i, j. Consider any F_N such that $F_N \in \mathcal{F}^{\alpha}_{\mathrm{CEG},N}$ eventually. Then clearly $\mathbb{E}_{F_N}[\xi_i] \to \mathbb{E}_F[\xi_i]$, $\mathbb{E}_{F_N}[\xi_i \xi_j] \to \mathbb{E}_F[\xi_i \xi_j]$.

Consider any F_N such that $F_N \in \mathcal{F}^{\alpha}_{\mathrm{DY},N}$ eventually. Because covariances exist, we may restrict to N large enough so that $\left|\left|\hat{\Sigma}_N\right|\right|_2 \leq M$ (operator norm) and $F_N \in \mathcal{F}^{\alpha}_{\mathrm{DY},N}$. Then we get

$$||\mathbb{E}_{F_N}[\xi] - \hat{\mu}_N|| \leq M \gamma_{1,N}(\alpha) \to 0$$

and

$$(\gamma_{3,N}(\alpha) - 1)\, \hat{\Sigma}_N \preceq \mathbb{E}_{F_0}[(\xi - \hat{\mu}_N)\,(\xi - \hat{\mu}_N)^T] - \hat{\Sigma}_N \preceq (\gamma_{2,N}(\alpha) - 1)\, \hat{\Sigma}_N,$$

which gives

$$\left|\left| \mathbb{E}_{F_0}[(\xi - \hat{\mu}_N)\,(\xi - \hat{\mu}_N)^T] - \hat{\Sigma}_N \right|\right|_2 \leq M \max\{\gamma_{2,N}(\alpha) - 1, 1 - \gamma_{3,N}(\alpha)\} \to 0.$$

Then again, we have $\mathbb{E}_{F_N}[\xi_i] \to \mathbb{E}_F[\xi_i]$, $\mathbb{E}_{F_N}[\xi_i\xi_j] \to \mathbb{E}_F[\xi_i\xi_j]$.

In either case we get $\mathbb{E}_{F_N}[c(x;\xi)] \to \mathbb{E}_{F_N}[c(x;\xi)]$ for any x due to factorability as in (4.25). This yields the contrapositive of the c-consistency condition. □

Proof of Proposition 4.20. If $F_0 \neq F$ then Theorem 1 of Scarsini (1998) yields that either $F_0 \not\preceq_{\text{LCX}} F$ or there is some $j = 1, \ldots, d$ such that $\mathbb{E}_{F_0}[\xi_j^2] \neq \mathbb{E}_F[\xi_j^2]$. If $F_0 \not\preceq_{\text{LCX}} F$ then power approaches one since $C_N > 0$ but $Q_{C_N}(\alpha_1) \to 0$. Otherwise, $F_0 \preceq_{\text{LCX}} F$ yields $\mathbb{E}_{F_0}[\xi_i^2] \leq \mathbb{E}_F[\xi_i^2]$ for all i via (4.11) using $a = e_i$ and $\phi(\zeta) = \zeta^2$. Then $\mathbb{E}_{F_0}[\xi_j^2] \neq \mathbb{E}_F[\xi_j^2]$ must mean that $\mathbb{E}_{F_0}\left[||\xi||_2^2\right] < \mathbb{E}_F\left[||\xi||_2^2\right]$ and power still goes to one. □

Proof of Proposition 4.21. Let $R = \sup_{\xi\in\Xi} ||\xi||_2 < \infty$. Restrict to the almost sure event that $\hat{F}_N \to F$. Consider F_N such that $F_N \in \mathcal{F}_N$ eventually. Let N be large enough so that it is so. Fix $||a||_2 = 1$. Let $a_1 = a$ and complete an orthonormal basis for $\mathbb{R}^d$: $a_1, a_2, \ldots, a_d$. On the one hand we have $Q_{R_N}(\alpha_2) \geq \mathbb{E}_{\hat{F}_N}\left[\sum_{i=1}^d (a_i^T\xi)^2\right] - \mathbb{E}_{F_N}\left[\sum_{i=1}^d (a_i^T\xi)^2\right]$. On the other hand, for each i,

$$\begin{aligned}
&\mathbb{E}_{\hat{F}_N}\left[(a_i^T\xi)^2\right] - \mathbb{E}_{F_N}\left[(a_i^T\xi)^2\right] = \\
&\qquad 2\int_{b=-R}^{0} \left(\mathbb{E}_{\hat{F}_N}[\max\left\{b - a_i^T\xi, 0\right\}] - \mathbb{E}_{F_N}[\max\left\{b - a_i^T\xi, 0\right\}]\right) db \\
&\qquad + 2\int_{b=0}^{R} \left(\mathbb{E}_{\hat{F}_N}[\max\left\{a_i^T\xi - b, 0\right\}] - \mathbb{E}_{F_N}[\max\left\{a_i^T\xi - b, 0\right\}]\right) db \\
&\geq 4\int_{b=0}^{R} (||a||_1 + |b|) Q_{C_N}(\alpha_1) db \geq 4\left(\sqrt{d} + R^2/2\right) Q_{C_N}(\alpha_1) = p_N.
\end{aligned}$$

Therefore, $q_N = Q_{R_N}(\alpha_2) + (d-1)p_N \geq \mathbb{E}_{\hat{F}_N}\left[(a^T\xi)^2\right] - \mathbb{E}_{F_N}\left[(a^T\xi)^2\right]$ and $Q_{R_N}(\alpha_2), Q_{C_N}(\alpha_1), p_N, q_N \to 0$. Let $G_N(t) = F_N(\left\{\xi : a^T\xi \leq t\right\}) \in [0,1]$ and $\hat{G}_N(t) = \hat{F}_N(\left\{\xi : a^T\xi \leq t\right\}) \in [0,1]$ be

the CDFs of $a^T\xi$ under F_N and $\hat{F}_N$, respectively. Then,

$$\begin{aligned}
q_N \geq\ & \mathbb{E}_{\hat{F}_N}\left[(a^T\xi)^2\right] - \mathbb{E}_{F_N}\left[(a^T\xi)^2\right] = \\
& 2\int_{b=-R}^{0} \left(\mathbb{E}_{\hat{F}_N}[\max\left\{b - a^T\xi, 0\right\}] - \mathbb{E}_{F_N}[\max\left\{b - a^T\xi, 0\right\}]\right) db \\
& + 2\int_{b=0}^{R} \left(\mathbb{E}_{\hat{F}_N}[\max\left\{a^T\xi - b, 0\right\}] - \mathbb{E}_{F_N}[\max\left\{a^T\xi - b, 0\right\}]\right) db \\
=\ & 2\int_{b=-R}^{0}\int_{t=-R}^{b} \left(\hat{G}_N(t) - G_N(t)\right) dt\, db \\
& + 2\int_{b=0}^{R}\int_{t=b}^{R} \left(G_N(t) - \hat{G}_N(t)\right) dt\, db \geq p_N, \\
\int_{t=-R}^{b} & \left(\hat{G}_N(t) - G_N(t)\right) dt \geq -(\sqrt{d} + R)Q_{C_N}(\alpha) \quad \forall b \in [-R, 0], \\
\int_{t=b}^{R} & \left(G_N(t) - \hat{G}_N(t)\right) dt \geq -(\sqrt{d} + R)Q_{C_N}(\alpha) \quad \forall b \in [0, R],
\end{aligned}$$

Because $\hat{F}_N \to F$, we get $\hat{G}_N(t) \to F(\left\{\xi : a^T\xi \leq t\right\})$ and therefore at every continuity point t we have $G_N(t) \to F(\left\{\xi : a^T\xi \leq t\right\})$. Because true for every a, the Cramer-Wold device yields $F_N \to F$. This is the contrapositive of the uniform consistency condition. □

Proof of Theorem 4.23. Problem (4.3) is equal to the optimization problems of Theorem 4.22 augmented with the variable $x \in X$ and weak optimization is polynomially reducible to weak separation (see Grotschel et al. (1993)). Tractable weak separation for all constraints except $x \in X$ and (4.27) is given by the tractable weak optimization over these standard conic-affine constraints. A weak separation oracle is assumed given for $x \in X$. We polynomially reduce separation over $c_j \geq \max_k c_{jk}(x)$ for fixed c'_j, x' to the oracles. We first call the evaluation oracle for each k to check violation and if there is a violation and $k^* \in \arg\max_k c_{jk}(x')$ then we call the subgradient oracle to get $s \in \partial c_{jk^*}(x')$ with $||s||_\infty \leq 1$ and produce the separating hyperplane $0 \geq c_{jk^*}(x') - c_j + s^T(x - x')$. □

Proof of Theorem 4.25. Substituting the given formulas for K_{S_N}, A_{S_N}, $b_{S_N,\alpha}$ for each $S_N \in \{D_N, V_N, W_N, U, N, A_N\}$ in $A_{S_N}\zeta - b_{S_N,\alpha} \in K_{S_N}$ we obtain $S_N(\zeta_1, \ldots, \zeta_N) \leq Q_{S_N}(\alpha)$ exactly for S_N as defined in (4.7). We omit the detailed arithmetic. □

Proof of Theorem 4.27. Under these assumptions (4.3) is equal to the optimization problems of Theorem 4.26 augmented with the variable x and weak optimization is polynomially reducible to weak separation (see Grotschel et al. (1993)). Tractable weak separation for all constraints except $x \in X$ and (4.29) is given by the tractable weak optimization over these standard conic-affine constraints. A weak separation oracle is assumed given for $x \in X$. By continuity and given structure of $c(x; \xi)$, we

may rewrite (4.29) as

$$c_i \geq \max_{\xi \in [\xi^{(i-1)}, \xi^{(i)}]} c_k(x;\xi) \quad \forall k = 1, \ldots, K. \tag{C.9}$$

We polynomially reduce weak δ-separation over the k^{th} constraint at fixed c_i', x' to the oracles. We call the δ-optimization oracle to find $\xi' \in [\xi^{(i-1)}, \xi^{(i)}]$ such that $c_k(x';\xi') \geq \max_{\xi \in [\xi^{(i-1)}, \xi^{(i)}]} c_k(x;\xi) - \delta$. If $c_i' \geq c_k(x';\xi')$ then $(c_i' + \delta, x')$ satisfy the constraint and is within δ of (c_i', x'). If $c_i' < c_k(x';\xi')$ then we call the subgradient oracle to get $s \in \partial_x c_k(x', \xi')$ with $||s||_\infty \leq 1$ and produce the hyperplane $c_i \geq c(x';\xi') + s^T(x - x')$ that is violated by (c_i', x') and for any (c_i, x) satisfying (C.9) (in particular if it is in the δ-interior) we have $c_i \geq \max_{\xi \in [\xi^{(i-1)}, \xi^{(i)}]} c_k(x;\xi) \geq c_k(x;\xi') \geq c_k(x';\xi') + s^T(x - x')$ since s is a subgradient. The case for constraints (4.30) is similar. □

Proof of Lemma 4.29. According to Theorem 4.26, the observations in Example 4.28, and by renaming variables, the DRO (4.3) is given by

$$\begin{aligned}
(P)\ \min\ & y + \sum_{i=1}^{N} \left(Q_{D_N}(\alpha) + \frac{i-1}{N} \right) s_i + \sum_{i=1}^{N} \left(Q_{D_N}(\alpha) - \frac{i}{N} \right) t_i \\
\text{s.t. } & x \in \mathbb{R}_+,\ y \in \mathbb{R},\ s \in \mathbb{R}_+^N,\ t \in \mathbb{R}_+^N \\
& (r-c)x + y + \sum_{i=j}^{N} (s_i - t_i) \geq (r-c)\xi^{(j)} && \forall j = 1, \ldots, N+1 \\
& -(c-b)x + y + \sum_{i=j}^{N} (s_i - t_i) \geq -(c-b)\xi^{(j-1)} && \forall j = 1, \ldots, N+1.
\end{aligned}$$

Applying linear optimization duality we get that its dual is

$$\begin{aligned}
(D)\ \max\ & (r-c) \sum_{i=1}^{N+1} \xi^{(i)} p_i - (c-b) \sum_{i=1}^{N+1} \xi^{(i-1)} q_i \\
\text{s.t. } & p \in \mathbb{R}_+^{N+1},\ q \in \mathbb{R}_+^{N+1} \\
& (r-c) \sum_{i=1}^{N+1} p_i - (c-b) \sum_{i=1}^{N+1} q_i \leq 0 \\
& \sum_{i=1}^{N+1} p_i + \sum_{i=1}^{N+1} q_i = 1 \\
& \sum_{i=1}^{j} p_i + \sum_{i=1}^{j} q_i \leq Q_{D_N}(\alpha) + \frac{j-1}{N} && \forall j = 1, \ldots, N \\
& -\sum_{i=1}^{j} p_i - \sum_{i=1}^{j} q_i \leq Q_{D_N}(\alpha) - \frac{j}{N} && \forall j = 1, \ldots, N.
\end{aligned}$$

It can be directly verified that the following primal and dual solutions are respectively

feasible

$$
\begin{aligned}
&x = (1-\theta)\xi^{(i_{\text{lo}})} + \theta\xi^{(i_{\text{hi}})}, \\
&y = (r-c)\xi^{(N+1)} - (r-c)x, \\
&s_i = \begin{cases} (c-b)\left(\xi^{(i)} - \xi^{(i-1)}\right) & i \le i_{\text{lo}} \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, N, \\
&t_i = \begin{cases} (r-c)\left(\xi^{(i+1)} - \xi^{(i)}\right) & i \ge i_{\text{hi}} \\ 0 & \text{otherwise} \end{cases} \qquad \forall i = 1, \ldots, N
\end{aligned}
$$

$$
p_i = \begin{cases} 0 & i \le i_{\text{hi}} - 1 \\ i/N - \theta - Q_{D_N}(\alpha) & i = i_{\text{hi}} \\ 1/N & N \ge i \ge i_{\text{hi}} + 1 \\ Q_{D_N}(\alpha) & i = N+1 \end{cases}, \qquad \forall i = 1, \ldots, N
$$

$$
q_i = \begin{cases} Q_{D_N}(\alpha) & i = 1 \\ 1/N & 2 \le i \le i_{\text{lo}} \\ \theta - Q_{D_N}(\alpha) - (i-2)/N & i = i_{\text{lo}} + 1 \\ 0 & i \ge i_{\text{lo}} + 2 \end{cases} \qquad \forall i = 1, \ldots, N
$$

and that both have objective cost in their respective programs of

$$
\begin{aligned}
z = &-(c-b)Q_{D_N}(\alpha)\xi^{(0)} - \frac{c-b}{N}\sum_{i=1}^{i_{\text{lo}}-1}\xi^{(i)} - (c-b)\left(\theta - Q_{D_N}(\alpha) - \frac{i_{\text{lo}}-1}{N}\right)\xi^{(i_{\text{lo}})} \\
&+ (r-c)Q_{D_N}(\alpha)\xi^{(N+1)} + \frac{r-c}{N}\sum_{i=i_{\text{hi}}+1}^{N}\xi^{(i)} + (r-c)\left(\frac{i_{\text{hi}}}{N} - Q_{D_N}(\alpha) - \theta\right)\xi^{(i_{\text{hi}})}.
\end{aligned}
$$

This proves optimality of x. Adding $0 = (c-b)\theta x - (r-c)(1-\theta)x$ to the above yields the form of the optimal objective given in the statement of the result.

□

Proof of Theorem 4.32. Fix x. Let $S = \{(a,b) \in \mathbb{R}^{d+1} : ||a||_1 + |b| \le 1\}$. Using the notation of Shapiro (2001), letting C be the cone of nonnegative measures on Ξ and C' the cone of nonnegative measures on S, we write the inner problem as

$$
\begin{aligned}
\sup_{F} \quad & \langle F, c(x;\xi)\rangle \\
\text{s.t.} \quad & F \in C,\ \langle 1, F\rangle = 1 \\
& \sup_{(a,b)\in S}\left(\left\langle F, \max\{a_j^T\xi - b, 0\}\right\rangle - \frac{1}{N}\sum_{i=\ell+1}^{N}\max\{a^T\xi^i - b, 0\} - Q_{C_N}(\alpha_1)\right) \le 0 \\
& \left\langle F, ||\xi||_2^2\right\rangle \ge Q_{R_N}^{\alpha_2}
\end{aligned}
$$

Invoking Proposition 2.8 of Shapiro (2001) (with the generalized Slater point equal to the empirical distribution), and using the representation (4.36) of the cost function,

we have that the strongly dual minimization problem is

$$\min_{G,\tau,\theta} \quad \theta + Q_{C_N}(\alpha)G\{S\} + \left\langle G, \frac{1}{N}\sum_{i=\ell+1}^{N} \max\{a^T\xi^i - b, 0\} \right\rangle - Q_{R_N}^{\alpha_2}\tau \tag{C.10}$$

$$\text{s.t.} \quad G \in C',\ \tau \in \mathbb{R}_+,\ \theta \in \mathbb{R}$$
$$\inf_{\xi\in\mathbb{R}^d} \left(\left\langle G, \max\{a^T\xi - b, 0\} \right\rangle - \left(p_{k2} + P_k^T x\right)^T \xi - \tau \left|\left|\xi\right|\right|_2^2 \right) \geq p_{k0} + p_{k1}^T x - \theta \quad \forall k. \tag{C.11}$$

We will now show that only $\tau = 0$ is feasible. Consider any feasible solution with $\tau > 0$. Notice that

$$\left\langle G, \max\{a^T\xi - b, 0\} \right\rangle - \left(p_{k2} + P_k^T x\right)^T \xi \leq \left(G\{S\} + \left|\left|p_{k2} + P_k^T x\right|\right|_\infty\right)\left(\left|\left|\xi\right|\right|_1 + 1\right),$$

which grows linearly with growing ξ. In contrast, $\tau\left|\left|\xi\right|\right|_2^2$ grows quadratically, i.e. strictly faster. Therefore, the left-hand side of (C.11) is negative infinity but the right hand side is finite. Therefore only $\tau = 0$ is feasible so fix it as such. Now rewrite the k^{th} constraint in (C.11) as follows

$$\begin{aligned} p_{k0} + p_{k1}^T x - \theta \leq \min_{\xi_k\in\mathbb{R}^d, g_k} \quad & \langle G, g_k \rangle - \left(p_{k2} + P_k^T x\right)^T \xi_k \\ \text{s.t.} \quad & \inf_{(a,b)\in S} \left(g_k(a,b) - a^T\xi_k + b\right) \geq 0 \\ & \inf_{(a,b)\in S} g_k(a,b) \geq 0. \end{aligned}$$

Again invoking Proposition 2.8 of Shapiro (2001) (with the generalized Slater point $\xi = e$, $g(a,b) = \max\{a^T e - b, 0\} + 1$) we see that the above is equivalent to

$$\begin{aligned} \exists H_k \text{ s.t.} \quad & \langle H_k, -b \rangle \geq p_{k0} + p_{k1}^T x - \theta \\ & H_k \in C',\ (G - H_k) \in C' \\ & \langle H_K, a \rangle = p_{k2} + P_k^T x. \end{aligned}$$

Thus, introducing these variables H_k into the problem (C.10) as well as the variable $x \in X$ and invoking Proposition 2.8 of Shapiro (2001) again (with the generalized Slater point being all variables zero except θ sufficiently large) we get the strongly

dual maximization problem

$$
\begin{aligned}
\max_{r,s,t,\psi} \quad & \sum_{k=1}^{K} (p_{k0} r_k + p_{k2} s_k) + h^T t \\
\text{s.t.} \quad & r \in \mathbb{R}^k_+,\ s \in \mathbb{R}^{k \times d},\ t \in \mathbb{R}^{d'} \\
& \inf_{(a,b) \in S} \psi_k(a,b) \geq 0 && \forall k = 1, \ldots, K \\
& \inf_{(a,b) \in S} \left(\psi_k(a,b) - a^T s_k + b r_k \right) \geq 0 && \forall k = 1, \ldots, K \\
& \sup_{(a,b) \in S} \left(\sum_{k=1}^{K} \psi_k(a,b) - \frac{1}{N} \sum_{i=1}^{N} \max\{a^T \xi^i - b, 0\} - Q_{C_N}(\alpha_1) \right) \leq 0 \\
& \sum_{k=1}^{K} r_k = 1 \\
& H^T t - \sum_{k=1}^{K} (r_k p_{k1} - P_k z_k) \leq 0 && \text{(C.12)}
\end{aligned}
$$

where x is the dual variable associated with (C.12). Recognizing that given any feasible solution we may set $\psi_k(a,b) = \max\{a^T s_k - b r_k, 0\}$ and remain feasible with the same objective, we arrive at the formulation in the statement of the theorem. $\square$

Appendix D

Appendix to Chapter 5

D.1 Omitted Proofs

Proof of Theorem 5.1. For the first part, let $\mathbf{x}^*$ be robust feasible in (5.2) and consider the closed, convex set $\{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$ where $t > 0$. That $\mathbf{x}^*$ is robust feasible implies $\max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{u}, \mathbf{x}^*) \leq 0$ which implies that $\mathcal{U}$ and $\{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$ are disjoint. From the separating hyperplane theorem, there exists a strict separating hyperplane $\mathbf{v}^T\mathbf{u} = v_0$ such that $v_0 > \mathbf{v}^T\mathbf{u}$ for all $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v}^T\mathbf{u} < v_0$ for all $\mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^d : f(\mathbf{u}, \mathbf{x}^*) \geq t\}$. Observe

$$v_0 > \max_{\mathbf{u}\in\mathcal{U}} \mathbf{v}^T\mathbf{u} = \delta^*(\mathbf{v}|\ \mathcal{U}) \geq \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}),$$

and

$$\mathbb{P}(f(\tilde{\mathbf{u}}, \mathbf{x}^*) \geq t) \leq \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}} > v_0) \leq \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}} > \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v})) \leq \epsilon.$$

Taking the limit as $t \downarrow 0$ and using the continuity of probability proves $\mathbb{P}(f(\tilde{\mathbf{u}}, \mathbf{x}^*) > 0) \leq \epsilon$ and that (5.2) is satisfied.

For the second part of the theorem, let $t > 0$ be such that $\delta^*(\mathbf{v}|\ \mathcal{U}) \leq \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) - t$. Define $f(\mathbf{u}, x) \equiv \mathbf{v}^T\mathbf{u} - x$. Then $x^* = \delta(\mathbf{v}|\ \mathcal{U})$ is robust feasible in (5.2), but

$$\mathbb{P}(f(\tilde{\mathbf{u}}, \mathbf{x}) > 0) = \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > \delta(\mathbf{v}|\ \mathcal{U})) \geq \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} \geq \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) - t) > \epsilon$$

by (5.6).

□

Proof. Proof of Theorem 5.2.

$$\begin{aligned}
&\mathbb{P}^*_{\mathcal{S}}(\mathcal{U}(\mathcal{S}, \epsilon, \alpha) \text{ implies a probabilistic guarantee at level } \epsilon \text{ for } \mathbb{P}^*) \\
&\qquad = \mathbb{P}^*_{\mathcal{S}}(\delta^*(\mathbf{v}|\ \mathcal{U}(\mathcal{S}, \epsilon, \alpha)) \geq \mathrm{VaR}_\epsilon^{\mathbb{P}^*}(\mathbf{v})\ \forall \mathbf{v} \in \mathbb{R}^d) && \text{(Theorem 5.1)} \\
&\qquad \geq \mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^* \in \mathcal{P}(\mathcal{S}, \epsilon, \alpha)) && \text{(Step 2 of schema)} \\
&\qquad \geq 1 - \alpha && \text{(Confidence region).}
\end{aligned}$$

□

Proof. Proof of Theorem 5.4. For the first part,

$$\begin{aligned}
&\mathbb{P}^*_{\mathcal{S}}(\{\mathcal{U}(\mathcal{S},\epsilon,\alpha):\ 0<\epsilon<1\} \text{ simultaneously implies a probabilistic guarantee})\\
&\quad= \mathbb{P}^*_{\mathcal{S}}(\delta^*(\mathbf{v}|\ \mathcal{U}(\mathcal{S},\epsilon,\alpha)) \geq \mathrm{VaR}^{\mathbb{P}^*}_{\epsilon}(\mathbf{v})\ \forall \mathbf{v}\in\mathbb{R}^d,\ 0<\epsilon<1) && \text{(Theorem 5.1)}\\
&\quad\geq \mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^*\in \bigcap_{\epsilon:0\leq\epsilon\leq 1} \mathcal{P}(\mathcal{S},\epsilon,\alpha)) && \text{(Step 2 of schema)}\\
&\quad= \mathbb{P}^*_{\mathcal{S}}(\mathbb{P}^*\in \mathcal{P}(\mathcal{S},\alpha)) && (\mathcal{P}(\mathcal{S},\alpha)=\mathcal{P}(\mathcal{S},\epsilon,\alpha))\\
&\quad\geq 1-\alpha && \text{(Confidence region).}
\end{aligned}$$

For the second part, let $\epsilon_1,\ldots,\epsilon_m$ denote any feasible ϵ_j's in (5.9).

$$\begin{aligned}
1-\alpha &\leq \mathbb{P}^*_{\mathcal{S}}(\{\mathcal{U}(\mathcal{S},\epsilon,\alpha):0<\epsilon<1\} \text{ simultaneously implies a probabilistic guarantee})\\
&\leq \mathbb{P}^*_{\mathcal{S}}(\mathcal{U}(\mathcal{S},\epsilon_j,\alpha) \text{ implies a probabilistic guarantee at level } \epsilon_j, j=1,\ldots,m).
\end{aligned}$$

Applying the union-bound and Theorem 5.2 yields the result. □

To prove Theorem 5.5 we require the following well-known result.

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

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

Then, $\delta^*(\mathbf{v}|\ \mathcal{U}^{CVaR^{\mathbb{P}}_{\epsilon}})=CVaR^{\mathbb{P}}(\mathbf{v})$.

Proof of Theorem 5.5: We prove the theorem for $\mathcal{U}^{\chi^2}_{\epsilon}$. The proof for $\mathcal{U}^{G}_{\epsilon}$ is similar. From Thm. 5.2, it suffices to show that $\delta^*(\mathbf{v}|\ \mathcal{U}^{\chi^2}_{\epsilon})$ is an upper bound to $\sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}}\mathrm{VaR}^{\mathbb{P}}_{\epsilon}(\mathbf{v})$:

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}}\mathrm{VaR}^{\mathbb{P}}_{\epsilon}(\mathbf{v}) &\leq \sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}}\mathrm{CVaR}^{\mathbb{P}}_{\epsilon}(\mathbf{v}) && \text{(CVaR is an upper bound to VaR)}\\
&= \sup_{\mathbb{P}\in\mathcal{P}^{\chi^2}}\max_{\mathbf{u}\in\mathcal{U}^{\mathrm{CVaR}^{\mathbb{P}}_{\epsilon}}}\mathbf{u}^T\mathbf{v} && \text{(Thm. D.1)}\\
&= \max_{\mathbf{u}\in\mathcal{U}^{\chi^2}_{\epsilon}}\mathbf{u}^T\mathbf{v} && \text{(Combining Eqs. (5.13) and (5.10)).}
\end{aligned}$$

To obtain the expression for $\delta^*(\mathbf{v}|\ \mathcal{U}^{\chi^2}_{\epsilon})$ observe,

$$\delta^*(\mathbf{v}|\ \mathcal{U}^{\chi^2}_{\epsilon})=\inf_{\mathbf{w}\geq 0}\left\{\max_{\mathbf{q}\in\Delta_n}\sum_{i=0}^{n-1}q_i(\mathbf{a}_i^T\mathbf{v}-w_i)+\frac{1}{\epsilon}\max_{\mathbf{p}\in\mathcal{P}^{\chi^2}}\mathbf{w}^T\mathbf{p}\right\},$$

from Lagrangian duality. The optimal value of the first max is $\beta=\max_i\mathbf{a}_i^T\mathbf{v}-w_i$. The second max is of the form studied in (Ben-Tal et al. 2013, Corollary 1) and has

optimal value

$$\eta + \frac{\lambda \chi^2_{n-1,1-\alpha}}{N} + 2\lambda - 2\sum_{i=0}^{n-1} \hat{p}_i \sqrt{\lambda}\sqrt{\lambda + \eta - w_i}.$$

Using the second-order cone representation of the hyperbolic constraint $s_i^2 \leq \lambda \cdot (\lambda + \eta - w_i)$ (Lobo et al. 1998) and simplifying we obtain the result. □

Proof of Proposition 5.9. 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,

$$D(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{j=0}^{n-1} p_j \left(\Delta_j + \frac{\Delta_j^2}{2} + O(\Delta_j^3)\right) = \sum_{j=0}^{n-1} \frac{(\hat{p}_j - p_j)^2}{2p_j} + \sum_{j=0}^{n-1} O(\Delta_j^3), \tag{D.2}$$

where the last eqDelageYe10uality follows by expanding out terms and observing that $\sum_{j=0}^{n-1} \hat{p}_j = \sum_{j=0}^{n-1} p_j = 1$. Next, note $\mathbf{p} \in \mathcal{P}^G \implies \hat{p}_j/p_j \leq \exp(\frac{\chi^2_{n-1,1-\alpha}}{2N\hat{p}_j})$. From the Strong Law of Large Numbers, for any $0 < \alpha' < 1$, there exists M such that $\hat{p}_j \geq p_j^*/2$ with probability at least $1 - \alpha'$ for all $j = 0, \ldots, n-1$, simultaneously. It follows that for N sufficiently large, with probability $1 - \alpha'$, $\mathbf{p} \in \mathcal{P}^G \implies \hat{p}_j/p_j \leq \exp(\frac{\chi^2_{n-1,1-\alpha}}{Np_j^*})$ which implies that $|\Delta_j| \leq \exp(\frac{\chi^2_{n-1,1-\alpha}}{Np_j^*}) - 1 = O(N^{-1})$. Substituting into (D.2) completes the proof. □

To prove Theorem 5.11 we first prove the following auxiliary result that will allow us to evaluate the inner supremum in (5.17).

Theorem D.2. *Suppose $g(u)$ is monotonic. Then,*

$$\sup_{\mathbb{P}_i \in \mathcal{P}_i^{KS}} \mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] = \max\left(\sum_{j=0}^{N+1} q_j^L(\Gamma^{KS}) g(\hat{u}_i^{(j)}), \sum_{j=0}^{N+1} q_j^R(\Gamma^{KS}) g(\hat{u}_i^{(j)})\right) \tag{D.3}$$

Proof. Observe that the discrete distribution which assigns mass $q_j^L(\Gamma^{KS})$ (resp. $q_j^R(\Gamma^{KS})$) to the point $\hat{u}^{(j)}$ for $j = 0, \ldots, N+1$ is an element of $\mathcal{P}_i^{KS}$. Thus, Eq. (D.3) holds with "=" replaced by "$\geq$".

For the reverse inequality, we have two cases. Suppose first that $g(u_i)$ is non-decreasing. Given $\mathbb{P}_i \in \mathcal{P}_i^{KS}$, consider the measure $\mathbb{Q}$ defined by

$$\begin{aligned} &\mathbb{Q}(\tilde{u}_i = \hat{u}_i^{(0)}) \equiv 0, \quad \mathbb{Q}(\tilde{u}_i = \hat{u}_i^{(1)}) \equiv \mathbb{P}_i(\hat{u}_i^{(0)} \leq \tilde{u}_i \leq \hat{u}_i^{(1)}), \\ &\mathbb{Q}(\tilde{u}_i = \hat{u}_i^{(j)}) \equiv \mathbb{P}_i(\hat{u}_i^{(j-1)} < \tilde{u}_i \leq \hat{u}_i^{(j)}), \quad j = 2, \ldots, N+1. \end{aligned} \tag{D.4}$$

Then, $\mathbb{Q} \in \mathcal{P}^{KS}$, and since $g(u_i)$ is non-decreasing, $\mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] \leq \mathbb{E}^{\mathbb{Q}}[g(\tilde{u}_i)]$. Thus, the measure attaining the supremum on the left-hand side of Eq. (D.3) has discrete support $\{\hat{u}_i^{(0)}, \ldots, \hat{u}_i^{(N+1)}\}$, and the supremum is equivalent to the linear optimization

problem:

$$\begin{aligned}\max_{\mathbf{p}} \quad & \sum_{j=0}^{N+1} p_j g(\hat{u}^{(j)}) \\ \text{s.t.} \quad & \mathbf{p} \geq \mathbf{0}, \quad \mathbf{e}^T\mathbf{p} = 1, \\ & \sum_{k=0}^{j} p_k \geq \frac{j}{N} - \Gamma^{KS}, \quad \sum_{k=j}^{N+1} p_k \geq \frac{N-j+1}{N} - \Gamma^{KS}, \quad j = 1, \ldots, N,\end{aligned} \tag{D.5}$$

(We have used the fact that $\mathbb{P}_i(\tilde{u}_i < \hat{u}_i^{(j)}) = 1 - \mathbb{P}_i(\tilde{u} \geq \hat{u}_i^{(j)})$.) 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 $\mathbf{y} = \mathbf{0}$, $t = g(\hat{u}_i^{(N+1)})$ and

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

constitute a primal-dual optimal pair. This proves (D.3) when g is non-decreasing. The case of $g(u_i)$ non-increasing is similar. □

Proof of Theorem 5.11. Notice by Theorem D.2, Eq. (5.17) is equivalent to the given expression for $\delta^*(\mathbf{v}|\, \mathcal{U}_\epsilon^I)$. By our schema, it suffices to show then that this expression is truly the support function of $\mathcal{U}_\epsilon^I$. By Lagrangian duality,

$$\delta^*(\mathbf{v}|\, \mathcal{U}_\epsilon^I) = \inf_{\lambda \geq 0} \left(\begin{array}{c} \lambda \log(1/\epsilon) + \max\limits_{\mathbf{q},\boldsymbol{\theta}} \sum\limits_{i=1}^{d} v_i \sum\limits_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i - \lambda \sum\limits_{i=1}^{d} D(\mathbf{q}^i, \theta_i \mathbf{q}^L + (1-\theta_i)\mathbf{q}^R) \\ \text{s.t.} \mathbf{q}^i \in \Delta_{N+2}, 0 \leq \theta_i \leq 1, \quad i = 1, \ldots, d. \end{array} \right)$$

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

$$\max_{\theta_i \in [0,1]} \lambda \left\{ \max_{\mathbf{q}_i \in \Delta_{N+2}} \left\{ \sum_{j=0}^{N+1} \frac{v_i \hat{u}_i^{(j)}}{\lambda} q_{ij} - D(\mathbf{q}^i, \theta_i \mathbf{q}^L + (1-\theta_i)\mathbf{q}^R) \right\} \right\}.$$

The inner maximization can be solved analytically (Boyd and Vandenberghe 2004, pg. 93), yielding:

$$q_j^i = \frac{p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda}}{\sum_{j=0}^{N+1} p_j^i e^{v_i \hat{u}_i^{(j)}/\lambda}}, \quad p_j^i = \theta_i q_j^L(\Gamma^{KS}) + (1-\theta_i) q_j^R(\Gamma^{KS}). \tag{D.6}$$

Substituting in this solution and recombining subproblems yields

$$\lambda \log(1/\epsilon) + \lambda \sum_{i=1}^{d} \log \left(\max_{\theta_i \in [0,1]} \sum_{j=0}^{N+1} (\theta_i q_j^L(\Gamma^{KS}) + (1-\theta_i) q_j^R(\Gamma^{KS})) e^{v_i \hat{u}_i^{(j)}/\lambda}. \right) \tag{D.7}$$

The inner optimizations over θ_i are all linear, and hence achieve an optimal solution at one of the end points, i.e., either $\theta_i = 0$ or $\theta_i = 1$. This yields the given expression for $\delta^*(\mathbf{v}|\,\mathcal{U})$.

Following this proof backwards to identify the optimal $\mathbf{q}^i$, and, thus, $\mathbf{u} \in \mathcal{U}^I$ also proves the validity of the procedure given in Remark 5.13 □

Proof Theorem 5.15. By inspection, (5.26) is the worst-case value of (5.24) over $\mathcal{P}^{FB}$. By Theorem 5.4, it suffices to show that this expression truly is the support function of $\mathcal{U}_\epsilon^{FB}$. First observe

$$\max_{\mathbf{u} \in \mathcal{U}_\epsilon^{FB}} \mathbf{u}^T \mathbf{v} = \min_{\lambda \geq 0} \left\{ \lambda \log(1/\epsilon) + \max_{\substack{\mathbf{m}_b \leq \mathbf{y}_1 \leq \mathbf{m}_b, \\ \mathbf{y}_2 \geq, \mathbf{y}_3 \geq \mathbf{0}}} \sum_{i=1}^{d} v_i (y_{1i} + y_{2i} - y_{3i}) - \lambda \sum_{i=1}^{d} \frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2} + \frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2} \right\}$$

by Lagrangian strong duality. The inner maximization decouples by i. The i^{th} subproblem further decouples into three sub-subproblems. The first is $\max_{m_{bi} \leq y_{i1} \leq m_{fi}} v_i y_{1i}$ with optimal solution

$$y_{1i} = \begin{cases} m_{fi} & \text{if } v_i \geq 0, \\ m_{bi} & \text{if } v_i < 0. \end{cases}$$

The second sub-subproblem is $\max_{y_{2i} \geq 0} v_i y_{2i} - \lambda \frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2}$. This is maximizing a concave quadratic function of one variable. Neglecting the non-negativity constraint, the optimum occurs at $y_{2i}^* = \frac{v_i \sigma_{fi}^2}{\lambda}$. If this value is negative, the optimum occurs at $y_{2i}^* = 0$. Consequently,

$$\max_{y_{2i} \geq 0} \; v_i y_{2i} - \lambda \frac{y_{2i}^2}{2\overline{\sigma}_{fi}^2} = \begin{cases} \frac{v_i \sigma_{fi}^2}{2\lambda} & \text{if } v_i \geq 0, \\ 0 & \text{if } v_i < 0. \end{cases}$$

Similarly, we can show that the third subproblem has the following optimum value

$$\max_{y_{3i} \geq 0} \; -v_i y_{3i} - \lambda \frac{y_{3i}^2}{2\overline{\sigma}_{bi}^2} = \begin{cases} \frac{v_i \sigma_{bi}^2}{2\lambda} & \text{if } v_i \leq 0, \\ 0 & \text{if } v_i > 0. \end{cases}$$

Combining the three sub-subproblems yields

$$\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{FB}) = \sum_{i:v_i>0} v_i m_{fi} + \sum_{i:v_i \leq 0} v_i m_{bi} + \min_{\lambda \geq 0} \lambda \log(1/\epsilon) + \frac{1}{2\lambda} \left(\sum_{i:v_i>0} v_i^2 \overline{\sigma}_{fi}^2 + \sum_{i:v_i \leq 0} v_i^2 \overline{\sigma}_{bi}^2 \right).$$

This optimization can be solved closed-form, yielding

$$\lambda^* = \sqrt{\frac{\sum_{i:v_i>0} v_i^2 \overline{\sigma}_{fi}^2 + \sum_{i:v_i\le 0} v_i^2 \overline{\sigma}_{bi}^2}{2\log(1/\epsilon)}}.$$

Simplifying yields the right hand side of (5.26). Moreover, following the proof backwards to identify the maximizing $\mathbf{u} \in \mathcal{U}_\epsilon^{FB}$ proves the validity of the procedure given in Remark 5.16. □

Proof of Theorem 5.18. Observe,

$$\sup_{\mathbb{P}\in\mathcal{P}^M} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) \le \sup_{\mathbb{P}\in\mathcal{P}^M} \sum_{i=1}^d \mathrm{VaR}_{\epsilon/d}^{\mathbb{P}}(v_i\mathbf{e}_i) = \sum_{i:v_i>0} v_i \hat{u}_i^{(s)} + \sum_{i:v_i\le 0} v_i \hat{u}_i^{(N-s+1)}, \tag{D.8}$$

where the equality follows rom the positive homogeneity of $\mathrm{VaR}_\epsilon^{\mathbb{P}}$, and this last expression is equivalent to (5.31) because $\hat{u}_i^{(N-s+1)} \le \hat{u}_i^{(s)}$. By Theorem 5.2, it suffices to show that $\delta^*(\mathbf{v}|\,\mathcal{U}^M)$ truly is the support function of $\mathcal{U}_\epsilon^M$, and this is immediate. □

Proof of Theorem 5.21. We first compute $\sup_{\mathbb{P}\in\mathcal{P}^{LCX}} \mathbb{P}(\mathbf{v}^T\tilde{\mathbf{u}} > t)$ for fixed $\mathbf{v}, t$. In this spirit of Shapiro (2001), Bertsimas et al. (2014b), this optimization admits the following strong dual:

$$\begin{aligned}
\inf_{\theta, w_\sigma, \lambda(\mathbf{a},b)} \quad & \theta + \left(\frac{1}{N}\sum_{j=1}^N \|\hat{\mathbf{u}}_j\|^2 - \Gamma_\sigma\right) w_\sigma + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \\
\text{s.t.} \quad & \theta - w_\sigma\|\mathbf{u}\|^2 + \int_{\mathcal{B}} (\mathbf{a}^T\mathbf{u} - b)^+ d\lambda(a,b) \ge \mathbb{I}(\mathbf{u}^T\mathbf{v} > t) \quad \forall \mathbf{u} \in \mathbb{R}^d, \\
& w_\sigma \ge 0, \quad d\lambda(\mathbf{a},b) \ge 0,
\end{aligned} \tag{D.9}$$

where $\Gamma(\mathbf{a},b) \equiv \frac{1}{N}\sum_{j=1}^N (\mathbf{a}^T\hat{\mathbf{u}}_j - b)^+ + \Gamma_{LCX}$. We claim that $w_\sigma = 0$ in any feasible solution. Indeed, suppose $w_\sigma > 0$ in some feasible solution. Note $(\mathbf{a},b) \in \mathcal{B}$ implies that $(\mathbf{a}^T\mathbf{u} - b)^+ = O(\|\mathbf{u}\|)$ as $\|\mathbf{u}\| \to \infty$. Thus, the left-hand side of eq. (D.9) tends to $-\infty$ as $\|\mathbf{u}\| \to \infty$ while the right-hand side is bounded below by zero. This contradicts the feasibility of the solution.

Since $w_\sigma = 0$ in any feasible solution, rewrite the above as

$$\begin{aligned}
\inf_{\theta, \lambda(\mathbf{a},b)} \quad & \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \\
\text{s.t.} \quad & \theta + \int_{\mathcal{B}} (\mathbf{a}^T\mathbf{u} - b)^+ d\lambda(a,b) \ge 0 \quad \forall \mathbf{u} \in \mathbb{R}^d, \\
& \theta + \int_{\mathcal{B}} (\mathbf{a}^T\mathbf{u} - b)^+ d\lambda(a,b) \ge 1 \quad \forall \mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^d : \mathbf{u}^T\mathbf{v} > t\}, \\
& d\lambda(\mathbf{a},b) \ge 0.
\end{aligned} \tag{D.10}$$

The two infinite constraints can be rewritten using duality. Specifically, the first

constraint is

$$-\theta \leq \min_{s(\mathbf{a},b)\geq 0, \tilde{\mathbf{u}}\in\mathbb{R}^d} \int_{\mathcal{B}} s(\mathbf{a},b) d\lambda(\mathbf{a},b)$$
$$\text{s.t.} \quad s(\mathbf{a},b) \geq (\mathbf{a}^T\tilde{\mathbf{u}} - b) \quad \forall (\mathbf{a},b) \in \mathcal{B},$$

which admits the dual:

$$-\theta \leq \max_{y_1(\mathbf{a},b)} \quad -\int_{\mathcal{B}} b \; dy_1(\mathbf{a},b)$$
$$\text{s.t.} \quad 0 \leq dy_1(\mathbf{a},b) \leq d\lambda(\mathbf{a},b) \quad \forall (\mathbf{a},b) \in \mathcal{B},$$
$$\int_{\mathcal{B}} \mathbf{a} \; dy_1(\mathbf{a},b) = 0.$$

The second constraint can be treated similarly using continuity to take the closure of $\{\mathbf{u} \in \mathbb{R}^d : \mathbf{u}^T\mathbf{v} > t\}$. Combining both constraints yields the equivalent representation of (D.10)

$$\begin{aligned}
\inf_{\substack{\theta,\tau,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} \quad & \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \\
\text{s.t.} \quad & \theta - \int_{\mathcal{B}} b \; dy_1(\mathbf{a},b) \geq 0, \quad \theta + t\tau - \int_{\mathcal{B}} b \; dy_2(\mathbf{a},b) \geq 1, \\
& 0 \leq dy_1(\mathbf{a},b) \leq d\lambda(\mathbf{a},b) \quad \forall (\mathbf{a},b) \in \mathcal{B}, \\
& 0 \leq dy_2(\mathbf{a},b) \leq d\lambda(\mathbf{a},b) \quad \forall (\mathbf{a},b) \in \mathcal{B}, \\
& \int_{\mathcal{B}} \mathbf{a} \; dy_1(\mathbf{a},b) = 0, \quad \tau\mathbf{v} = \int_{\mathcal{B}} \mathbf{a} \; dy_2(\mathbf{a},b), \\
& \tau \geq 0.
\end{aligned} \tag{D.11}$$

Now the worst-case Value at Risk can be written as

$$\sup_{\mathbb{P}\in\mathcal{P}^{LCX}} \text{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) = \inf_{\substack{\theta,\tau,t,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} t$$
$$\text{s.t.} \quad \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \leq \epsilon,$$
$$(\theta, \tau, \lambda(\mathbf{a},b), y_1(\mathbf{a},b), y_2(\mathbf{a},b), t) \text{ feasible in (D.11)} .$$

We claim that $\tau > 0$ in an optimal solution. Suppose to the contrary that $\tau = 0$ in some solution. Let $t \to -\infty$ in this solution. The resulting solution remains feasible, implying that $\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > -\infty) \leq \epsilon$ for all $\mathbb{P} \in \mathcal{P}^{LCX}$. However, the empirical distribution $\hat{\mathbb{P}} \in \mathcal{P}^{LCX}$, a contradiction.

Since $\tau > 0$, apply the transformation

$$(\theta/\tau, 1/\tau, \lambda(\mathbf{a},b)/\tau, \mathbf{y}(\mathbf{a},b)/\tau) \to (\theta, \tau, \lambda(\mathbf{a},b), \mathbf{y}(\mathbf{a},b))$$

yielding

$$\begin{aligned}
\inf_{\substack{\theta,\tau,t,\lambda(\mathbf{a},b),\\ y_1(\mathbf{a},b),y_2(\mathbf{a},b)}} \quad & t \\
\text{s.t.} \quad & \theta + \int_{\mathcal{B}} \Gamma(\mathbf{a},b) d\lambda(\mathbf{a},b) \leq \epsilon\tau \\
& \theta - \int_{\mathcal{B}} b \; dy_1(\mathbf{a},b) \geq 0, \quad \theta + t - \int_{\mathcal{B}} b \; dy_2(\mathbf{a},b) \geq \tau, \\
& 0 \leq dy_1(\mathbf{a},b) \leq d\lambda(\mathbf{a},b) \quad \forall (\mathbf{a},b) \in \mathcal{B}, \\
& 0 \leq dy_2(\mathbf{a},b) \leq d\lambda(\mathbf{a},b) \quad \forall (\mathbf{a},b) \in \mathcal{B}, \\
& \int_{\mathcal{B}} \mathbf{a} \; dy_1(\mathbf{a},b) = 0, \quad \mathbf{v} = \int_{\mathcal{B}} \mathbf{a} \; dy_2(\mathbf{a},b), \\
& \tau \geq 0.
\end{aligned}$$

Eliminate the variable t, and make the transformation $(\tau\epsilon, \theta - \int_{\mathcal{B}} b dy_1(\mathbf{a},b)) \to (\tau, \theta)$ to yield the righthand side of (5.34).

By Theorem 5.4, it suffices to show that the right hand side of (5.34) is indeed the support function of $\mathcal{U}_\epsilon^{LCX}$. Take the dual of (5.34) and simplify to yield the given description of $\mathcal{U}_\epsilon^{LCX}$.

□

Proof of Theorem 5.26. By Theorem 5.4, it suffices to show that $\delta^*(\mathbf{v}|\, \mathcal{U}_\epsilon^{CS})$ is given by (5.36), which follows immediately from two applications of the Cauchy-Schwartz inequality. □

To prove Theorem 5.31 we require the following proposition.

Proposition D.3.

$$\sup_{\mathbb{P} \in \mathcal{P}^{DY}} \mathbb{P}(\tilde{\mathbf{u}}^T \mathbf{v} > t) \tag{D.12}$$

$$\begin{aligned}
= \min_{r,s,\theta,\mathbf{y}_1,\mathbf{y}_2,\mathbf{Z}} \quad & r + s \qquad (D.13) \\
s.t. \quad & \begin{pmatrix} r + \mathbf{y}_1^{+T} \hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^{-T} \hat{\mathbf{u}}^{(N+1)} & \frac{1}{2}(\mathbf{q} - \mathbf{y}_1)^T, \\ \frac{1}{2}(\mathbf{q} - \mathbf{y}_1) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& \begin{pmatrix} r + \mathbf{y}_2^{+T} \hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^{-T} \hat{\mathbf{u}}^{(N+1)} + \theta t - 1 & \frac{1}{2}(\mathbf{q} - \mathbf{y}_2 - \theta\mathbf{v})^T, \\ \frac{1}{2}(\mathbf{q} - \mathbf{y}_2 - \theta\mathbf{v}) & \mathbf{Z} \end{pmatrix} \succeq \mathbf{0}, \\
& s \geq (\gamma_2^B \hat{\mathbf{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T \mathbf{q} + \sqrt{\gamma_1^B} \|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
& \mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-, \quad \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-, \quad \mathbf{y}_1^+, \mathbf{y}_1^-, \mathbf{y}_2^+, \mathbf{y}_2^- \theta \geq \mathbf{0}.
\end{aligned}$$

Proof. We claim that $\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > t)$ has the following dual representation:

$$\begin{aligned}
\min_{r,s,\mathbf{q},\mathbf{Z},\mathbf{y}_1,\mathbf{y}_2,\theta} \quad & r+s \\
\text{s.t.} \quad & r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 0 \quad \forall \mathbf{u} \in [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}], \\
& r + \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \geq 1 \quad \forall \mathbf{u} \in [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}] \cap \{\mathbf{u} : \mathbf{u}^T\mathbf{v} > t\}, \\
& s \geq (\gamma_2^B\hat{\mathbf{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T) \circ \mathbf{Z} + \hat{\boldsymbol{\mu}}^T\mathbf{q} +, \sqrt{\gamma_1^B}\|\mathbf{q} + 2\mathbf{Z}\hat{\boldsymbol{\mu}}\|_{\hat{\mathbf{\Sigma}}^{-1}}, \\
& \mathbf{Z} \succeq \mathbf{0}.
\end{aligned} \tag{D.14}$$

See the proof of Lemma 1 in Delage and Ye (2010) for details. Since $\mathbf{Z}$ is positive semidefinite, we can use strong duality to rewrite the two semi-infinite constraints:

$$\begin{aligned}
& \min_{\mathbf{u}} \quad \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \\
& \text{s.t.} \quad \hat{\mathbf{u}}^{(0)} \leq \mathbf{u} \leq \hat{\mathbf{u}}^{(N+1)}, \\
& \qquad\qquad \Longleftrightarrow \\
\max_{\mathbf{y}_1,\mathbf{y}_1^+,\mathbf{y}_1^-} \quad & -\frac{1}{4}(\mathbf{q}-\mathbf{y}_1)^T\mathbf{Z}^{-1}(\mathbf{q}-\mathbf{y}_1) + \mathbf{y}_1^+\hat{\mathbf{u}}^{(0)} - \mathbf{y}_1^-\hat{\mathbf{u}}^{(N+1)} \\
\text{s.t.} \quad & \mathbf{y}_1 = \mathbf{y}_1^+ - \mathbf{y}_1^-, \quad \mathbf{y}_1^+, \mathbf{y}_1^- \geq \mathbf{0},
\end{aligned}$$

$$\begin{aligned}
& \min_{\mathbf{u}} \quad \mathbf{u}^T\mathbf{Z}\mathbf{u} + \mathbf{u}^T\mathbf{q} \\
& \text{s.t.} \quad \hat{\mathbf{u}}^{(0)} \leq \mathbf{u} \leq \hat{\mathbf{u}}^{(N+1)}, \\
& \qquad \mathbf{u}^T\mathbf{v} \geq t, \\
& \qquad\qquad \Longleftrightarrow \\
\max_{\mathbf{y}_2,\mathbf{y}_2^+,\mathbf{y}_2^-} \quad & -\frac{1}{4}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{v})^T\mathbf{Z}^{-1}(\mathbf{q}-\mathbf{y}_2-\theta\mathbf{v}) + \mathbf{y}_2^+\hat{\mathbf{u}}^{(0)} - \mathbf{y}_2^-\hat{\mathbf{u}}^{(N+1)} + \theta t \\
\text{s.t.} \quad & \mathbf{y}_2 = \mathbf{y}_2^+ - \mathbf{y}_2^-, \quad \mathbf{y}_2^+, \mathbf{y}_2^- \geq \mathbf{0}, \ \theta \geq 0.
\end{aligned}$$

Then, by using Schur-Complements, we can rewrite Problem (D.14) as in the proposition. □

We can now prove the theorem.

Proof of Theorem 5.31. Using Proposition D.3, we can characterize the worst-case VaR by

$$\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \text{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v}) = \inf\left\{t : r+s \leq \epsilon, (r,s,t,\theta,\mathbf{y}_1,\mathbf{y}_2,\mathbf{Z}) \text{ are feasible in problem (D.12)}\right\}. \tag{D.15}$$

We claim that $\theta > 0$ in any feasible solution to the infimum in Eq. (D.15). Suppose to the contrary that $\theta = 0$. Then this solution is also feasible as $t \downarrow \infty$, which

implies that $\mathbb{P}(\tilde{\mathbf{u}}^T\mathbf{v} > -\infty) \leq \epsilon$ for all $\mathbb{P} \in \mathcal{P}^{DY}$. On the other hand, the empirical distribution $\hat{\mathbb{P}} \in \mathcal{P}^{DY}$, a contradiction.

Since $\theta > 0$, we can rescale all of the above optimization variables in problem (D.12) by θ. Substituting this into Eq. (D.15) yields the given expression for $\sup_{\mathbb{P}\in\mathcal{P}^{DY}} \mathrm{VaR}_\epsilon^{\mathbb{P}}(\mathbf{v})$. Rewriting this optimization problem as a semidefinite optimization problem and taking its dual yields $\mathcal{U}_\epsilon^{DY}$ in the theorem. By Theorem 5.4, this set simultaneously implies a probabilistic guarantee. □

Proof of Theorem 5.34. For each part, the convexity in $(\mathbf{v}, t)$ is immediate since $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon)$ is a support function of a convex set. For the first part, note that from the second part of Theorem 5.26, $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) \leq t$ will be convex in ϵ for a fixed $(\mathbf{v}, t)$ whenever $\sqrt{1/\epsilon - 1}$ is convex. Examining the second derivative of this function, this occurs on the interval $0 < \epsilon < .75$. Similarly, for the second part, note that from the second part of Theorem 5.15, $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{FB}) \leq t$ will be convex in ϵ for a fixed $(\mathbf{v}, t)$ whenever $\sqrt{2\log(1/\epsilon)}$ is convex. Examining the second derivative of this function, this occurs on the interval $0 < \epsilon < 1\sqrt{e}$.

From the representations of $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{\chi^2})$ and $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{G})$ in Theorem 5.5, we can see they will be convex in ϵ whenever $1/\epsilon$ is convex, i.e., $0 < \epsilon < 1$. From the representation of $\delta^*(\mathbf{v}|\mathcal{U}_\epsilon^{I})$ in Theorem 5.11 and since $\lambda \geq 0$, we see this function will be convex in ϵ whenever $\log(1/\epsilon)$ is convex, i.e., $0 < \epsilon < 1$.

Finally, examining the support functions of $\mathcal{U}_\epsilon^{LCX}$ and $\mathcal{U}_\epsilon^{DY}$ shows that ϵ occurs linearly in each of these functions. □

D.2 Omitted Figures

This section contains additional figures omitted from the main text.

D.3 Optimizing ϵ_j's for Multiple Constraints

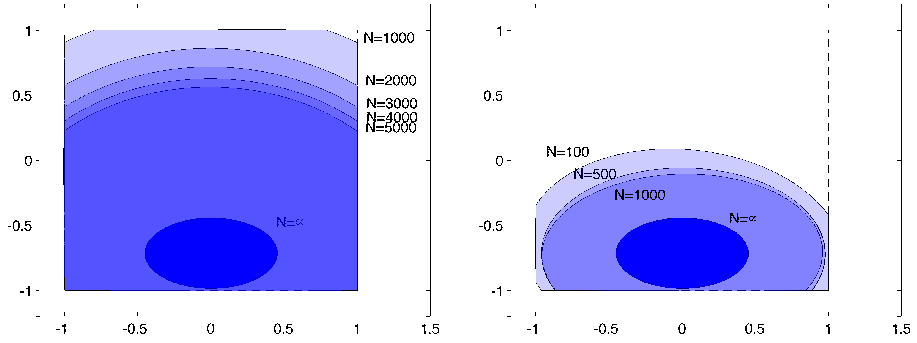
In this section we specify the optimization problem that we solve in ϵ_j's as part of our alternating optimization heuristic for treating multiple constraints. We first present our approach using m constraints of the form $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS}) \leq t$. Without loss of generality, assume the overall optimization problem is a minimization. Consider the j^{th} constraint, and let $(\mathbf{v}', t')$ denote the subset of the solution to the original optimization problem at the current iterate pertaining to the j^{th} constraint. Let ϵ_j', $j = 1, \ldots, m$ denote the current iterate in ϵ. Finally, let λ_j denote the shadow price of the j^{th} constraint in the overall optimization problem.

Notice from the second part of Theorem 5.26 that $\delta^*(\mathbf{v}|\ \mathcal{U}_\epsilon^{CS})$ is decreasing in ϵ. Thus, for all $\epsilon_j \geq \underline{\epsilon}_j$, $\delta^*(\mathbf{v}'|\ \mathcal{U}_{\epsilon_j}^{CS}) \leq t'$, where,

$$\underline{\epsilon}_j \equiv \left[\frac{(t' - \hat{\boldsymbol{\mu}}^T\mathbf{v}' - \Gamma_1\|\mathbf{v}'\|_2)^2}{\mathbf{v}'^T(\boldsymbol{\Sigma} + \Gamma_2\mathbf{I})\mathbf{v}'} + 1\right]^{-1}.$$

Motivated by the shadow-price λ_j, we define the next iterates of ϵ_j, $j = 1, \ldots, m$

Figure D-1: $\mathcal{U}_\epsilon^{CS}$ With and Without Bootstrapping for the Example from Fig. 5-3



(a) Not Bootstrapped (b) Bootstrapped

Note: $N_B = 10,000$, $\alpha = 10\%$, $\epsilon = 10\%$. Notice that for $N = 1,000$, the non-bootstrapped set is almost as big as the full support and shrinks slowly to its infinite limit. The bootstrapped set with $N = 100$ points is smaller than the non-bootstrapped version with 50 times as many points.

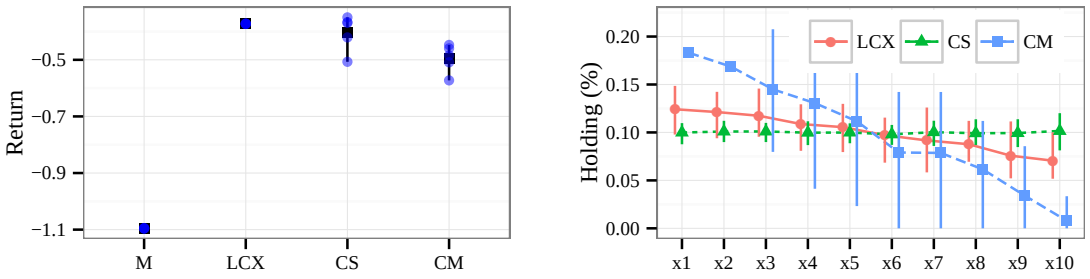
to be the solution of the linear optimization problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\epsilon}} & -\sum\limits_{j=1}^{d}\left(\dfrac{\sqrt{\mathbf{v}'^T(\boldsymbol{\Sigma}+\Gamma_2\mathbf{I})\mathbf{v}'}}{2\epsilon'^2\sqrt{\frac{1}{\epsilon'}-1}}\right)\lambda_j\cdot\epsilon_j \\ \text{s.t.} & \underline{\epsilon}_j \le \epsilon_j \le .75, \quad j=1,\ldots,m, \\ & \sum\limits_{j=1}^{m}\epsilon_j \le \bar{\epsilon}, \quad \|\boldsymbol{\epsilon}'-\boldsymbol{\epsilon}\|_1 \le \kappa. \end{array} \tag{D.16}$$

The coefficient of ϵ_j in the objective function is $\lambda_j \cdot \partial_{\epsilon_j}\delta^*(\mathbf{v}'|\,\mathcal{U}_{\epsilon_j}^{CS})$ which is intuitively a first-order approximation to the improvement in the overall optimization problem for a small change in ϵ_j. The norm constraint on $\boldsymbol{\epsilon}$ ensures that the next iterate is not too far away from the current iterate, so that the shadow-price λ_j remains a good approximation. (We use $\kappa = .05$ in our experiments.) The upper bound ensures that we remain in a region where $\delta^*(\mathbf{v}|\,\mathcal{U}_{\epsilon_j}^{CS})$ is convex in ϵ_j. Finally, the lower bounds on ϵ_j ensure that the previous iterate of the original optimization problem $(\mathbf{v}', t')$ will still be feasible for the new values of ϵ_j. Consequently, the objective value of the original optimization problem is non-increasing. We terminate the procedure when the objective value no longer makes significant progress.

With the exception of $\mathcal{U}_\epsilon^{LCX}$, we can follow an entirely analogous procedure, simply adjusting the formulas for $\underline{\epsilon}_j$, the upper bounds, and the objective coefficient appropriately. We omit the details. Computing the relevant objective coefficient for $\mathcal{U}^{LCX}$ is more subtle. From (5.34), we require the optimal τ corresponding to $\mathbf{v}'$. This τ is dual to the constraint $z \le \frac{1}{\epsilon}$. Thus, our strategy is to evaluate $\delta^*(\mathbf{v}'|\,\mathcal{U}_{\epsilon_j'})$

Figure D-2: The Case $N = 2000$ for the Experiment Outlined in Sec. 5.11.1



Note: The left panel shows the cross-validation results. The right panel shows the average holdings by method. $\alpha = \epsilon = 10\%$.

by generating $(\mathbf{a}, b)$'s via the separation routine of Remark 5.24. At termination, we let τ be the dual variable to the constraint $z \leq \frac{1}{\epsilon}$, and, finally, we set the objective coefficient of ϵ_j in (D.16) to be $-\frac{\tau}{\epsilon_j'}^2$. Again, intuitively, this coefficient corresponds to the change in the overall optimization problem for a small change in ϵ_j.

D.4 Additional Portfolio Results

Fig. D-2 summarizes the case $N = 2000$ for the experiment outlined in Sec. 5.11.1.

D.5 Additional Queueing Results

We first derive the bound $W_n^{3,FB}$. Notice that in (5.41), the optimizing index j represents the most recent customer to arrive when the queue was empty. Let $\tilde{n}$ denote the number of customers served in a typical busy period. Intuitively, it suffices to truncate the recursion (5.41) at customer $\min(n, n^{(k)})$ where, with high probability, $\tilde{n} \leq n^{(k)}$. More formally, considering only the first half of the data $\hat{x}^1, \ldots, \hat{x}^{\lceil N/2 \rceil}$ and $\hat{t}^1, \ldots, \hat{t}^{\lceil N/2 \rceil}$, 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 α_1, we observe that with probability at least $1 - \alpha$ with respect to the sampling,

$$\mathbb{P}(\tilde{n} > \hat{n}^{(k)}) \leq 1 - \frac{k}{K} + \Gamma^{KS}(\alpha), \quad \forall k = 1, \ldots, K. \tag{D.17}$$

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

Next, calculate the constants $\mathbf{m}_f, \mathbf{m}_b, \boldsymbol{\sigma}_f, \boldsymbol{\sigma}_b$ using only the second half of the data. Then, truncate the sum in (5.46) at $\min(n, n^{(k)})$ and replace the righthand side by $\bar{\epsilon} - 1 + \frac{k}{K} - \Gamma^{KS}(\alpha/2)$. Denote the solution of this equation by $W_n^{2,FB}(k)$. Finally, let $W_n^{3,FB} \equiv \min_{1 \leq k < K} W_n^{2,FB}(k)$, obtained by grid-search.

We claim that with probability at least $1 - 2\alpha$ with respect to the sampling,

$\mathbb{P}(\tilde{W}_n > W_n^{3,FB}) \leq \bar{\epsilon}$. Namely, from our choice of parameters, eqs.(5.46) and (D.17) hold simultaneously with probability at least $1-2\alpha$. Restrict attention to a sample path where these equations hold. Since (D.17) holds for the optimal index k^*, recursion (5.41) truncated at $n^{(k^*)}$ is valid with probability at least $1-\frac{k^*}{K}+\Gamma^{KS}(\alpha)$. Finally, $\mathbb{P}(\tilde{W}_n > W_n^{3,FB}) \leq \mathbb{P}(\text{ (5.41) is invalid }) + \mathbb{P}((\tilde{W}_n > W_n^{2,FB}(k^*) \text{ and(5.41) is valid }) \leq \bar{\epsilon}$. This proves the claim.

We observe in passing that since the constants $\mathbf{m}_f, \mathbf{m}_b, \boldsymbol{\sigma}_f, \boldsymbol{\sigma}_b$ are computed using only half the data, it may not be the case that $W_n^{3,FB} < W_n^{2,FB}$, particularly for small N, but that typically $W_n^{3,FB}$ is a much stronger bound than $W_n^{2,FB}$.

Applying a similar analysis with set $\mathcal{U}_{\bar{\epsilon}}^{CS}$, yields the following bounds:

$$W_n^{1,CS} \leq \begin{cases} (\hat{\mu}_1 - \hat{\mu}_2)n + \left(\Gamma_1 + \sqrt{(\frac{n}{\bar{\epsilon}}-1)(\sigma_1^2+\sigma_2^2+2\Gamma_2)}\right)\sqrt{n} & \text{if } n < n_c^{1,CS} \\ & \text{or } \hat{\mu}_1 > \hat{\mu}_2, \\ \frac{\left(\Gamma_1+\sqrt{(\frac{n}{\bar{\epsilon}}-1)(\sigma_1^2+\sigma_2^2+2\Gamma_2)}\right)^2}{4(\hat{\mu}_2-\hat{\mu}_1)} & \text{otherwise,} \end{cases}$$

where $n_c^{1,CS} = \frac{\left(\Gamma_1+\sqrt{(\frac{n}{\bar{\epsilon}}-1)(\sigma_1^2+\sigma_2^2+2\Gamma_2)}\right)^2}{4(\hat{\mu}_1-\hat{\mu}_2)^2}$, $W_n^{2,CS}$ is the solution to

$$\sum_{j=1}^{n-1}\left[\left(\frac{W_n^{2,CS}-(\hat{\mu}_1-\hat{\mu}_2)(n-j)}{\sqrt{n-j}\sqrt{\sigma_1^2+\sigma_2^2+2\Gamma_2}} - \frac{\Gamma_1}{\sqrt{\sigma_1^2+\sigma_2^2+2\Gamma_2}}\right)^2 + 1\right]^{-1} = \bar{\epsilon}, \qquad \text{(D.18)}$$

and $W_n^{3,CS}$ defined analogously to $W_n^{3,FB}$ but using (D.18) in lieu of (5.46).

D.6 Constructing $\mathcal{U}_\epsilon^I$ from Other EDF Tests

In this section we show how to extend our constructions for $\mathcal{U}_\epsilon^I$ to other EDF tests. We consider several of the most popular, univariate goodness-of-fit, empirical distribution function test. Each test below considers the null-hypothesis $H_0 : \mathbb{P}_i^* = \mathbb{P}_{0,i}$.

Kuiper (K) Test: The K test rejects the null hypothesis at level α if

$$\max_{j=1,\ldots,N}\left(\frac{j}{N} - \mathbb{P}_{0,i}(\tilde{u}_i \leq \hat{u}_i^{(j)})\right) + \max_{j=1,\ldots,N}\left(\mathbb{P}_{0,i}(\tilde{u}_i < \hat{u}_i^{(j)}) - \frac{j-1}{N}\right) > V_{1-\alpha}.$$

Cramer von-Mises (CvM) Test: The CvM test rejects the null hypothesis at level α if

$$\frac{1}{12N^2} + \frac{1}{N}\sum_{j=1}^{N}\left(\frac{2j-1}{2N} - \mathbb{P}_{0,i}(\tilde{u}_i \leq \hat{u}_i^{(j)})\right)^2 > (T_{1-\alpha})^2.$$

Watson (W) Test: The W test rejects the null hypothesis at level α if

$$\frac{1}{12N^2} + \frac{1}{N}\sum_{j=1}^{N}\left(\frac{2j-1}{2N} - \mathbb{P}_{0,i}(\tilde{u}_i \leq \hat{u}_i^{(j)})\right)^2 - \left(\frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_{0,i}(\tilde{u}_i \leq \hat{u}_i^{(j)}) - \frac{1}{2}\right)^2 > (U_{1-\alpha})^2.$$

Anderson-Darling (AD) Test: The AD test rejects the null hypothesis at level α if

$$-1 - \sum_{j=1}^{N} \frac{2j-1}{N^2} \left(\log\left(\mathbb{P}_{0,i}(\tilde{u}_i \le \hat{u}_i^{(j)}) \right) + \log\left(1 - \mathbb{P}_{0,i}(\tilde{u}_i \le \hat{u}_i^{(N+1-j)}) \right) \right) > (A_{1-\alpha})^2$$

Tables of the thresholds above are readily available (e.g., Stephens 1974, and references therein).

As described in Bertsimas et al. (2014b), the confidence regions of these tests can be expressed in the form

$$\mathcal{P}_i^{EDF} = \{\mathbb{P}_i \in \theta[\hat{u}_i^{(0)}, \hat{u}_i^{(N+1)}] : \exists \zeta \in \mathbb{R}^N, \quad \mathbb{P}_i(\tilde{u}_i \le \hat{u}_i^{(j)}) = \zeta_i, \quad \mathbf{A}_S \boldsymbol{\zeta} - \mathbf{b}_S \in \mathcal{K}_S\},$$

where the the matrix $\mathbf{A}_S$, vector $\mathbf{b}_S$ and cone $\mathcal{K}_S$ depend on the choice of test and are given by Theorem 4.25.

Let $\mathcal{K}^*$ denote the dual cone to $\mathcal{K}$. By specializing Theorem 10 of Bertsimas et al. (2014b), we obtain the following theorem, paralleling Theorem D.2.

Theorem D.4. *Suppose $g(u)$ is monotonic and right-continuous, and let $\mathcal{P}^S$ denote the confidence region of any of the above EDF tests.*

$$\begin{aligned}
\sup_{\mathbb{P}_i \in \mathcal{P}_i^{EDF}} \mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] = \min_{\mathbf{r},\mathbf{c}} \quad & \mathbf{b}_S^T \mathbf{r} + c_{N+1} \\
s.t. \quad & -\mathbf{r} \in \mathcal{K}_S^*, \quad \mathbf{c} \in \mathbb{R}^{N+1}, \\
& (\mathbf{A}_S^T \mathbf{r})_j = c_j - c_{j+1} \quad \forall j = 1, \ldots, N, \\
& c_j \ge g(\hat{u}_i^{(j-1)}), \quad c_j \ge g(\hat{u}_i^{(j)}), \quad j = 1, \ldots, N+1. \qquad \text{(D.19)} \\
= \max_{\mathbf{z},\mathbf{q}^L,\mathbf{q}^R,\mathbf{p}} \quad & \sum_{j=0}^{N+1} p_j g(\hat{u}_i^{(j)}) \\
s.t. \quad & \mathbf{A}_S \mathbf{z} - \mathbf{b}_S \in \mathcal{K}_S, \quad \mathbf{q}^L, \mathbf{q}^R, \mathbf{p} \in \mathbb{R}_+^{N+1} \\
& q_j^L + q_j^R = z_j - z_{j-1}, \quad j = 1, \ldots, N, \qquad \text{(D.20)} \\
& q_{N+1}^L + q_{N+1}^R = 1 - z_N \\
& p_0 = q_1^L, \quad p_{N+1} = q_{N+1}^R, \quad p_j = q_{j+1}^L + q_j^R, \quad j = 1, \ldots, N,
\end{aligned}$$

where $\mathbf{A}_S, \mathbf{b}_S, \mathcal{K}_S$ are the appropriate matrix, vector and cone to the test. Moreover, when $g(u)$ is non-decreasing (resp. non-increasing), there exists an optimal solution where $\mathbf{q}^L = \mathbf{0}$ (resp. $\mathbf{q}^R = \mathbf{0}$) in (D.20).

Proof. Proof. Apply Theorem 10 of Bertsimas et al. (2014b) and observe that since $g(u)$ is monotonic and right continuous,

$$c_j \ge \sup_{u \in (\hat{u}_i^{(j-1)}, \hat{u}_i^{(j)}]} g(u) \iff c_j \ge g(\hat{u}_i^{(j-1)}), \; c_j \ge g(\hat{u}_i^{(j)}).$$

Take the dual of this (finite) conic optimization problem to obtain the given maximization formulation.

To prove the last statement, suppose first that $g(u)$ is non-decreasing and fix some j. If $g(\hat{u}_i^{(j)}) > g(\hat{u}_i^{(j-1)})$, then by complementary slackness, $\mathbf{q}^L = 0$. If $g(\hat{u}_i^{(j)}) = g(\hat{u}_i^{(j-1)})$, then given any feasible (q_j^L, q_j^R), the pair $(0, q_j^L + q_j^R)$ is also feasible with the same objective value. Thus, without loss of generality, $\mathbf{q}^L = 0$. The case where $g(u)$ is non-increasing is similar. □

Remark D.5. At optimality of (D.20), $\mathbf{p}$ can be considered a probability distribution, supported on the points $\hat{u}_i^{(j)}$ $j = 0, \ldots, N+1$. This distribution is analogous to $\mathbf{q}^L(\Gamma), \mathbf{q}^R(\Gamma)$ for the KS test.

In the special case of the K test, we can solve (D.20) explicitly to find this worst-case distribution.

Corollary D.6. *When $\mathcal{P}_i^{EDF}$ refers specifically to the K test in Theorem D.4 and if g is monotonic, we have*

$$\sup_{\mathbb{P}_i \in \mathcal{P}_i^{EDF}} \mathbb{E}^{\mathbb{P}_i}[g(\tilde{u}_i)] = \max\left(\sum_{j=0}^{N+1} q_j^L(\Gamma^K) g(\hat{u}_i^{(j)}), \sum_{j=0}^{N+1} q_j^R(\Gamma^K) g(\hat{u}_i^{(j)})\right). \tag{D.21}$$

Proof. Proof. One can check that in the case of the K test, the maximization formulation given is equivalent to (D.5) with Γ^{KS} replaced by Γ^K. Following the proof of Theorem D.2 yields the result. □

Remark D.7. One an prove that $\Gamma^K \geq \Gamma^{KS}$ for all N, α. Consequently, $\mathcal{P}_i^{KS} \subseteq \mathcal{P}_i^K$. For practical purposes, one should thus prefer the KS test to the K test, as it will yield smaller sets.

We can now generalize Theorem 5.11. For each of K, CvM, W and AD tests, define the (finite dimensional) set

$$\mathcal{P}_i^{EDF} = \{\mathbf{p} \in \mathbb{R}_+^{N+2} : \exists \mathbf{q}^L, \mathbf{q}^R \in \mathbb{R}_+^{N+2}, \mathbf{z} \in \mathbb{R}^N,\ \mathbf{p}, \mathbf{q}^L, \mathbf{q}^R, \mathbf{z} \text{ are feasible in (D.20)}\}, \tag{D.22}$$

using the appropriate $\mathbf{A}_S, \mathbf{b}_S, \mathcal{K}_S$.

Theorem D.8. *Suppose $\mathbb{P}^*$ has independent components, with supp$(\mathbb{P}^*) \subseteq [\hat{\mathbf{u}}^{(0)}, \hat{\mathbf{u}}^{(N+1)}]$.*

i) With probability at least $1-\alpha$ over the sample, the family $\{\mathcal{U}_\epsilon^I : 0 < \epsilon < 1\}$ simultaneously implies a probabilistic guarantee, where

$$\mathcal{U}_\epsilon^I = \Bigg\{ \mathbf{u} \in \mathbb{R}^d : \exists \mathbf{p}^i \in \mathcal{P}_i^{EDF},\ \mathbf{q}^i \in \Delta_{N+2},\ i = 1 \ldots, d, \\ \sum_{j=0}^{N+1} \hat{u}_i^{(j)} q_j^i = u_i\ i = 1, \ldots, d,\ \ \sum_{i=1}^{d} D(\mathbf{q}^i, \mathbf{p}^i) \leq \log(1/\epsilon) \Bigg\}. \tag{D.23}$$

ii) In the special case of the K test, the above formulation simplifies to (5.20) with Γ^{KS} replaced by Γ^K.

The proof of the first part is entirely analogous to Theorem 5.11, but uses Theorem D.4 to evaluate the worst-case expectations. The proof of the second part follows by applying Corollary D.6. We omit the details.

Remark D.9. In contrast to our definition of $\mathcal{U}_\epsilon^I$ using the KS test, we know of no simple algorithm for evaluating $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^I)$ when using the CvM, W, or AD tests. (For the K test, the same algorithm applies but with Γ^K replacing Γ^{KS}.) Although it still polynomial time to optimize over constraints $\delta^*(\mathbf{v}|\,\mathcal{U}_\epsilon^I) \leq t$ for these tests using interior-point solvers for conic optimization, it is more challenging numerically.

Appendix E

Appendix to Chapter 7

E.1 A Priori Balance in Estimating Treatment Effect on Compliers

In many experimental endeavors involving human subjects the researcher does not fully control the treatment actually administered. Consider two treatments, "treatment" ($k = 1$) and "control" ($k = 2$). Situations where a subject receives a treatment different from their assignment include refusal of surgery, ethical codes that allow subjects assigned to control to demand treatment, or the leakage of information to some control subjects in a teaching intervention. This issue is termed non-compliance. In such situations, W represents initial assignment intent and our estimator $\hat{\tau}$ estimates the effect of the *intent* to treat (ITT). Often a researcher is interested in the compliers' average treatment effect in the sample (CSATE) or population (CPATE), disregarding all non-compliers. Subjects that always demand treatment are known as always-takers, those that always refuse treatment as never-takers, and those that always choose the opposite of their assignment as defiers (this is exhaustive if subjects comply based only on their own assignment). Denote by π_c and Π_c the unknown fraction of compliers in the sample and population, respectively. In the absence of defiers we can observe the identity of never-takers in the treatment group and of always-takers in the control group. We can estimate the fraction of compliers as the complement of those:

$$\hat{\pi}_c = 1 - \frac{2}{n} \sum_{i:W_i=1} \mathrm{NT}_i - \frac{2}{n} \sum_{i:W_i=2} \mathrm{AT}_i$$

where $\mathrm{NT}_i = 1$ if i is a never-taker and $\mathrm{AT}_i = 1$ if i is an always-taker (both 0 for compliers). Under an assignment that blinds the identity of treatment, such as complete randomization, $\hat{\pi}_c$ is conditionally (for π_c) and marginally (for Π_c) unbiased if there are no defiers. Moreover, without defiers,

$$\mathrm{CSATE} = \mathrm{SATE} \,/ \pi_c \qquad \mathrm{CPATE} = \mathrm{PATE} \,/ \Pi_c$$

since the individual ITT effect for an always- or never-taker is identically 0. It has been often advocated (see Imbens and Rubin (1997), Little and Yau (1998)) in completely randomized trials to estimate the compliers' average treatment effect by a ratio estimator $\hat{\tau}_c = \hat{\tau}/\hat{\pi}_c$. Such an estimator need not be unbiased but because it is the ratio of two unbiased estimators it has been argued to be approximately unbiased (ibid.). Under a design that blinds the identity of treatments the two estimators remain unbiased and the very same approach can be taken.

We can do even better if we use a priori balance to improve the precision of the compliance fraction estimator. The difference between the sample compliance fraction and our estimator of it can be seen to be

$$\hat{\pi}_c - \pi_c = \frac{2}{N}\sum_{i:W_i=1}(\mathrm{AT}_i - \mathrm{NT}_i) - \frac{2}{n}\sum_{i:W_i=2}(\mathrm{AT}_i - \mathrm{NT}_i) = \frac{2}{n}\sum_{i=1}^{n} u_i C_i$$

$$\text{where} \quad C_i = \begin{cases} 1 & i \text{ is always-taker} \\ 0 & i \text{ is complier} \\ -1 & i \text{ is never-taker} \end{cases} \quad \text{is } i\text{'s compliance status.}$$

Therefore, matching the means of $f_c(x) = \mathbb{E}[C_i | X_i = x]$ will eliminate variance in estimating the compliance fraction and get us closer to the true CSATE and CPATE. Moreover, if the two unbiased estimators, $\hat{\tau}$ and $\hat{\pi}_c$, are both more precise, their ratio $\hat{\tau}_c$ is both more precise and less biased. To achieve this through our framework we need only incorporate our belief $\mathcal{F}_c$ about f_c into the larger $\mathcal{F}$ and proceed as before. (See also supplemental Section E.2 for a discussion about combining spaces.)

E.2 Generalizations of $\mathcal{F}$

In this supplemental section we consider more general forms of the space $\mathcal{F}$. For the most part, the theorems presented in the main text will still apply. We deferred this discussion to this supplement to avoid overly cumbersome notation in the main text.

First, we consider the restriction to cones in $\mathcal{F}$. A cone is a set $C \subset \mathcal{F}$ such that $f \in C \implies cf \in C \; \forall c > 0$. We may then further restrict to $f \in C$, $||f|| \leq 1$ in the definitions of $M_{\mathrm{p}}^2(W)$ and $M_{\mathrm{m}}^2(\sigma)$. By symmetry, this is the same as restricting to $C \cup (-C)$. Since it is still the case that $||cf|| = c\,||f||$, Theorems 7.14 and 7.18 still apply. One example of a cone is the cone of monotone functions (either nondecreasing or nonincreasing). In a single dimension and for two treatments, this will result in a pure-strategy optimal design that sorts the data and assigns subjects in an alternating fashion. This is also a feasible assignment for pairwise matching in one dimension. More generally and in higher dimensions, we can consider a directed acyclic graph (DAG) on the nodes $V = \{1, \ldots, n\}$ with edge set $E \subset V^2$ and its associated topological cone $C = \{f : f(X_i) \leq f(X_j) \; \forall (i,j) \in E\}$. Other cones include nonnegative/positive functions and $\pm$-sum-of-squares polynomials.

Second, we consider re-centering the norms. We might have a nominal regression function g that we believe is approximately right, perhaps due to a prior regression analysis or based on models from the literature. In this case, it would make sense to

solve the minimax problem against perturbations around this g. Given a norm $||\cdot||'$ on $\mathcal{F}$ we can formally define the magnitude

$$||f|| = \max\left\{\min\left\{||f-g||', ||f+g||'\right\}, 1\right\}. \tag{E.1}$$

We consider both g and $-g$ because it has no effect on the imbalance metrics due to symmetry of the objective while it can only reduce magnitudes. Using this alternate definition of $||\cdot||$ in (E.1), Theorem 7.14 still applies and Theorem 7.18 applies if its conditions apply to the Banach space $\mathcal{F}$ with its usual norm and $\mathbb{E}\,|g(X_1)| < \infty$. In the Bayesian interpretation discussed in Section 7.2.4, this is equivalent to making the prior mean of $f(x)$ be $g(x)$.

Third, we consider combining multiple spaces $\mathcal{F}_1, \ldots, \mathcal{F}_b$. There are two ways. The first way is to combine these via an algebraic sum. The space $\mathcal{F} = \mathcal{F}_1 + \cdots + \mathcal{F}_b = \{\phi_1 + \cdots + \phi_b : \phi_j \in \mathcal{F}_j \ \forall j\}$ endowed with the norm $||f|| = \min_{\phi_j \in \mathcal{F}_j : f = \phi_1 + \cdots + \phi_b} \max_{j=1,\ldots,b} ||\phi_j||_{\mathcal{F}_j}$ is Banach space and as such a valid choice. In particular, the algebraic sum $\mathcal{F}$ can be identified with the quotient of the direct sum $\mathcal{F}' = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_b$ by its subspace $\{(\phi_1, \ldots, \phi_b) \in \mathcal{F}' : \phi_1 + \cdots + \phi_b = 0\}$. We can decompose the pure-strategy imbalance metric corresponding to this new choice as follows:

$$M_{\mathrm{p}}^2(W) = \max_{k \neq k'} \left(\sum_{j=1}^{b} \sup_{||\phi_j||_{\mathcal{F}_j} \leq 1} B_{kk'}(W, \phi_j) \right)^2.$$

Theorems 7.14 and 7.18 still apply (in particular the conditions of Theorem 7.18 hold for $\mathcal{F}$ if they hold for each $\mathcal{F}_j$).

The second way is to combine these formally via a union. Consider the space $\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_b = \{f : f \in \mathcal{F}_j \text{ for some } j\}$. This is not a vector space but we can formally define the magnitude $||f|| = \min_{j=1,\ldots,b} ||f||_{\mathcal{F}_j}$. We can then decompose the pure-strategy imbalance metric corresponding to this new choice as follows:

$$M_{\mathrm{p}}^2(W) = \max_{k \neq k'} \max_{j=1,\ldots,b} \sup_{||\phi_j||_{\mathcal{F}_j} \leq 1} B_{kk'}^2(W, \phi_j).$$

Theorem 7.14 still applies and Theorem 7.18 applies if its conditions hold for each Banach space $\mathcal{F}_j$.

We can even take several spaces $\mathcal{F}_1, \ldots, \mathcal{F}_b$, re-center each norm with its own g_j as in (E.1), and then combine them in either of the two ways, defining the combined magnitudes strictly formally. In this way, we can have multiple centers to represent various beliefs about the same or different regression functions f_k. Theorem 7.14 still applies and Theorem 7.18 applies if its conditions hold for each $\mathcal{F}_j$ and $\mathbb{E}\,|g_j(X_1)| < \infty$ for for each j.

E.3 Inference for Mixed-Strategy Designs

As noted in Section 7.5 Algorithm 7.5.1 can be used to answer inferential questions for mixed-strategy designs as well, but their additional randomization allows for the standard randomization and exact permutation tests to be used instead. The following is the standard permutation test when applied to a non-completely randomized design, including the mixed-strategy optimal design.

Algorithm E.3.1. Let σ be given. For a confidence level $0 < 1 - \alpha < 1$:

1: Draw W^0 from σ, assign subjects, apply treatments, measure $Y_{iW_i^0}$, and compute $\hat{\tau}$. Let $\mathcal{W}' = \{W \in \mathcal{W} : \sigma(W) > 0\}$.

2: For $W \in \mathcal{W}'$ compute $\tilde{\tau}^W = \frac{1}{p}\sum_{i:W_i=1} Y_{iW_i^0} - \frac{1}{p}\sum_{i:W_i=2} Y_{iW_i^0}$.

3: The p-value of H_0 is $p = \sum_{W\in\mathcal{W}'} \sigma(W)\mathbb{I}\left[\left|\tilde{\tau}^W\right| \geq |\hat{\tau}|\right]$.
If $p \leq \alpha$ then reject H_0.

The above exact test requires that we have a full description of σ and that we iterate over all feasible assignments. This works well for the output of Algorithm 7.4.2 but can be prohibitive for the output of Algorithm 7.4.1. The standard randomization test eschews these issues.

Algorithm E.3.2. Let σ be given. For a confidence level $0 < 1 - \alpha < 1$:

1: Draw W^0 from σ, assign subjects, apply treatments, measure $Y_{iW_i^0}$, and compute $\hat{\tau}$.

2: For $t = 1, \ldots, T$ do:

2.1: Draw W^t from σ.

2.2: Compute $\tilde{\tau}^t = \frac{1}{p}\sum_{i:W_i^t=1} Y_{iW_i^0} - \frac{1}{p}\sum_{i:W_i^t=2} Y_{iW_i^0}$.

3: The p-value of H_0 is $p = \left(1 + |\{t \ : \ |\tilde{\tau}^t| \geq |\hat{\tau}|\}|\right) \big/ (1 + T)$.
If $p(H_0) \leq \alpha$ then reject H_0.

E.4 Omitted Proofs

Proof of Theorem 7.1. Simple arithmetic yields,

$$\hat{\tau} - \text{SATE} = \frac{2}{n}\sum_{i:W_i=1}\left(\frac{Y_{i1}+Y_{i2}}{2}\right) - \frac{2}{n}\sum_{i:W_i=2}\left(\frac{Y_{i1}+Y_{i2}}{2}\right) = \frac{2}{n}\sum_{i=1}^{n} u_i\hat{Y}_i.$$

By conditional unbiasedness, we have

$$\text{Var}\left(\hat{\tau}|X,Y\right) = \mathbb{E}\left[\left(\hat{\tau} - \text{SATE}\right)^2 \Big| X, Y\right] = \mathbb{E}\left[\left(\frac{2}{n}\sum_{i=1}^{n} u_i\hat{Y}_i\right)^2 \Bigg| X, Y\right].$$

Consider any feasible $\sigma \in \Delta$ and let W be drawn from it. Because shifting Y_1 by one constant and Y_2 by another amounts to shifting $\hat{Y}$ by a constant, which does not change $\hat{\tau}$, by minimizing norms we have that

$$\max_{Y \in \mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|Y_1\|^2 + \|Y_2\|^2} = \max_{Y \in \mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|\overline{Y}\|^2} = \max_{Y \in \mathbb{R}^{n\times 2}} \frac{\mathrm{Var}\left(\hat{\tau}|X,Y\right)}{\|\hat{Y}\|^2}$$
$$= \max_{\hat{Y}\in\mathbb{R}^n : \|\hat{Y}\| \le 1} \sum_{W \in \mathcal{W}} \sigma\left(W\right) \left(\frac{2}{n}\sum_{i=1}^{n} u_i \hat{Y}_i\right)^2. \tag{E.2}$$

Suppose $\sigma \in \Delta$ minimizes (E.2). For any $\pi \in S_n$ a permutation of $\{1,\ldots,n\}$, define $\sigma_\pi((W_1,\ldots,W_n)) = \sigma((W_{\pi(1)},\ldots,W_{\pi(n)}))$. Then by the symmetry of $||\cdot||$, σ_π is also optimal. Next note that (E.2) is a maximum over linear forms in σ and is therefore convex. Therefore, $\sigma^*(W) = \frac{1}{n!}\sum_{\pi\in S_n}\sigma_\pi(W)$ is also optimal. By construction we get $\sigma^*((W_1,\ldots,W_n)) = \sigma^*((W_{\pi(1)},\ldots,W_{\pi(n)}))$ for any $\pi \in S_n$. Hence, $\sigma^*((W_1,\ldots,W_n)) = \sigma^*((1,2,1,2,\ldots,1,2))$ is constant for every $W \in \mathcal{W}$, and therefore σ^* is complete randomization. □

Proof of Theorem 7.5. First note that by (7.2), for any i,j,k,k',

$$\sigma\left(\{W_i = W_j,\, W_i \in \{k,k'\},\, W_j \in \{k,k'\}\}\right) = \frac{2}{m}\sigma\left(\{W_i = W_j\}\right),$$
$$\sigma\left(\{W_i \neq W_j,\, W_i \in \{k,k'\},\, W_j \in \{k,k'\}\}\right) = \frac{2}{m}\frac{1}{m-1}\sigma\left(\{W_i \neq W_j\}\right).$$

Therefore, by squaring and interchanging sums, we have

$$M_{\mathrm{m}}^2(\sigma) = \max_{||f||\le 1}\max_{k\neq k'} \frac{1}{p^2}\sum_{i,j=1}^{n} f(X_i)f(X_j)\sum_{W\in\mathcal{W}}\sigma(W)(w_{ik}-w_{ik'})(w_{jk}-w_{jk'})$$
$$= \max_{||f||\le 1}\max_{k\neq k'} \frac{2}{pn}\sum_{i,j=1}^{n} P_{ij}(\sigma)f(X_i)f(X_j)$$
$$= \max_{||f||\le 1} \frac{2}{pn}\sum_{i,j=1}^{n} P_{ij}(\sigma)f(X_i)f(X_j).$$ □

Proof of Theorem 7.6. Let $\{x_1,\ldots,x_\ell\}$ be the set of values taken by the baseline covariates $X_1,\ldots,X_n$ ($\ell \le n$). Let an assignment W be given. Let $\{i_1,i_1'\},\ldots,\{i_q,i_q'\}$ denote a maximal perfect exact match across the two groups ($W_{i_j}=1$, $W_{i_j'}=2$, $X_{i_j}=X_{i_j'}$, and q maximal) with $\{i_1'',\ldots,i_{q'}''\}$, $\{i_1''',\ldots,i_{q'}'''\}$ being the remaining unmatched subjects ($W_{i_j''}=1$, $W_{i_{j'}'''}=2$, $X_{i_j''} \neq X_{i_{j'}'''}$). For $i=1,\ldots,\ell$, if there are more x_i's in group 1 set $f'(x_i)=1$ otherwise set $f'(x_i)=-1$. This f' is feasible ($||f'||_\infty \le 1$) and hence

$$\max_{||f||\le 1}|B(W,f)| \ge |B(W,f')| = \frac{2}{n}\times q' \times 2 = 2 - \frac{4}{n}q.$$

At the same time, we have

$$\begin{aligned}\max_{||f||\le 1} |B(W,f)| &= \max_{||f||\le 1} \left|\sum_{i=1}^{n} u_i f(X_i)\right| \\ &\le \frac{2}{n}\sum_{j=1}^{q} \max_{||f||\le 1} \left|f(X_{i_j}) - f(X_{i'_j})\right| + \frac{2}{n}\sum_{j=1}^{q'} \max_{||f||\le 1} \left|f(X_{i''_j}) - f(X_{i'''_j})\right| \\ &= 0 + \frac{2}{n} \times q' \times 2 = 2 - \frac{4}{n}q.\end{aligned}$$

To summarize,

$$\sqrt{M_{\mathrm{p}}^2(W)} = 2 - \frac{4}{n}\begin{pmatrix}\text{number of perfect exact matches}\\ \text{across the experimental groups}\end{pmatrix}. \quad \square$$

Proof of Theorem 7.7. Let $D_{ij} = \delta(X_i, X_j)$. The pure-strategy optimal design solves the optimization problem

$$\min_{W\in\mathcal{W}} \max_{||f||_{\mathrm{lip}}\le 1} |B(W,f)| = \frac{2}{n} \min_{\substack{u\,\in\,\{-1,1\}^n \\ \sum_{i=1}^n u_i = 0}} \max_{\substack{y\,\in\,\mathbb{R}^n \\ y_i - y_j \le D_{ij}}} u^T y. \tag{E.3}$$

We will show that the set of optimal solutions u to (E.3) is equal to the set of assignments of $+1, -1$ to the pairs in any minimal-weight pairwise match. Since the pure-strategy optimal design randomizes over these, this will show that it is equivalent to pairwise matching, which randomly splits pairs.

Consider any non-bipartite matching $\mu = \{\{i_1, j_1\}, \ldots, \{i_{n/2}, j_{n/2}\}\}$ and any $t \in \{-1, +1\}^{n/2}$. Let $u_{i_l} = t_l$, $u_{j_l} = -t_l$. Enforcing only a subset of the constraints on y, the cost of u in (E.3) is bounded above as follows

$$\max_{y_i - y_j \le D_{ij}} u^T y = \max_{y_i - y_j \le D_{ij}} \sum_{l=1}^{n/2} t_l(y_{i_l} - y_{j_l}) \le \sum_{l=1}^{n/2} D_{i_l j_l},$$

which is the matching cost of μ. Now let instead a feasible solution u to (E.3) be given. Let $S = \{i : u_i = +1\} = \{i_1, \ldots, i_{n/2}\}$ and its complement $S^C = \{i : u_i = -1\} = \{i'_1, \ldots, i'_{n/2}\}$. By linear programming duality we have

$$\max_{y_i - y_j \le D_{ij}} u^T y = \min_{Fe - F^T e = u,\, F \ge 0} \sum_{i,j=1}^{n} D_{ij} F_{ij} \tag{E.4}$$

since the LHS is bounded ($\le D_{i_1 i'_1} + \cdots + D_{i_{n/2} i'_{n/2}}$) and feasible ($y_i = 0\ \forall i$). The RHS is an uncapacitated min-cost transportation problem with sources S (with inputs 1) and sinks S^C (with outputs 1). Consider any $j_s \in S$, $j_t \in S^C$ and any path

$j_s, j_1, \ldots, j_p, j_t$. By the triangle inequality,

$$D_{j_s j_t} \leq D_{j_s j_1} + D_{j_1 j_2} + \cdots + D_{j_p j_t}.$$

Therefore, it is always preferable to send flow along edges between S and S^C only. Thus, erasing all edges within S or S^C, the problem is seen to be a bipartite matching problem. The min-weight bipartite matching is also a non-bipartite matching and by (E.4) its matching cost is the same as the cost of the given u in the objective of (E.3). □

Proof of Theorem 7.9. The argument is similar to the above. This time the network flow problem has an additional node with zero external flow (neither sink nor source), uncapacitated edges into it from every other node with a unit cost of δ_0, and uncapacitated edges out of it to every other node with a unit cost of δ_0. □

Proof of Theorem 7.10. For the pure-strategy case we have,

$$\begin{aligned}
M_{\mathrm{p}}^2(W) &= \max_{k \neq k'} \max_{||f|| \leq 1} \left(\frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) \right)^2 \\
&= \max_{k \neq k'} \left|\left| \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot) \right|\right|^2 \\
&= \frac{1}{p^2} \max_{k \neq k'} \left\langle \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot), \quad \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot) \right\rangle \\
&= \frac{1}{p^2} \max_{k \neq k'} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'})
\end{aligned}$$

Now, consider the maximum over f in $M_{\mathrm{m}}^2(P)$. Let f_0 be a feasible solution. Write $f_0 = f + f^{\perp}$ with $f \in S = \mathrm{span}\{\mathcal{K}(X_i, \cdot) \ : \ i = 1, \ldots, n\}$ and $f^{\perp} \in S^{\perp}$, its orthogonal complement. By orthogonality $f^{\perp}(X_i) = \left\langle \mathcal{K}(X_i, \cdot), f^{\perp} \right\rangle = 0$ and $||f||^2 = ||f_0||^2 - \left|\left|f^{\perp}\right|\right|^2 \leq 1$ so that f achieves the same objective value as f_0 and remains feasible. Therefore we may restrict to S and assume that $f = \sum_i \beta_i \mathcal{K}(X_i, \cdot)$ such that $\beta^T K \beta \leq 1$.

By positive semi-definiteness of K and P, we get

$$\begin{aligned}
M_{\mathrm{m}}^2(P) &= \frac{2}{np} \sup_{\beta^T K \beta \leq 1} \sum_{i,j=1}^{n} P_{ij} \left(K\beta\right)_i \left(K\beta\right)_j = \frac{2}{np} \sup_{\beta^T K \beta \leq 1} \beta^T K P K \beta \\
&= \frac{2}{np} \sup_{\gamma^T \gamma \leq 1} \gamma^T \sqrt{K} P \sqrt{K} \gamma = \frac{2}{np} \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right).
\end{aligned}$$ □

Proof of Theorem 7.13. By blinding of treatments (7.2), each W_i by itself (but *not*

the vector W) is statistically independent of X, Y so that

$$\mathbb{E}\left[w_{ik}Y_{ik}\middle|X,Y\right] = \mathbb{E}\left[w_{ik}\right]Y_{ik} = \frac{1}{m}Y_{ik}, \text{ and therefore}$$
$$\mathbb{E}\left[\hat{\tau}_{kk'}\middle|X,Y\right] = \frac{1}{p}\sum_{i=1}^{n}\frac{1}{m}Y_{ik} - \frac{1}{p}\sum_{i=1}^{n}\frac{1}{m}Y_{ik'} = \mathrm{SATE}_{kk'}\,.$$

Note that we can rewrite $E_{kk'}$ as

$$E_{kk'} = \frac{1}{m}\sum_{l\neq k}\Xi_{kl} - \frac{1}{m}\sum_{l\neq k'}\Xi_{k'l} \quad \text{where} \quad \Xi_{kl} := \frac{1}{p}\sum_{i=1}^{n}\left(w_{ik} - w_{il}\right)\epsilon_{ik}.$$

Using the notation $A_{kk'} = \frac{1}{p}\sum_{i:W_i=k'} Y_{ik}$, $C_{kl} = B_{kl}\left(f_k\right) + \Xi_{kl}$, we have

$$\begin{aligned}\hat{\tau}_{kk'} - \mathrm{SATE}_{kk'} &= A_{kk} - A_{k'k'} - \frac{1}{m}\sum_{l=1}^{m}A_{kl} + \frac{1}{m}\sum_{l=1}^{m}A_{k'l} \\ &= \frac{m-1}{m}A_{kk} - \frac{1}{m}A_{kk'} + \frac{1}{m}A_{k'k} - \frac{m-1}{m}A_{k'k'} \\ &\quad - \frac{1}{m}\sum_{l\neq k,k'}\left(A_{kk} - C_{kl}\right) + \frac{1}{m}\sum_{l\neq k,k'}\left(A_{k'k'} - C_{k'l}\right) = D_{kk'} + E_{kk'}.\end{aligned}$$

Let i, j be equal or unequal, k, k', l, l' equal or unequal. Then,

$$\begin{aligned}\mathrm{Cov}(w_{il}f_k(X_i),\, w_{jl'}\epsilon_{jk'}) &= \mathbb{E}\left[w_{il}w_{jl'}f_k(X_i)\mathbb{E}\left[\epsilon_{jk'}|X,Z\right]\right] \\ &\quad - \mathbb{E}\left[w_{il}f_k(X_i)\right]\mathbb{E}\left[w_{jl'}\mathbb{E}\left[\epsilon_{jk'}|X,Z\right]\right] = 0 - 0 = 0, \\ \mathrm{Cov}((w_{ik} - w_{il})f_k(X_i),\, f_{k'}(X_j)) &= \mathbb{E}\left[w_{ik} - w_{il}\right]\mathrm{Cov}\left(f_k(X_i), f_{k'}(X_j)\right) = 0, \\ \mathrm{Cov}((w_{ik} - w_{il})\epsilon_{ik},\, f_{k'}(X_j)) &= \mathbb{E}\left[w_{ik} - w_{il}\right]\mathrm{Cov}\left(\epsilon_{ik}, f_{k'}(X_j)\right) = 0,\end{aligned}$$

where the latter two equalities are due to the independence of W_i due to blinding treatments. This proves uncorrelateness. The rest follows from an application of the law of total variance and rearranging terms. □

Proof of Theorem 7.14. Define

$$Z(f,g) = \mathbb{E}\left[\left(\frac{1}{p}\sum_{i=1}^{n}\left(w_{i1} - w_{i2}\right)f(X_i)\right)\left(\frac{1}{p}\sum_{i=1}^{n}\left(w_{i1} - w_{i2}\right)g(X_i)\right)\right].$$

By construction, $Z(f,f) \leq ||f||^2 \mathbb{E}\left[M_{\text{opt}}^2\right]$. By condition (7.2),

$$\begin{aligned}
&\text{Var}\left(B_{kl}(f)\right) = Z(f,f) \quad \text{for } l \neq k,\\
&\text{Cov}\left(B_{kl}(f), B_{kl'}(f)\right) = \frac{1}{2}Z(f,f) \quad \text{for } k,l,l' \text{ distinct},\\
&\text{Cov}\left(B_{kl}(f), B_{k'l'}(g)\right) = \begin{cases} \frac{1}{2}Z(f,g) & \text{for } l = l' \notin \{k,k'\},\\ -\frac{1}{2}Z(f,g) & \text{for } l = k',\ l' \neq k,\\ -\frac{1}{2}Z(f,g) & \text{for } l \neq k',\ l' = k,\\ -Z(f,g) & \text{for } l = k',\ l' = k,\\ 0 & \text{for } k,k',l,l' \text{ distinct.}\end{cases}
\end{aligned}$$

It follows that

$$\begin{aligned}
\text{Var}\left(D_{kk'}\right) &= \frac{1}{m^2}\left(\frac{m^2-m}{2}Z(f_k,f_k) + \frac{m^2-m}{2}Z(f_{k'},f_{k'})\right)\\
&\quad + \frac{m+2}{m^2}Z(f_k,f_{k'})\\
&= \frac{1}{m^2}\left(\frac{m^2}{2} - m - 1\right)\left(Z(f_k,f_k) + Z(f_{k'},f_{k'})\right)\\
&\quad + \frac{1}{m^2}\left(\frac{m+2}{2}\right)Z(f_k+f_{k'},f_k+f_{k'})\\
&\leq \frac{1}{m^2}\left(\frac{m^2}{2} - m - 1\right)\left(\mathbb{E}\left[M_{\text{opt}}^2\right]||f_k||^2 + \mathbb{E}\left[M_{\text{opt}}^2\right]||f_{k'}||^2\right)\\
&\quad + \frac{1}{m^2}\left(\frac{m+2}{2}\right)\mathbb{E}\left[M_{\text{opt}}^2\right]||f_k+f_{k'}||^2\\
&\leq \frac{\left(||f_k|| + ||f_{k'}||\right)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M_{\text{opt}}^2\right]
\end{aligned}$$

since $||f+g||^2 \leq \left(||f|| + ||g||\right)^2$ and $\left(||f||^2 + ||g||^2\right) \leq \left(||f|| + ||g||\right)^2$. □

Proof of Theorem 7.16. Fix f and g. Using $\left(\sum_{i=1}^b z_i\right)^2 \leq b\sum_{i=1}^b z_i^2$,

$$\begin{aligned}
Z(f,f) &= \mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)(f-g)(X_i) + \frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)g(X_i)\right)^2\\
&\leq 2\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)(f-g)(X_i)\right)^2 + 2\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)g(X_i)\right)^2\\
&\leq \frac{2}{p^2} \times p \times p \times \frac{2}{m} \times \mathbb{E}\left((f-g)(X_1)\right)^2 + 2Z(g,g) = \frac{4}{m}||f-g||_2^2 + 2Z(g,g)
\end{aligned}$$

The rest is as in the proof of Theorem 7.14, choosing $g \in \mathcal{F}$. □

Proof of Theorem 7.18. Fix the assignment $W_i' = (i \mod p) + 1$ and let $\xi_i^{(k,k')} : f \mapsto \left(f\left(X_{m(i-1)+k}\right) - f\left(X_{m(i-1)+k'}\right)\right)$. Then, since $\xi_i^{(k,k')}$ is in the continuous dual space

$\mathcal{F}^*$, we can write $M_{\mathrm{p}}^2(W') = \max_{k \neq k} T_n^{(k,k')}$ where

$$T_n^{(k,k')} = \sup_{||f|| \leq 1} \left(\frac{1}{p} \sum_{i=1}^{p} \xi_i^{(k,k')}(f) \right)^2 = \left| \left| \frac{1}{p} \sum_{i=1}^{p} \xi_i^{(k,k')} \right| \right|_{\mathcal{F}^*}^2 .$$

Note $\xi_i^{(k,k')}$ are independent and identically distributed with expectation (Bochner integral) equal 0. B-convexity of $\mathcal{F}$ implies the B-convexity of $\mathcal{F}^*$. By B-convexity and the main result of Beck (1962) (or by Chen and Zhu (2011) for the Hilbert case),

$$T_n^{(k,k')} \to 0 \text{ almost surely as } n \to \infty.$$

As there are only finitely many k, k', we have $M_{\mathrm{p}}^2(W') \to 0$ almost surely. By construction, $M_{\text{m-opt}}^2 \leq M_{\text{p-opt}}^2 \leq M_{\mathrm{p}}^2(W')$. Hence, the distance between $\hat{\tau}_{kk'}$ and $\mathrm{SATE}_{kk'} + E_{kk'}$ is $|D_{kk'}| \leq \left(1 - \frac{1}{m}\right)(||f_k|| + ||f_{k'}||)\sqrt{M_{\text{opt}}^2} \to 0$ almost surely. Therefore, as $\mathrm{SATE}_{kk'} + E_{kk'}$ is strongly consistent, so is $\hat{\tau}_{kk'}$. □

Bibliography

C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

M. Ahsanullah, V. B. Nevzorov, and M. Shakil. *An Introduction to Order Statistics*. Atlantis Press, 2013.

N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

S. Asur and B. Huberman. Predicting the future with social media. In *WI-IAT*, pages 492–499, 2010.

C. Bandi and D. Bertsimas. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical programming*, 134(1):23–70, 2012.

C. Bandi, B. Bertsimas, and N. Youssef. Robust queueing theory. Submitted for publication to Operations Research, 2012.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.

A. Beck. A convexity condition in Banach spaces and the strong law of large numbers. *Proc. Amer. Math. Soc.*, 13(2):329–334, 1962.

A. Belloni and V. Chernozhukov. $\ell 1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

A. Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(3):411–424, 2000.

A. Ben-Tal, B. Golany, A. Nemirovski, and J. Vial. Retailer-supplier flexible commitments contracts: a robust optimization approach. *Manufacturing & Service Operations Management*, 7(3):248–271, 2005.

A. Ben-Tal, D. Den Hertog, and J. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, pages 1–35, 2012.

A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2001.

A. Ben-Tal and A. Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

J. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975.

R. Beran. Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley, 1981.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

D. Bertsekas, A. Nedić, and A. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, Belmont, 2003.

D. P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, 1999.

D. Bertsimas and D. Brown. Constructing uncertainty sets for robust linear optimization. *Operations Research*, 57(6):1483–1495, 2009.

D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

D. Bertsimas, D. Gamarnik, and A. Rikun. Performance analysis of queueing networks via robust optimization. *Operations research*, 59(2):455–466, 2011a.

D. Bertsimas, I. Dunning, and M. Lubin. Reformulations versus cutting planes for robust optimization. 2014a. URL `http://www.optimization-online.org/DB_HTML/2014/04/4336.html`.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics in multi-period decision problems. 2015a.

D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. 2015b.

D. Bertsimas and G. Perakis. Dynamic pricing: A learning approach. In S. Lawphongpanich, D. W. Hearn, and M. J. Smith, editors, *Mathematical and Computational Models for Congestion Charging*, volume 101 of *Applied Optimization*, pages 45–79. Springer, 2006.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific, Belmont, 1997.

D. Bertsimas and Y. Ye. Semidefinite relaxations, multivariate normal distributions, and order statistics. In *Handbook of Combinatorial Optimization*, pages 1473–1491. Springer, 1999.

D. Bertsimas, M. Johnson, and N. Kallus. The power of optimization over randomization in designing experiments involving small samples. Submitted for publication.

D. Bertsimas, X. V. Doan, K. Natarajan, and C.-P. Teo. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, 35(3):580–602, 2010.

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011b.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. 2013. Submitted to *Operations Research*.

D. Bertsimas, V. Gupta, and N. Kallus. Robust SAA. 2014b. Submitted to *Mathematical Programming*.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 2015.

O. Besbes, R. Phillips, and A. Zeevi. Testing the validity of a demand model: An operations perspective. *Manufacturing & Service Operations Management*, 12(1):162–183, 2010.

T. H. Bijmolt, H. J. v. Heerde, and R. G. Pieters. New empirical generalizations on the determinants of price elasticity. *Journal of marketing research*, 42(2):141–156, 2005.

P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.

J. R. Birge and R. J. B. Wets. Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse. *Math. Programming Study*, 27:54–102, 1986.

J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, 2011.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10):P10008, 2008.

C. Borgs, J. Chayes, S. Mertens, and C. Nair. Proof of the local REM conjecture for number partitioning. I: Constant energy scales. *Random Structures & Algorithms*, 34 (2):217–240, 2009a.

C. Borgs, J. Chayes, S. Mertens, and C. Nair. Proof of the local REM conjecture for number partitioning. II. growing energy scales. *Random Structures & Algorithms*, 34 (2):241–284, 2009b.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

R. Bradley. Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, pages 165–192. Birkhausser, 1986.

R. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2(107-44):37, 2005.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. CRC press, 1984.

L. Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

G. Calafiore and B. Monastero. Data-driven asset allocation with guaranteed short-fall probability. In *American Control Conference (ACC), 2012*, pages 3687–3692. IEEE, 2012.

G. Calafiore and M. C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

G. C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

M. Campi and A. Carè. Random convex programs with l_1-regularization: Sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, 2013.

M. Campi and S. Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

M. Carrasco and X. Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.

W. Chen, M. Sim, J. Sun, and C. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Operations Research*, 58(2):470–485, 2010.

X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

Y.-X. Chen and W.-J. Zhu. Note on the strong law of large numbers in a Hilbert space. *Gen. Math.*, 19(3):11–18, 2011.

H. Choi and H. Varian. Predicting the present with google trends. *Econ. Rec.*, 88(s1):2–9, 2012.

S.-C. Chow, M. Chang, et al. Adaptive design methods in clinical trials-a review. *Orphanet J Rare Dis*, 3(11), 2008.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

M. C. Cohen, N.-H. Z. Leung, K. Panchamgam, G. Perakis, and A. Smith. The impact of linear optimization on promotion planning. *Available at SSRN 2382251*, 2014.

Columbia University Center for Pricing and Revenue Management. Dataset cprm-12-001: On-line auto lending, 2012.

M. Conforti, M. Rao, and A. Sassano. The equipartition polytope. i: formulations, dimension and basic facets. *Mathematical Programming*, 49(1):49–70, 1990a.

M. Conforti, M. Rao, and A. Sassano. The equipartition polytope. ii: valid inequalities and facets. *Mathematical Programming*, 49(1-3):71–90, 1990b.

D. R. Cox. *Planning of Experiments*. Wiley, 1958.

Z. Da, J. Engelberg, and P. Gao. In search of attention. *J. Finance*, 66(5):1461–1499, 2011.

R. B. D'Agostino and M. A. Stephens. *Goodness-of-fit techniques*. Dekker, New York, 1986.

H. David and H. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 55(3):98–112, 2010.

L. P. Devroye and T. Wagner. On the l 1 convergence of kernel estimators of regression functions with applications in discrimination. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 51(1):15–25, 1980.

Y. Dodge. *The Oxford dictionary of statistical terms*. Oxford University Press, 2006.

P. Doukhan. *Mixing: Properties and Examples*. Springer, 1994.

R. M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, Cambridge, 2002.

J. Dupačová. The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of Probability and Stochastic Processes*, 20(1):73–88, 1987.

E. Edgington and P. Onghena. *Randomization tests*. CRC Press, Boca Raton, 2007.

J. Edmonds. Paths, trees, and flowers. *Canad. J. Math.*, 17(3):449–467, 1965.

B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapmen and Hall, 1993.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

W. Elmaghraby and P. Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10):1287–1309, 2003.

P. Embrechts, A. Höing, and A. Juri. Using copulae to bound the value-at-risk for functions of dependent risks. *Finance and Stochastics*, 7(2):145–167, 2003. ISSN 0949-2984.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216, 1993.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

C. A. Flores. *Estimation of Dose-Response Functions and Optimal Treatment Doses with a Continuous Treatment*. PhD thesis, University of California, Berkeley, 2005.

G. Gallego and G. Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

M. R. Garey and D. S. Johnson. *Computers and Intractability*, volume 174. Freeman, 1979.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63 (1):3–42, 2006.

S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts. Predicting consumer behavior with web search. *PNAS*, 107(41):17486–17490, 2010.

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

A. S. Goldberger. Structural equation methods in the social sciences. *Econometrica*, pages 979–1001, 1972.

D. Goldfarb and G. Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

R. Greevy, B. Lu, J. H. Silber, and P. Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

A. Gretton, K. Borgwardt, B. Schölkopf, M. Rasch, and E. Smola. A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems 19*, 2007.

M. Grotschel, L. Lovasz, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Springer, New York, 1993.

D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Syst. J.*, 43(1):64–77, 2004.

D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *SIGKDD*, pages 78–87, 2005.

X. S. Gu and P. R. Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4): 405–420, 1993.

Gurobi Optimization Inc. Gurobi optimizer reference manual. `http://www.gurobi.com`, 2013.

G. A. Hanasusanto and D. Kuhn. Robust data-driven dynamic programming. In *Advances in Neural Information Processing Systems*, pages 827–835, 2013.

L. Hannah, W. Powell, and D. M. Blei. Nonparametric density estimation for stochastic optimization with an observable state variable. In *Advances in Neural Information Processing Systems*, pages 820–828, 2010.

B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(03):726–748, 2008.

W. Hardle. *Applied nonparametric regression*, volume 27. Cambridge Univ Press, 1990.

J. M. Harrison, N. B. Keskin, and A. Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 1. Springer, 2001.

J. J. Heckman. Econometric causality. *International Statistical Review*, 76(1):1–27, 2008.

K. Hirano and G. W. Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

T. K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.

J. L. Horowitz. The bootstrap. *Handbook of econometrics*, 5:3159–3228, 2001.

W. T. Huh, R. Levi, P. Rusmevichientong, and J. B. Orlin. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4):929–941, 2011.

J. K. Hunter and B. Nachtergaele. *Applied Analysis*. World Scientific, 2001.

K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467), 2004.

G. W. Imbens and D. B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. Technical report, Technical report, University of Florida. Available at: Optimization Online www. optimization-online. org, 2013.

K. Johnson, B. H. A. Lee, and D. Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. 2014.

J. B. Kadane and T. Seidenfeld. Randomization in a bayesian perspective. *Journal of statistical planning and inference*, 25(3):329–345, 1990.

V. Kaibel, M. Peinhardt, and M. E. Pfetsch. Orbitopal fixing. *Discrete Optimization*, 8(4): 595–610, 2011.

S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793–800, 2008.

N. Kallus. Predicting crowd behavior with big public data. In *WWW*, number 23, pages 625–630, 2014a.

N. Kallus. Optimal a priori balance in the design of controlled experiments. 2014b. URL `http://arxiv.org/abs/1312.0531`.

Y.-h. Kao, B. V. Roy, and X. Yan. Directed regression. In *Advances in Neural Information Processing Systems*, pages 889–897, 2009.

A. Kapelner and A. Krieger. Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 2014.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

N. Karmarkar and R. M. Karp. The difierencing method of set partitioning. Technical report, Technical Report UCB/CSD 82/113, University of California, Berkeley, 1982.

N. Karmarkar, R. M. Karp, G. S. Lueker, and A. M. Odlyzko. Probabilistic analysis of optimum partitioning. *Journal of Applied Probability*, pages 626–645, 1986.

G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1970.

A. J. King and R. J. B. Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2):83–92, 1991.

J. Kingman. Some inequalities for the queue GI/G/1. *Biometrika*, 49(3/4):315–324, 1962.

D. Klabjan, D. Simchi-Levi, and M. Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2): 479–502, 2002a.

A. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, 12(2):479–502, 2002b.

R. Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

S. Kotz, N. Johnson, and D. Boyd. Series representations of distributions of quadratic forms in normal variables. I. central case. *Ann. Math. Statist.*, 38(3):823–837, 1967.

S. Kullback. A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13(1):126–127, 1967.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

S. Lee, T. Homem-de Mello, and A. J. Kleywegt. Newsvendor-type models with decision-dependent uncertainty. *Mathematical Methods of Operations Research*, 76(2):189–221, 2012.

E. Lehmann and J. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics, 2010.

E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998.

R. Levi, G. Perakis, and J. Uichanco. The data-driven newsvendor problem: new bounds and insights. 2012. Working paper.

D. Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

R. J. Little and L. H. Yau. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, 3(2):147, 1998.

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear algebra and its applications*, 284(1):193–228, 1998.

P. McCullagh, J. A. Nelder, and P. McCullagh. *Generalized linear models*. Chapman and Hall, London, 2 edition, 1989.

D. Mcdonald, C. Shalizi, and M. Schervish. Estimating beta-mixing coefficients. In *AISTATS*, pages 516–524, 2011.

R. McHugh and J. Matts. Post-stratification in the randomized clinical trial. *Biometrics*, pages 217–225, 1983.

S. Mertens. A physicist's approach to number partitioning. *Theoretical Computer Science*, 265(1):79–108, 2001.

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *NIPS*, pages 1097–1104, 2008.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

A. Mokkadem. Mixing properties of arma processes. *Stochastic Process. Appl.*, 29(2):309–315, 1988.

K. L. Morgan, D. B. Rubin, et al. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

C. Morris. A finite selection model for experimental design of the health insurance study. *Journal of Econometrics*, 11(1):43–61, 1979.

Mosek, APS. The MOSEK optimization software. `http://www.mosek.com/`, 2009.

A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.

E. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.

K. Natarajan, P. Dessislava, and M. Sim. Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science*, 54(3):573–585, 2008.

A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

W. K. Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(02):1–21, 1994.

S. Oren, S. Smith, and R. Wilson. Pricing a product line. *Journal of Business*, pages S73–S99, 1984.

E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076, 1962.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

J. Pearl. Remarks on the method of propensity score. *Statistics in Medicine*, 28(9):1415–1416, 2009.

R. Phillips. *Pricing and revenue optimization*. Stanford University Press, 2005.

R. Phillips, A. S. Simsek, and G. VanRyzin. Endogeneity and price sensitivity in customized pricing. *Columbia University Center for Pricing and Revenue Management Working Paper*, 4, 2012.

S. J. Pocock and R. Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115, 1975.

I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.

A. Prékopa. *Stochastic Programming*. Kluwer Academic Publishers, Dordrecht, 1995.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

J. Rice. *Mathematical statistics and data analysis*. Thomson/Brooks/Cole, Belmont, 2007.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

R. T. Rockafellar. *Convex analysis*. Princeton university press, 1997.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer, 1998.

T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *J. Risk*, 2:21–42, 2000.

P. R. Rosenbaum. Interference between units in randomized experiments. *J. Amer. Statist. Assoc.*, 102(477), 2007.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

G. G. Roussas. Exact rates of almost sure convergence of a recursive kernel estimate of a probability densiy function: Application to regression and hazard rate estimation. *Journal of Nonparametric Statistics*, 1(3):171–195, 1992.

H. L. Royden. *Real Analysis*, volume 3. Prentice Hall, 1988.

D. B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a): 318–328, 1979.

D. B. Rubin. Comment: Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

D. B. Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484), 2008.

D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

C. Rudin and G.-Y. Vahn. The big data newsvendor: Practical insights from machine learning. 2014.

P. Rusmevichientong and H. Topaloglu. Robust assortment optimization in revenue management under the multinomial logit choice model. *Operations Research*, 60(4):865–882, 2012.

L. J. Savage. The foundations of statistics reconsidered. In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 575–586. University of California Press, 1961.

H. Scarf. A min-max solution of an inventory problem. In K. J. Arrow, S. Karlin, and H. Scarf, editors, *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Sanford University Press, Stanford, 1958.

M. Scarsini. Multivariate convex orderings, dependence, and stochastic equality. *Journal of applied probability*, 35(1):93–103, 1998.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statis.*, 41(5):2263–2291, 2013.

J. S. Sekhon. The neyman-rubin model of causal inference and estimation via matching methods, 2008.

M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

A. Shapiro. On duality theory of conic linear problems. In M. A. Goberna and M. A. López, editors, *Semi-Infinite Programming: Recent Advances*, pages 135–165. Kluwer Academic Publishers, Dordrecht, 2001.

A. Shapiro. Monte carlo sampling methods. *Handbooks Oper. Res. Management Sci.*, 10: 353–425, 2003.

A. Shapiro and P. R. Andrzej. *Stochastic programming*. Elsevier, 2003.

A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.

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`.

I. Shrier. Propensity scores. *Statistics in Medicine*, 28(8):1317–1318, 2009.

P. Spirtes. Introduction to causal inference. *The Journal of Machine Learning Research*, 11: 1643–1662, 2010.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006.

M. A. Stephens. Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistics Society B*, 32(1):115–122, 1970.

M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.

Student. Comparison between balanced and random arrangements of field plots. *Biometrika*, pages 363–378, 1938.

O. Thas. *Comparing Distributions*. Springer, New York, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

V. Vapnik. *The nature of statistical learning theory*. springer, 2000.

A. Wächter and L. Biegler. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

A. Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, pages 165–205, 1949.

H. Walk. Strong laws of large numbers and nonparametric estimation. In *Recent Developments in Applied Probability and Statistics*, pages 183–214. Springer, 2010.

Z. Wang, P. W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven newsvendor problems. Technical report, Working paper, 2009.

J. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58 (301):236–244, 1963.

G. Watson. Smooth regression analysis. *Sankhyā A*, pages 359–372, 1964.

T. E. Wheldon. *Mathematical models in cancer research*. Adam Hilger, Bristol, 1988.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. 2013. Working paper.

J. Žáčková. On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky*, 91(4):423–430, 1966.

K. Ziegler. On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric Statistics*, 14(6):749–774, 2002.