Stochastic and Convex Optimization in Statistical Estimation

By

ROBERT L. BASSETT

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Roger J-B Wets, Chair

James Sharpnack

Matthias Koeppe

Committee in Charge

2018

Contents

Robert L. Bassett
June 2018
Mathematics

Stochastic and Convex Optimization in Statistical Estimation

Abstract

This dissertation is an exploration at the interface of optimization and statistics. It focuses on four specific applications. We begin by considering nonparametric density estimation, and formulate a new estimator by applying a fusion penalty to a maximum log-likelihood objective. We place particular emphasis on the computation of these estimators, because ease of calculation is essential to density estimation techniques being useful in practice. This leads us naturally into convex optimization, as we compare and contrast the performance of different algorithms on this problem. From density estimation, we move next into Bayesian point estimation. We consider the relationship between Bayes estimators and Maximum-a-Posteriori (MAP) estimators. It is commonly accepted that one of these estimators is a limiting case of the other. In fact, this is not the case without some additional assumptions. We correct this inaccuracy and then provide conditions for the traditional results to hold. Because both estimators are defined in terms of optimization problems, we use the theory of variational analysis–a convergence theory with origin in the optimization community–as a foundation for these corrections. The Bayesian point-estimation theme continues in our next application: MAP estimation in a dynamical system with log-concave noise. We formulate this problem as a convex program, and show that its convex-analytic dual can be interpreted as an problem of optimal control. This connection relates to the celebrated Kalman filter, where a "duality" between the solutions of optimal control and estimation problems was an essential component of its original derivation. We continue with the theme of convex-analytic duality in our final application: set-valued portfolio optimization. Set-valued portfolio optimization is natural when modeling an uncertain financial exchange where there is no clear notion of a *numeraire*, or universal measure of value. We show that multistage portfolio optimization in a set-valued conic market model has a set-valued dual. This dual can be interpreted as one of finding a consistent pricing system, analogous to historical portfolio duality results. Furthermore, we prove that strong duality holds between these two set-valued problems. These four applications demonstrate the value of research at the

intersection of statistics and optimization. Taken together, the applications in this dissertation make the case for unifying statistical estimation and optimization in order to further research in both fields.

Acknowledgments

I would like express my sincere gratitude to the mentors who have supported me through this journey. This manuscript is evidence of their infinite patience. I would like to thank Roger J-B Wets for his mentorship and words of wisdom, and James Sharpnack for his guidance and intellectual enthusiasm. Both men were excellent role models–both personally and professionally–for a young graduate student. I am grateful to have had their support. I thank Matthias Koeppe for his careful comments and corrections. I wish also to thank my friends and collaborators, Johannes Royset and Julio Deride, for fruitful conversation and valuable insight.

I would also like to thank my educators, whose dedication to their craft kept the spark of curiousity alive along the way: Javier Trigos, Sophia Raczcowski, Charles Lam, Brian Jean, Tim Cater, and Anthony Perone. You do make a difference.

I would like to thank my family. Their unconditional love and support provides the firm foundation on which I build. Especially, I would like to thank Grace, for keeping my heart full and making this time so pleasant.

Finally, I would like to acknowledge financial support from the Graduate Assistance in Areas of National Need (GAANN) Fellowship, the Programme Gaspard Monge pour l'Optimisation et la recherche opérationnelle (PGMO), and the Science of Autonomy Program at the Office of Naval Research under grant number N00014-17-2372.

CHAPTER 1

Introduction

Decision-making under uncertain conditions is a universal experience, and because of this universality, there has been an enormous amount of scientific effort devoted to the topic. From the ancient Greeks' consideration of the arithmetic mean, to modern methods in machine learning, each generation has found ways to cope with this issue. The process of decision-making in the face of uncertainty can be roughly divided into two steps: defining and then computing an appropriate decision. Statistical estimation is the theory of defining decisions in the presence of uncertainty–often by minimizing or maximizing a context-dependent function. Mathematical optimization, on the other hand, is concerned with computation: constructing efficient methods to minimize or maximize objective functions. Mathematical optimization and statistical estimation were rightfully acknowledged as important subdisciplines of applied mathematics in the early and mid 20th century, through the creation of distinct statistics and operations research departments at many universities. Though these research areas are natural partners in progress for making decisions under uncertainty, their establishment as separate disciplines can make communication between the fields difficult.

This dissertation is one attempt to unify the research efforts of these communities, towards the goal of establihing new theory and methods for decision-making under uncertainty. It is divided into five chapters, where we consider four different projects which rely jointly on contributions from statistical estimation and optimization. In the remainder of this introduction, we outline the remaining chapters and highlight our contributions to each area. In each of the next chapters, we present a project and derive our results. The second chapter considers a problem in nonparametric density estimation, where sparse quadratic programming plays an important role in our ability to compute solutions. The third chapter is an application of tools from variational analysis–a mathematical theory developed for optimization problems–to the convergence of Bayesian point estimators. In the fourth chapter we consider a data-driven estimation problem, and then apply convex duality to connect it to the theory of optimal control. In the fifth and final chapter, we consider a set-valued portfolio optimization problem, and derive a duality result in that setting.

1.1. Fused Density Estimation

In chapter 2 we define a new piecewise constant density estimator, called the *fused density estimator* (FDE). We show that fused density estimation, though originally formulated as a variational problem, has a finite-dimensional representation as the solution to a sparse quadratic program. We also give rates of convergence for this estimator. The rate we derive achieves the minimax rate (in Hellinger distance) over all densities with logarithm of bounded variation. The FDE formulation has a natural extension to geometric networks, a network model useful in infrastructure applications, so we also extend our computational and theoretical results to this setting.

The definition of a fused density estimator is motivated by the success of total variation penalties for image denoising [**ROF92**] and fusion penalties for model-selection and regularization [**TSR+05**]. In both of these examples, a quadratic term, with minimum that fits the observed data, is added to a total variation, or fusion penalty term, to form an objective function. The result is an estimator that balances data fidelity and "simple" solution structure. We leverage this same intuition in the FDE definition: a maximum log-likelihood term is added to the total variation of the logarithm of the density. The resulting objective is then minimized over all densities for which the penalty term is finite. Though this nonparametric problem is formulated as an optimization problem over an infinite-dimensional function space, we reduce it to a finite-dimensional sparse quadratic program, thus gaining all of the computational tractability of that problem class.

Though this work is not the first to consider total variation penalized maximum likelihood density estimation [**ST10**, **KM06**], our penalization of the log-density is a new contribution. This choice of penalty is computationally justified by the QP reformulation outlined in the previous paragraph. It is theoretically justified by the rate of convergence we derive: a fused density estimator has squared-Hellinger rate of $n^{-2/3}$. To our knowledge, this is the first result of its kind for nonparametric total-variation penalized density estimation. Furthermore, this rate achieves the minimax rate over all densities for which the penalty is finite, which provides strong theoretical support for the use of FDEs.

Total-variation penalties can be naturally extended to settings involving networks [**WSST16**]. We leverage this strength to extend FDEs to the setting of geometric networks–networks where edges are identified with compact intervals. Geometric network models have extensive applications to infrastructure networks, in areas such as transportation studies, water resource management, and national defense. In this framework,

observations occur along the edges of a geometric network, and densities are defined along these edges as well. FDEs can then be used to find densities from data on a geometric network–such as traffic accidents on a road network or contaminant detection in a water network. This extension of FDEs to the setting of geometric networks makes them applicable to a variety of application areas, where a network-valued density estimate can be a qualitative tool to benefit decision makers.

1.2. MAP Estimators as a Limit of Bayes Estimators

In chapter 3, we investigate a folklore theorem relating two methods of Bayesian point estimation. We provide a counterexample to this theorem, even though it commonly appears in Bayesian statistics literature. The issue with the theorem is subtle, yet cannot be overlooked. Both estimators are defined in terms of optimization problems–one in terms of maximizing a posterior likelihood and another as minimizing posterior risk. The relationship between them is claimed to be one of limits, but in order for limits of functions to imply limits of their maximizers we must use an appropriate topology for convergence of functions. These topological considerations lead us into the theory of variational analysis, which we use to correct the theorem and specify the additional assumptions required.

The estimators we consider are the Maximum-a-Posteriori estimator (MAP) estimator, and the Bayes estimator with shrinking 0-1 loss. Minimizing Bayes risk under 0-1 loss corresponds to choosing an estimator with high posterior likelihood averaged value over a small neighborhood. MAP estimators, on the other hand, are simply maximizers of the posterior distribution. The folklore theorem is as follows: when the size of the neighborhood over which the average value is taken shrinks to zero, the corresponding Bayes estimators converge to a MAP estimator. This statement seems intuitive, which contributes to its widespread acceptance. An argument which is sometimes used to justify the theorem is that convergence of local averaging to the posterior density gives continuous convergence [**Gew05**], but in fact this is not enough to guarantee that the limit of maximizers is itself a maximizer [**RW09**]. We provide a counterexample which serves to make this point.

The notion of function convergence required to guarantee convergence of maximizers is hypo-convergence, which can be interpreted as set convergence of the functions' hypographs [**RW09**]. Hypo-convergence is an important component in the theory of variational analysis: a mathematical framework for approximating optimization problems. Though many statistical estimators are defined in terms of optimization problems,

variational analysis is not often used by statisticians. We use this theory to prove hypo-convergence of the sequence of mollifying approximates which define Bayes estimators. We then provide a very mild level-set condition on the posterior density for the folklore theorem to hold. The level-set condition covers all distributions of practical interest, including quasi-concave and log-concave families. In correcting this theorem, we hope to demonstrate the value of applying variational-analytic tools to problems in statistical estimation.

1.3. Log-Concave Duality in Estimation and Control

In chapter 4, we consider the problem of estimating the state of a dynamical system. The system is in discrete time, with linear, possibly time-dependent dynamics, and an initial condition and additive noise that are generated according to a (possibly time-dependent) log-concave density. We assume, furthermore, that we receive a noisy linear measurement of the system at each time step, again corrupted by additive log-concave noise. Our goal is to reconstruct the past and present states of the system. We formulate this as a Maximum-a-Posteriori (MAP) estimation problem. In his seminal paper [**Kal60**], Kalman considered an online version of this problem, where each noise term was restricted to be Gaussian, a special case of the log-concave setting proposed here. In order to derive his solution, the celebrated Kalman filter, he relied on a notion of "duality" between this estimation problem and the Linear-Quadratic Regulator from optimal control. Taking inspiration from Kalman, we investigate the duality structure in our log-concave estimation problem.

Kalman's duality referred to an duality of solutions to equations: solutions to the algebraic Ricatti equations of the control problem are in one-to-one correspondence with the solutions to the system of equations governing covariance propagation. Instead, we consider convex-analytic duality. Solution-duality can often be derived as a result of convex-analytic duality, whenever one is able to prove strong duality between the primal and dual problems. We derive a dual problem for our log-concave estimation problem. This dual corresponds to an optimal control problem. We prove that strong-duality holds between the problems under a mild constraint qualification (as in [**Roc97**]), yielding a bijection between solutions to the estimation and optimal control problems.

We also consider the special case that the noise terms are *log-piecewise linear quadratic*. This is a special case of the more general log-concave noise, when the dynamical system, measurement, and initialization noise each have densities of the form $e^{-\rho_{U,M}}$, where $\rho_{U,M}$ is a piecewise-linear quadratic function in

the sense of [RW09]. In this case, the dual optimal control problem has a simple closed-form expression. We also show that strong duality holds between the estimation and control problems without any additional constraint qualification. We hope that this result serves to motivate researchers in estimation and control to unify their efforts–our work indicates that researchers specializing in the computation for optimal control problems can use their expertise to solve problems in estimation, and vice versa. Though the log-concave case is quite general, this project provides inspiration for further investigation of the setting where noise is not log-concave, in order to extend these theoretical and computational insights to more general noise terms.

1.4. Duality in Set-Valued Conic Market Models

In chapter 5, we consider a multi-stage portfolio optimization problem in a financial market with proportional transaction costs. The financial market is described by Kabanov's model of foreign exchange markets [Kab99], formulated over a finite probability space, and in discrete time steps with a finite horizon. We assume proportional transaction costs, so that exchanging between assets is penalized proportional to the size of the transaction. Furthermore, our framework allows us to consider vector-valued portfolios under a partial ordering. The primary advantage of this model is that it allows us to directly compare portfolios based on their number of assets, without having to translate into currency for the sake of comparison. In the context of international exchange markets, a model that requires liquidation into some numeraire at terminal time can be a significant burden, particularly when modeling transaction costs. Treating one currency as metric of comparison places portfolios dominated by foreign assets at a disadvantage. However, the generality of this model comes with additional complexity in its analysis.

We embed this problem in a set-optimization framework [HHL$^+$ng], bypassing a purely vector-valued model for the sake of additional generality and mathematical novelty. An agent's utility function is then a convex, set-valued function, as opposed to a strict ordering in the traditional setting. In set-optimization, one attempts to optimize over sets, where the ordering is a cone-induced partial ordering. This setting can be seen as a direct extension of traditional convex optimization over the extended real numbers. The drawback of a set-optimization framework is that it requires additional mathematical subtlety. As an example, defining appropriate notions of minimum and minimizer requires is not a direct extension of the real-valued setting [HL14]–these pave the way for the definitions of a solution and a full solution (one which contains all other

solutions). The drawback of constructing new mathematical theory is compensated by the additional realism the complexity allows us to incorporate into our model.

Our contribution is the formulation of a set-valued dual to the portfolio optimization problem. The dual solutions can be interpreted as consistent pricing systems, analogous to the interpretation of dual vectors as prices in traditional portfolio optimization. We prove that strong duality holds between the two problems, under a suitable extension of the Slater condition to the set-valued framework. Lastly, we include an illustrative example, working out the dual problem in detail. This example serves to provide tangibility to the mathematical abstraction presented in this chapter.

CHAPTER 2

Fused Density Estimation: Theory and Methods

2.1. Introduction

In the pantheon of statistical tools, the histogram remains the primary way to explore univariate empirical distributions. Since its introduction by Karl Pearson in the late 19th century, the form of the histogram has remained largely unchanged. In practice, the regular histogram, with its equal bin widths chosen by simple heuristic formulas, remains one of the most ubiquitous statistical methods. Most methodological improvements on the regular histogram have come from the selection of bin widths—this includes varying bin widths to construct irregular histograms—motivated by thinking of the histogram as a piecewise constant density estimate. In this work, we study a piecewise constant density estimation technique based on total variation penalized maximum likelihood. We call this method fused density estimation (FDE). We extend FDE from irregular histogram selection to density estimation over geometric networks, which can be used to model observations on infrastructure networks like road systems and water supply networks. The use of fusion penalties for density estimation is inspired by recent advances in theory and algorithms for the fused lasso over graphs [**PSST16**, **WSST16**]. Our thesis, that FDE is an important algorithmic primitive for statistical modeling, compression, and exploration of stochastic processes, is supported by our development of fast implementations, minimax statistical theory, and experimental results.

In 1926, [**Stu26**] provided a heuristic for regular histogram selection where, naturally, the bin width increases with the range and decreases with the number of points. The regular histogram is an efficient density estimate when the underlying density is uniformly smooth, but irregular histograms can 'zoom in' to regions where there is more data and better capture the local smoothness of the density. A simple irregular histogram, known as the equal-area histogram, is constructed by partitioning the domain so that each bin has the same number of points. [**DM09**] noted that the equal-area histogram can often split bins unnecessarily

This chapter is based on joint work with James Sharpnack.

when the density is smooth and merge bins when the density is variable, and proposed a heuristic method to correct this oversight. Recently, [**LMSW16**] proposed the essential histogram, an irregular histogram constructed such that it has the fewest number of bins and lies within a confidence band of the empirical distribution. While theoretically attractive, in practice its complex formulation is intractable and requires approximation. If the underlying density is nearly constant over a region, then the empirical distribution is well approximated locally by a constant, and hence the essential histogram will tend to not split this region into multiple bins. Such a method is called *locally adaptive*, because it adapts to the local smoothness of the underlying density.

In Figure 2.1, we compare FDE to the regular histogram, both of which have 70 bins. Because FDE can be thought of as a bin selection procedure, in this example, we recompute the restricted MLE after the bin selection, which is common practice for model selection with lasso-type methods. We see that with 70 bins the regular histogram can capture the variability in the left-most region of the domain but under-smooths in the right-most region. We can compare this to FDE which adapts to the local smoothness of the true density. As a natural extension of 1-dimensional data, we will consider distributions that lie on geometric networks—graphs where the edges are continuous line segments—such as is common in many infrastructure networks. Another motivation to use total variation penalties is that they are easily defined over any geometric network, in contrast to other methods, such as the essential histogram and multiscale methods. Figure 2.2 depicts the FDE for data in downtown San Diego. The geometric network is generated from the road network in the area, and observations on the geometric network are the locations of eateries (data extracted from the OpenStreetMap database [**Ope17**]).

Without any constraints, maximum likelihood will select histograms that have high variation (as in Figure 2.1), so to regularize the problem, we bias the solution to have low total variation. Total variation penalization is a popular method for denoising images because solutions tend to identify homogeneous regions in the underying image. This procedure, also known as the 2-D fused lasso, [**TSR+05**], is a least-squares method with a fusion penalty that pulls adjacent pixels toward one another. Fast optimization methods, such as alternating direction method of multipliers, projected Newton methods, and split Bregman iteration, have been developed for total variation denoising [**ROF92**, **BT09**]. Total variation penalized methods like the 2-D fused lasso can be naturally extended to signal processing over graphs by thinking of vertices as pixels [**WSST16**]. Signal processing over graphs refers to methods that denoise and infer signal within

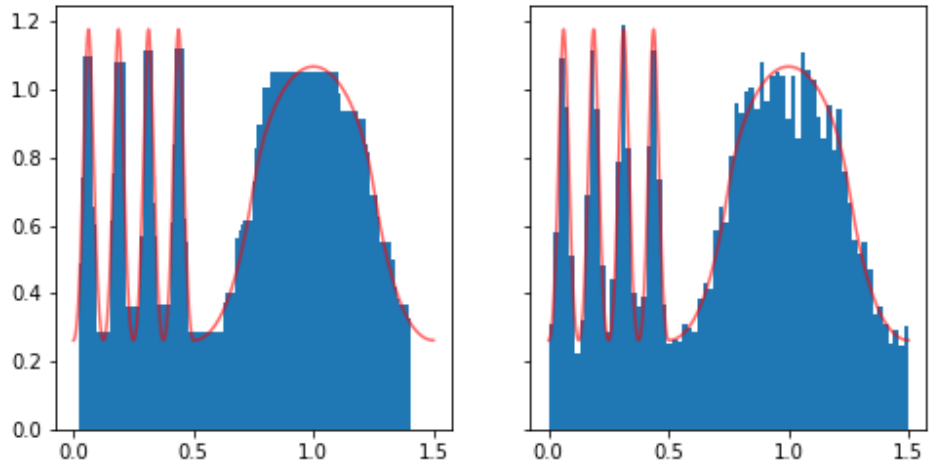


FIGURE 2.1. A comparison of FDE (left) and the regular histogram (right) of 10,000 data points from a density (red) with varying smoothness—both have 70 bins.

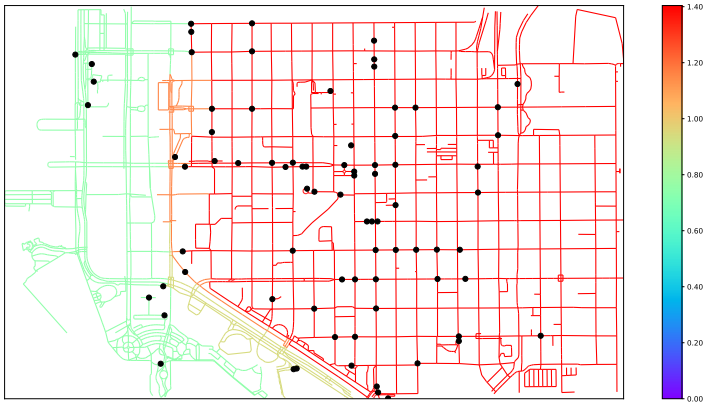


FIGURE 2.2. FDE for the location of eateries in downtown San Diego.

noisy observations over vertices of a general graph [SNF+13]. In general, sensor networks can be viewed as graphs where the vertices, the sensors, produce observations. Distributions over geometric networks, which we consider here, are distinguished from this literature by the fact that observations can occur at any point along an edge of the network. This leads to a variational density estimation problem, which we reduce to a finite dimensional formulation.

> Contribution 1. We show that the FDE is equivalent to a total variation penalized weighted least squares problem, conferring fast optimization tools to the density estimation setting.

In order to justify the use of FDE, we will analyze the statistical performance of FDE for densities of log-bounded variation over geometric networks. The majority of statistical guarantees for density estimates control some notion of divergence between the estimate and the true underlying density. Several authors have used the L_2 loss (mean integrated square error) to evaluate their methods for tuning the bin width for the regular histogram [**Sco79**, **FD81**, **BM97**]. While it is appealing to use L_2 loss, it is not invariant to choice of base measure, and divergence measures such as L_1, Hellinger loss, and the Kullback-Leibler (KL) divergence are preferred for maximum likelihood—an idea pioneered by Le Cam [**LCY12**] and furthered by [**DG85**, **HW88**]. By appealing to Hellinger loss, [**BR06**] proposed a method for optimal choice of the number of bins in a regular histogram, and we will similarly focus on Hellinger loss.

> Contribution 2. We provide a non-parametric Hellinger distance rate guarantee for FDE in the univariate case, over densities of log-bounded variation, which achieves the minimax lower bound rate.

When the log-density lies in a Sobolev space, an appropriate non-parametric approach to density estimation is maximum likelihood with a smoothing splines penalty [**Sil82**]. The smoothing spline method is not locally adaptive because it does not adjust to the local smoothness of the density or log-density. Epi-splines, [**RW13a**], are density estimates formed by maximizing the likelihood such that the density, or log-density, has a representation in a local basis and lies in a prespecified constraint set. [**DJKP96**] studied wavelet thresholding for density estimation and proved L_p rate guarantees. [**KK96**] studied maximum likelihood with a log-density wavelet estimator and showed that it achieves minimax rates for KL-divergence when the log-density is in a Besov space. In a related work, [**KK00**] considered log-spline density estimation from binned data with stepwise knot selection. FDE differs from this work because we minimize a variational objective directly, through a representer theorem, and this can be solved via a weighted fused lasso, as opposed to a stepwise selection procedure. [**WN07**] used a recursive partitioning approach to form adaptive polynomial estimates of the density, a similar approach to wavelet decomposition. Such multiscale methods have well known local adaptivity properties, but extending wavelets to geometric networks is a cumbersome task, while total variation penalties extends very naturally to the geometric network setting.

Contribution 3. We prove that the same Hellinger distance rate guarantee for the univariate case also holds for any connected geometric network.

So, it also turns out that the theoretical results in the univariate setting can be extended to the geometric network case. Before we can consider the theoretical performance of FDE, we will define the setting and methodology in more detail.

2.1.1. Problem Statement. When considering road systems and water networks, we observe that individual roads or pipes can be modeled as line segments, and the entire network constructed by joining these segments at nodes of intersection. Mathematically, we model this as a *geometric network* G, a finite collection of nodes V and edges E, where each edge is identified with a closed and bounded interval of the real line. Each edge in the network has a well-defined notion of length, inherited from the length of the closed interval. We fix an orientation of G by assigning, for each edge $e = \{v_i, v_j\}$, a bijection between $\{v_i, v_j\}$ and the endpoints of the closed interval associated with e. This corresponds to the intuitive notion of "gluing" edges together to form a geometric network. A *point* in a geometric network G is an element of one of the closed intervals identified with edges in G, modulo the equivalence of endpoints corresponding to the same node. Because we only discuss geometric networks in this chapter, we will often refer to them as networks.

A real-valued function g, defined on a geometric network G, is a collection of univariate functions $\{g|_e\}_{e\in E}$, defined on the edges of G. We require that the function respects the network structure, by which we mean that for any two edges e_1 and e_2 which are incident at a node v, $g|_{e_1}(v) = g|_{e_2}(v)$. We abuse notation slightly–by referring to $g|_e(v)$, we mean g evaluated at the endpoint of the interval identified to v. A geometric network G inherits a measure from its univariate segments in a natural way, as the sum of the Lebesgue measure along each segment. With this measure we have a straight-forward extension of the Lebesgue measure to G, making G a measurable space.

For any random variable taking values on the network G, we will assume that the measure induced by the random variable is absolutely continuous with respect to the base measure, dx, and so has density f. We will abuse notation by using dx to refer to both the Lebesgue measure and the base measure on a geometric graph; which of these we mean will be clear from its context. Furthermore, we assume that the density is non-zero everywhere, so that its logarithm is well defined. Moreover, we will assume the log-density is not arbitrarily variable, and for this purpose we will use the notion of total variation. Let $B \subseteq \mathbb{R}$. The *total*

variation of a function $g : B \to \mathbb{R}$ is defined as

$$\mathrm{TV}(g) = \sup_{P \subset B} \sum_{z_i \in P} |g(z_i) - g(z_{i+1})| .$$

The supremum is over all partitions, or finite point-subsets P, of B. For a real-valued function g defined on a network G, we extend the univariate definition to

$$\mathrm{TV}(g) = \sum_{e \in E} \mathrm{TV}(g|_e).$$

One advantage of the use of the TV penalty is that is it invariant to the choice of the segment length in the geometric network, so scaling the edge by a constant multiplier leaves the total variation unchanged. As a consequence fused density estimation will be invariant to the choice of edge length.

Let f_0 be a density on a geometric network G, and $x_1, \ldots, x_n$ an independent sample identically distributed according to f_0. Let $P_n = \frac{1}{n} \sum \delta_{x_i}$ be the empirical measure associated to the sample. We let P_n act on a function, by which we mean that we take the expectation of that function with respect to P_n. So for any function f,

$$P_n(f) = \int f \, dP_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i).$$

We will also use $P(f)$ to denote $\int f \, dP$ for non-empirical measures P.

Fix $\lambda \in \mathbb{R}^+$. A *fused density estimator* (FDE) of f_0 is a density $\hat{f} = \exp(\hat{g})$, such that the log-density $\hat{g}$ is minimizer or the following program,

$$\min \; -P_n(g) + \lambda \, \mathrm{TV}(g) \text{ s.t. } \int e^g \, dx = 1 \tag{2.1}$$

where the minimum is taken over all functions $g : G \to \mathbb{R}$ for which the expression is finite and the resulting f is a valid density. That is, $f \in \mathscr{F}$ and $g \in \mathscr{G}$ where

$$\mathscr{F} = \{e^g : g \in \mathscr{G}\}, \quad \mathscr{G} = \left\{ g : \mathrm{TV}(g) < \infty, \; \int_G e^g \, dx = 1 \right\}.$$

The set $\mathscr{F}$ will be referred as the set of densities with *log-bounded variation*. Indeed, the integration constraint on elements of $\mathscr{G}$ makes them log-densities. Note that densities in $\mathscr{F}$ are necessarily bounded above and away from zero, as a result of the total variation condition.

The program in (2.1) is variational, because it is a minimizer over an infinite dimensional function space. It is quite common for variational problems in non-parametric statistics to involve a reproducing kernel Hilbert space (RKHS) penalty, as opposed to a total variation penalty [**Wah90**]. In the RKHS setting setting, the Hilbert space allows us to establish representer theorems, which reduce the variational program to an equivalent finite dimensional one, so that it can be solved numerically. Functions of bounded variation, on the other hand, is an example of a more general Banach space, so RKHS results cannot be applied to this setting. In the next section, we discuss representer theorems for (2.1), and further show that it can be solved using a sparse quadratic program.

2.2. Computation

In this section we provide results toward the computation of fused density estimators. The key challenge is the variational formulation of the Fused Density Estimator (2.1). To this end, we prove that solutions to the variational problem can be finitely parametrized. Moreover, we show that after applying this representer theorem, the finite-dimensional analog of (2.1) has an equivalent formulation as a total variation penalized least-squares problem. Our main theorem of this section, which reduces the computation of a fused density estimator to a weighted fused-lasso problem, follows.

THEOREM 2.2.1. *Fix $\lambda \in \mathbb{R}^+$. The corresponding FDE $\hat{f}$ can be finitely represented as vectors $\hat{z}$ and $\hat{h}$. Furthermore, there is a diagonal, data-dependent matrix S with nonnegative entries, sparse matrices D_1 and D_2 depending on both the graph structure and choice of λ, and a vector w such that $\hat{z}$ and $\hat{p}$ are given as the minimizer of the following sparse, total variation regularized quadratic program.*

$$\min_{z,h} \frac{1}{2} z^\top S z + w^\top z + ||D_1 z + D_2 h||_1 .$$

The details of this theorem, by which we mean the constructions of S, D_1, D_2, and w, and the connection between $\hat{z}$, $\hat{h}$ and $\hat{f}$, will be given later in this section.

Theorem 2.2.1 demonstrates that the FDE (2.1) can be solved as a specific incarnation of the generalized lasso, for which there are well known fast implementations [**AT16**]. In practice we will solve the dual to this problem, which we discuss in Theorem 2.2.5. Theorem 2.2.4 is a precise restatement of Theorem 2.2.1. In order to prove it, we proceed through a series of important lemmas. Lemma 2.2.2 transforms the

FDE problem from a constrained to unconstrained one by removing the integration constraint. From this new formulation, Lemma 2.2.3 shows that the search space for the fused density estimator problem can be reduced from functions of bounded variation to an equivalent, finite-dimensional version. Theorem 2.2.4 performs the final step in the proof–demonstrating that the previously derived finite-dimensional problem can be solved using a ℓ_1 penalized quadratic program. The last subsection in this section is tangential, but sheds further light on the structure of fused density estimators. Proposition 2.2.6, which we refer to as the Ordering Property, qualifies the local-adaptivity of fused density estimators by describing their local structure. Omitted proofs can be found in Appendix 2.6.

2.2.1. Main Computational Results. Our first lemma reduces the fused density estimator problem, (2.1), to an unconstrained program where the integral constraint is incorporated into the objective. This result is originally due to Silverman [**Sil82**], who proved the result in the context of univariate density estimation and Sobolev-norm penalties. Minor modifications allow us to extend it to geometric networks and the non-Sobolev total variation penalty.

LEMMA 2.2.2. *Let $\hat{g}$ be a solution to*

$$\min -P_n(g) + \lambda \, \mathrm{TV}(g) + \int_G e^g \, dx \tag{2.2}$$

Then $\hat{g}$ satisfies $\int_G e^{\hat{g}} \, dx = 1$.

We remark that the objective in Lemma 2.2.2 is equivalent to total variation penalized Poisson process likelihood, where the log-intensity is g, so our computations also apply to that setting. Lemma 2.2.2 gives that the fused density estimator definition (2.1) can instead be solved by the unconstrained problem 2.2 over all functions g on G of bounded variation. An alternative interpretation of the lemma is that the Lagrange multiplier associated to the constraint in (2.1) is 1. The next lemma reduces the unconstrained problem (2.2) to an equivalent finite-dimensional version. The proof technique is analogous to similar results in [**MvdG+97**]. In the context of Reproducing Kernel Hilbert Spaces, results that reduce variational problem formulations to finite-dimensional analogs are referred to as *representer theorems*, eg. [**Wah90**]. We will also use this language to describe our result, even though we are in a more general Banach space setting. The result demonstrates that FDEs have large, piecewise constant regions, which is a well known property of fusion penalties [**TSR+05**, **KKBG09**, **WSST16**].

LEMMA 2.2.3 (Representer Theorem). *A fused density estimator $\hat{g}$ is piecewise constant. All discontinuities are contained in the set $\{x_1, ..., x_n\} \cup V$, the observations and the points of G identified with nodes.*

Using Lemma 2.2.3, we can parametrize fused density estimators with three finite-dimensional vectors: the fused density estimator at the observation points, p, the fused density estimator at the vertices of G, k, and the piecewise constant values of the fused density estimator, c. Furthermore, let s be a vector, of the same length as c, such that s_i is the length of the segment corresponding to the value c_i. For simplicity, we will assume that no two observations occur at the same location, a condition that we can and will relax in the remark following Theorem 2.2.4.

Let n_e denote the number of observations along edge e. We will denote by $p_{e,i}$ the value, in the vector p, at the ith observation along edge e. We adopt similar conventions for the vectors c and s. We denote by k_v the value in k at the vertex v. For a given node v, let $\mathrm{inc}(v)$ denote the set of edges which are incident to v and denote by $c_{e,v}$ the segment in c which is incident to v. The problem (2.2) becomes

$$\min_{p,c,k} \sum_{e \in E} \left\{ -\frac{1}{n} \sum_{i=1}^{n_e} p_{e,i} + \lambda \sum_{i=1}^{n_e} |p_{e,i} - c_{e,i}| + |p_{e,i} - c_{e,i+1}| + \sum_{i=1}^{n_e+1} s_{e,i} e^{c_{e,i}} \right\} + \lambda \sum_{v \in V} \sum_{e \in \mathrm{inc}(v)} |k_v - c_{e,v}|$$

The first summand, over the edges in E, gives the log-likelihood term, the total variation along an edge, and the integration term. The second summand gives the total variation at nodes of the geometric graph.

Let F denote the objective function in this optimization problem. Assume that an optimal solution $\hat{p}, \hat{c}, \hat{n}$ exists. Of course, $\hat{p} \in \mathrm{argmin}_p \; F(p, \hat{c}, \hat{k})$. Define $\tilde{F}(p) = F(p, \hat{c}, \hat{k})$. Because the function P is convex, a necessary and sufficient condition for optimality is that zero lies within the subdifferential of $\tilde{F}$, $\partial \tilde{F}$, which

consists of all subgradients. The subdifferential can be computed exactly. By [Roc15, Theorem 23.8],

$$\left(\partial\tilde{F}(p)\right)_{e,i} = \begin{cases} -\frac{1}{n} + 2\lambda & p_i > c_{e,i} \text{ and } p_i > c_{e,i+1} \\ \{-\frac{1}{n} + \lambda + \lambda\alpha : \alpha \in [-1,1]\} & p_i = c_{e,i} \text{ and } p_i > c_{e,i+1} \\ \{-\frac{1}{n} + \lambda + \lambda\alpha : \alpha \in [-1,1]\} & p_i > c_{e,i} \text{ and } p_i = c_{e,i+1} \\ \{-\frac{1}{n} + \lambda\alpha + \lambda\beta : \alpha,\beta \in [-1,1]\} & p_i = c_{e,i} = c_{e,i+1} \\ \{-\frac{1}{n} - \lambda + \lambda\alpha : \alpha \in [-1,1]\} & p_i < c_{e,i} \text{ and } p_i = c_{e,i+1} \\ \{-\frac{1}{n} - \lambda + \lambda\alpha : \alpha \in [-1,1]\} & p_i = c_{e,i} \text{ and } p_i < c_{e,i+1} \\ -\frac{1}{n} - 2\lambda & p_i < c_{e,i} \text{ and } p_i < c_{e,i+1} \end{cases} \tag{2.3}$$

For the zero subgradient condition to hold, we must have that 0 lies between the upper and lower bound of the subdifferential, that is, the top and bottom lines of equation (2.3). This requires $\lambda \geqslant \frac{1}{2n}$, and in this case, $p_{e,i} = \max\{c_{e,i}, c_{e,i+1}\}$ is a solution for p. Indeed, if $\lambda < \frac{1}{2n}$, then $0 \notin \partial\tilde{F}(p)$ for any choice of p because it is strictly less than the smallest possible value of the subdifferential. When $\lambda > \frac{1}{2n}$, $p_{e,i} = \max\{c_{e,i}, c_{e,i+1}\}$ gives a subdifferential that contains 0, and in this case it is the *only* solution. We therefore make the following assumption that $\lambda > \frac{1}{2n}$, where n is number of observations on the geometric network.

Hence, we can reduce the program to

$$\min_{c,k} \sum_{e\in E} \left\{ -\frac{1}{n}\sum_{i=1}^{n_e} \max\{c_{e,i}, c_{e,i+1}\} + \lambda \sum_{i=1}^{n_e} |c_{e,i} - c_{e,i+1}| + \sum_{i=1}^{n_e+1} s_{e,i} e^{c_{e,i}} \right\} + \lambda \sum_{v\in V} \sum_{e\in \text{inc}(v)} |k_v - c_{e,v}| .$$

Because $2 \cdot \max\{c_{e,i}, c_{e,i+1}\} = c_{e,i} + c_{e,i+1} + |c_{e,i} - c_{e,i+1}|$, we have the further equivalence,

$$\min_{c,k} \sum_{e\in E} \left\{ -\frac{1}{2n}\sum_{i=1}^{n_e} (c_{e,i} + c_{e,i+1}) + (\lambda - \frac{1}{2n}) \sum_{i=1}^{n_e} |c_{e,i} - c_{e,i+1}| + \sum_{i=1}^{n_e+1} s_{e,i} e^{c_{e,i}} \right\} + \lambda \sum_{v\in V} \sum_{e\in \text{inc}(v)} |k_v - c_{e,v}| .$$

Again, by [Roc15, Theorem 23.8], a necessary and sufficient condition for $\hat{c}, \hat{k}$ to solve this problem is

$$\begin{aligned} 0 \in \sum_{e\in E} \Bigg\{ & -\frac{1}{2n}\sum_{i=1}^{n_e} \partial\left(c_{e,i} + c_{e,i+1}\right) + (\lambda - \frac{1}{2n}) \sum_{i=1}^{n_e} \partial\left(|c_{e,i} - c_{e,i+1}|\right) + \sum_{i=1}^{n_e+1} \partial\left(s_{e,i} e^{c_{e,i}}\right) \Bigg\} \\ & + \lambda \sum_{v\in V} \sum_{e\in \text{inc}(v)} \partial\left(|k_v - c_{e,v}|\right) . \end{aligned}$$

Here we make an important point. The subdifferential of each $(c_{e,i} + c_{e,i+1})$ term is constant. The subdifferential of $|c_{e,i} - c_{e,i+1}|$ is piecewise constant, and *depends only on the ordering* of the terms $c_{e,i}$ and $c_{e,i+1}$. Similarly, the subdifferential of the $|k_v - c_{e,v}|$ term is piecewise constant and again only depends on the ordering of its terms. Lastly, the subdifferential of $s_{e,i}e^{c_{e,i}}$ is given by its gradient: the (e,i)th coordinate of the subdifferential is $s_{e,i}e^{c_{e,i}}$.

Consider the transformation $\hat{z} = e^{\hat{c}}$, $\hat{h} = e^{\hat{k}}$. This transformation preserves ordering of elements of $\hat{c}$ and $\hat{k}$, so the subdifferential of each absolute value term is invariant under this transformation. We can use this invariance to establish an equivalence of optimality conditions for two different problems, under an exponential transformation. Pursuing this line of reasoning gives the following theorem. In order to facilitate its statement, we briefly establish some notation.

The total variation of a density f on G, which has been parametrized into vectors z and h, can be expressed as a sum of pairwise distances between values in z and h. That is, there are sets J_1 and J_2 of index pairs such that

$$\mathrm{TV}(f) = \sum_{(i,j)\in J_1} |z_i - z_j| + \sum_{(i,j)\in J_2} |z_i - h_j| \,.$$

This formulation depends on the underlying graph structure and the locations of the observations. The right-hand side of this expression can be written as the ℓ_1 norm of a vector $C_1 z + C_2 h$, where C_1 and C_2 are matrices with elements in $\{-1, 0, 1\}$, each having $|J_1| + |J_2|$ rows. We will use the matrices C_1 and C_2, which satisfy $\mathrm{TV}(f) = ||C_1 z + C_2 h||_1$ and C_2 is zero in its first $|J_1|$ rows, in the statement of the following theorem.

THEOREM 2.2.4. *Let $x_1, ..., x_n$ be distinct locations of observations on a geometric network G. Partition these observations into the edges they occur on and the order in which they occur, so that $x_{e,i}$ denotes the ith observation along edge e. Choose λ to satisfy $\lambda > \frac{1}{2n}$.*

- *Let z be a vector with indices enumerating the constant portions of the fused density estimator $\hat{f}$, such that $z_{e,i}$ denotes the value of the fused density estimator on the open interval between $x_{e,i}$ and $x_{e,i-1}$, or between an observation and the end of the edge if $i = 1$ or $n_e + 1$.*
- *Let $s_{e,i}$ be the length of the segment that determines $z_{e,i}$ and $S = \mathrm{diag}(s)$.*
- *Let h be a vector with indices enumerating the nodes in G, such that h_v denotes the value of $\hat{f}$ at node v.*

- *Let C_1 and C_2 be as defined above, and $n_i = |J_i|$. That is, C_1 and C_2 are matrices with $n_1 + n_2$ rows and elements in $\{-1, 0, 1\}$. We have that $\mathrm{TV}(f) = ||C_1 z + C_2 h||_1$, and C_2 is identically zero on its first n_1 rows while having a nonzero index in each of the remaining rows. Let*

$$B = \begin{pmatrix} (\lambda - 1/2n) I_{n_1 \times n_1} & 0_{n_1 \times n_2} \\ 0_{n_2 \times n_1} & \lambda I_{n_2 \times n_2} \end{pmatrix}.$$

- *Let D_1 and D_2 denote the matrices BC_1 and BC_2, respectfully.*
- *Lastly, define the vector w such that*

$$w_{e,i} = \begin{cases} -\frac{1}{2n} & i = 1 \text{ or } i = n_e + 1 \\ -\frac{1}{n} & \text{otherwise.} \end{cases}.$$

Then the fused density estimator $\hat{f}$ for this sample is the minimizer of

$$\min_{z,h} \frac{1}{2} z^\top S z + w^\top z + ||D_1 z + D_2 h||_1 . \tag{2.4}$$

PROOF. The proof follows directly from the line of reasoning before the theorem's statement. Details can be found in Appendix 2.6. □

Remark. With a slight modification of the assumption on λ, Theorem 2.2.4 can be extended to the setting where multiple observations are allowed at a single location. This extension also allows observations to occur at nodes of the geometric network. In practice, this extension may be useful when dealing with imperfect data, though we will not focus on it here because it is a measure zero event in the density estimation paradigm. For completeness, we include the extension in Theorem 2.6.1 of Appendix 2.6.

Methods for computing solutions to the objective in Theorem 2.2.4–a total-variation regularized quadratic program–are well established. As in [**KKBG09**], we rely on solving the dual quadratic program.

PROPOSITION 2.2.5. *The dual problem to (2.4) is*

$$\begin{aligned} \min_y \quad & \frac{1}{2} y^\top D_1 S^{-1} D_1^\top y + w^\top S^{-1} D_1^\top y \\ & ||y||_\infty \leqslant 1 \\ & D_2^\top y = 0. \end{aligned} \tag{2.5}$$

The primal solution $\hat{z}$ can be recovered from the dual $\hat{y}$ through the expression

$$\hat{z} = -S^{-1}(D_1^\top \hat{y} + w).$$

A more general statement of Proposition 2.2.5 which suits the more general statement of Theorem 2.2.4, can be found in Appendix 2.6. It is worth noting that strong duality between the primal and dual problems in (2.4) and (2.5) follows immediately. Indeed, both are extended linear-quadratic programs in the sense of [**RW09**]. By Theorem 11.42 in [**RW09**], strong duality holds, and in addition both the primal and dual problem attain their minimum and maximum values, respectively, if and only if (2.4) is bounded. This is guaranteed by the assumption on λ in Theorem 2.2.4.

2.2.2. Additional Properties of FDEs. In this section, we state a result on the local structure of an FDE and provide additional comments on its implementation details. The result is intuitive: along an edge, the value of piecewise constant segments is inversely related to the length of the segment, relative to adjacent segments. Since smaller segments suggest higher probability in the corresponding region, this property demonstrates local structure of the estimator which aligns with essential global behavior.

PROPOSITION 2.2.6 (Ordering Property). *Let $s_{e,i}$ and $s_{e,i+1}$ be the lengths of two segments interior to an edge e, in the sense that $2 \leqslant i \leqslant n-1$. Assume further that only one observation occurs at $p_{e,j}$ for $j = i-2, \ldots, i+2$. Then $s_{e,i} \leqslant s_{e,i+1}$ implies that $\hat{z}_{e,i} \geqslant \hat{z}_{e,i+1}$. Similarly, $s_{e,i} \geqslant s_{e,i+1}$ implies $\hat{z}_{e,i} \leqslant z_{e,i+1}$.*

PROOF. We will prove that $s_{e,i} \leqslant s_{e,i+1}$ implies $z_{e,i} \geqslant z_{e,i+1}$. The second claim follows symmetrically. Assume, for contradiction, that $s_{e,i} \leqslant s_{e,i+1}$ and $\hat{z}_{e,i} < \hat{z}_{e,i+1}$.

The condition for optimality in (2.4) is

$$0 \in \partial \left(\frac{1}{2} z^\top S z + w^\top z + ||D_1 z + D_2 h||_1 \right) \Big|_{\hat{z},\hat{h}}.$$

The value of the subdifferential in the index corresponding to $z_{e,i}$ is

$$\partial \left(\frac{1}{2} s_{e,i} z_{e,i}^2 + \frac{1}{n} z_{e,i} + \left(\lambda - \frac{1}{2n} \right) \left(|z_{e,i-1} - z_{e,i}| + |z_{e,i} - z_{e,i+1}| \right) \right).$$

Under the assumption that $\hat{z}_{e,i} < \hat{z}_{e,i+1}$, its evaluation at $\hat{z}$ is

$$\hat{z}_{e,i} s_{e,i} + \frac{1}{n} - \left(\lambda - \frac{1}{2n} \right) + (\lambda - \frac{1}{2n}) \partial (|z_{e,i-1} - z_{e,i}|)|_{\hat{z}}.$$

Similarly, the $e, i+1$ index evaluates to

$$\hat{z}_{e,i+1}s_{e,i+1} + \frac{1}{n} + \left(\lambda - \frac{1}{2n}\right) + (\lambda - \frac{1}{2n})\partial(|z_{e,i+2} - z_{e,i+1}|)|_{\hat{z}}.$$

Since $-1 \leqslant \partial|\cdot| \leqslant 1$, we have that

$$\begin{aligned}\hat{z}_{e,i}s_{e,i} + \frac{1}{n} - 2\left(\lambda - \frac{1}{2n}\right) &\leqslant \hat{z}_{e,i}s_{e,i} + \frac{1}{n} - \left(\lambda - \frac{1}{2n}\right) + (\lambda - \frac{1}{2n})\partial(|z_{e,i-1} - z_{e,i}|)|_{\hat{z}} \\ &\leqslant \hat{z}_{e,i}s_{e,i} + \frac{1}{n}\end{aligned}$$

and

$$\begin{aligned}\hat{z}_{e,i+1}s_{e,i+1} + \frac{1}{n} &\leqslant \hat{z}_{e,i+1}s_{e,i+1} + \frac{1}{n} + \left(\lambda - \frac{1}{2n}\right) + (\lambda - \frac{1}{2n})\partial(|z_{e,i} - z_{e,i+1}|)|_{\hat{z}} \\ &\leqslant \hat{z}_{e,i+1}s_{e,i+1} + \frac{1}{n} + 2\left(\lambda - \frac{1}{2n}\right).\end{aligned}$$

Under the assumption that $\hat{z}$ solves this problem, we have that 0 is in the (e,i) index of the subdifferential. This implies

$$\hat{z}_{e,i}s_{e,i} + \frac{1}{n} - 2\left(\lambda - \frac{1}{2n}\right) \leqslant 0 \leqslant \hat{z}_{e,i}s_{e,i} + \frac{1}{n}.$$

But this inequality gives that 0 is not in the $(e, i+1)$ index of the subdifferential, since

$$\hat{z}_{e,i}s_{e,i} + \frac{1}{n} < \hat{z}_{e,i+1}s_{e,i+1} + \frac{1}{n}.$$

This contradicts $\hat{z}$ as solving (2.4), so the result is proven. □

Up to this point in our analysis, we have discussed the computation of the FDE without consideration for preprocessing the data or postprocessing our resulting FDE. Since the computation and rates of convergence of the FDE represent the bulk of our contribution, we will maintain this perspective in the remainder of the chapter. It is worth mentioning, however, that FDE is amenable to pre and postprocessing. Handling multiple observations at a single location in Theorem 2.2.4 makes initial binning or minor discretizations of data (such as projecting observations *onto* a geometric network) straightforward. Moreover, the FDE can be viewed exclusively as a method for generating adaptive bin widths, where the resulting bins can then be fit to the data as in a regular histogram. This approach performs model selection (via FDE) and model fit (via a post-selection MLE) of the histogram separately, and is common practice in model selection using lasso

and related methods [MD17, FST14]. When FDE is used exclusively to find bins, it becomes a change point localization method, instead of a nonparametric density estimator as in its original formulation. Though FDE is amenable to these examples of pre and postprocessing, we will examine the FDE as a density estimator in the remaining sections.

We also make some suggestions into the selection of λ. The choice of λ leads to a fixed number of piecewise constant portions of the fused density estimator. In this sense, the choice of the λ parameter is analogous to choosing the number of bins in histogram estimation. One can tune this selection with information-criteria (IC) such as AIC or BIC by selecting the FDE over a grid of λ values that minimizes the IC. Each of these ICs requires the specification of the degrees of freedom, which can be set to the number of selected piecewise constant regions in the graph, as is done in the Gaussian case [WSST16, TT+12]. Alternatively, one could use cross-validation as a selection criterion. Implementing cross-validation is often practical for large problems because the sparse QP in (2.5) can be very quickly, as we will see in the next section.

2.3. Experiments

We have established a tractable formulation of the fused density estimator in (2.5). Quadratic programming is a mature technology, so computing FDEs via quadratic programming dramatically improves its computation. Quadratic program solvers designed to leverage sparsity in the D_1 and D_2 matrices allow the optimization portion of fused density estimation to scale to large networks and many observations.

In this section we compute FDEs on a number of synthetic and real-world examples.[1] We evaluate the performance of different optimization methods and provide recommendations for solvers which implement those methods. To facilitate accessibility and customization of these tools, each of the solvers we consider is open source and compare favorably with commercial alternatives.

2.3.1. Univariate Examples. We first evaluate fused density estimators in the context of univariate density estimation–where the geometric network G is simply a single edge connecting two nodes. The operator $D_1 + D_2$ is especially simple in this setting, corresponding to an oriented edge-incidence matrix of a chain graph.

[1]These examples can be found at github.com/rbassett3/FDE-Tools, which also includes a Python package for fused density estimation on geometric networks.

Figure 2.3 contains fused density estimators of the standard normal, exponential, and uniform densities, each derived from 100 sample points. The λ parameter in these experiments was selected by 20 fold cross-validation.

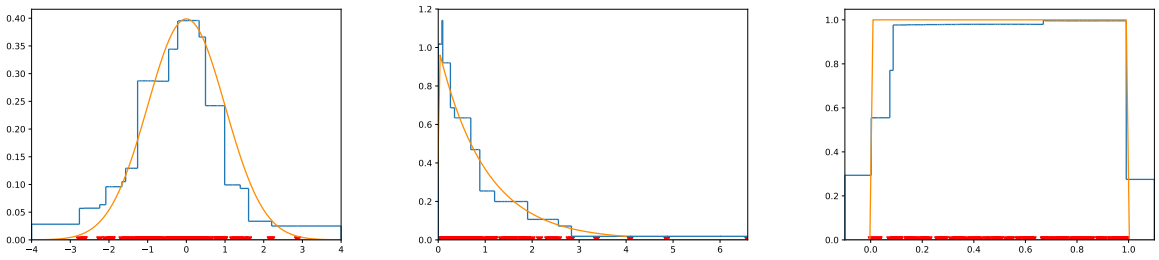


FIGURE 2.3. Univariate densities and fused density estimators

2.3.2. Geometric Network Examples. We next evaluate FDEs on geometric networks. For each of these examples, the underlying geometric network is extracted from OpenStreetMap (OSM) database [**Ope17**].

Figure 2.4 is a fused density estimator with domain taken to be the road network in a region of the city of Baghdad. Observations are the locations of terrorist incidents which occurred in this region from 2013 to 2016, according to the Global Terrorism Database [**LD07**]. The density we attempt to infer is the distribution for the location of terrorist attacks in this region of the city.

Figure 2.5 is an FDE on the road network in Monterey, California. The observations were generated according to a multivariate normal distribution, and projected onto the nearest waypoint in the OpenStreetMap dataset.

Figure 2.6 is the largest example we consider–a fused density estimator run on the entire city of Davis, California. The observations on the network are restaurants and cafes in the town. The corresponding optimization problem has 19000 variables and 25000 constraints (corresponding to the dual formulation in (2.5)). The region with elevated density value is the downtown region of Davis, which was automatically detected by the FDE. Despite its size, the dual quadratic problem was solved in 1.14 seconds in our FDE implementation.

These examples of FDEs on geometric networks illustrate some important properties of the estimator. The FDEs clearly respect the network topology. This is most obviously demonstrated in the Monterey

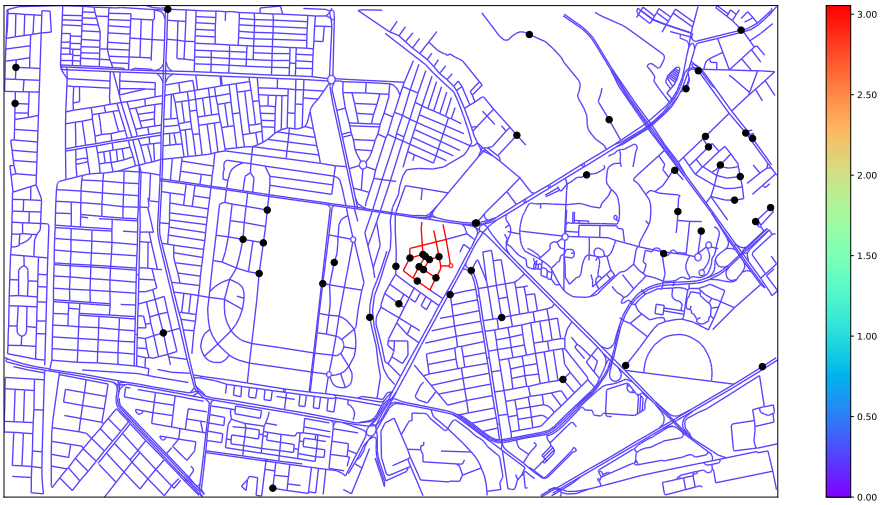


FIGURE 2.4. An FDE for the location of terrorist attacks in Baghdad

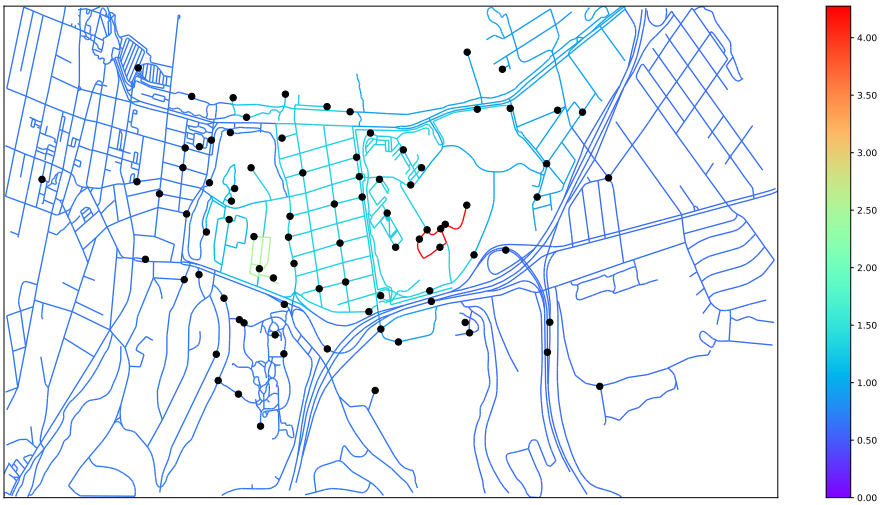


FIGURE 2.5. A fused density estimator for artificial observations on Monterey's road network

example, where the red and light green regions, which correspond to elevated portions of the density, are chosen to be sparsely connected regions of the network. This is intuitive because the sparsely connected regions impact the fusion penalty less severely than a highly connected region, but it is one way that FDEs

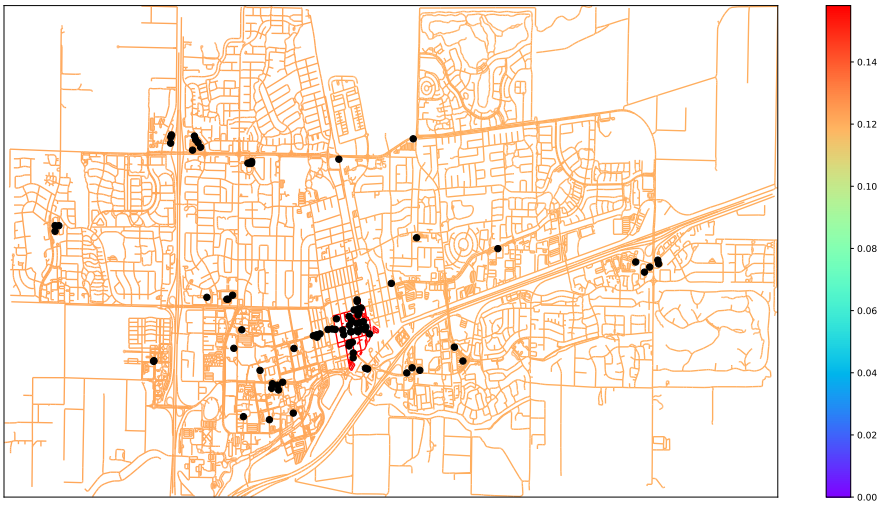


FIGURE 2.6. A fused density estimator on the entire city of Davis, California

reflect the underlying network structure. The Baghdad and Davis examples demonstrate that FDEs can also be used for hot spot localization, and especially in low-data circumstances. Lastly, we note that FDEs partition the geometric network into level sets, thereby forming various regions of the network into clusters. This clustering is an interesting aspect of FDEs, and suggests they could be used to classify regions into areas of high and low priority.

2.3.3. Algorithmic Concerns. The two most prevalent methods for solving sparse quadratic programs are interior point algorithms and the alternating direction method of multipliers. Interior point methods to solve problems of the form (2.4) were introduced by [**KKBG09**]. Interior point approaches have the benefit of requiring few iterations for convergence. The cost per iteration, however, depends crucially on the structure of D_1 and D_2 when performing a Newton step on the relaxed KKT system. In the case of univariate fused density estimators, the Newton step requires inversion of a banded matrix, one which has its nonzero elements concentrated along the diagonal. Leveraging the banded structure allows inversion to be performed in linear time, which is crucial to the performance of the algorithm. For further details of interior point methods, we refer the reader to [**Wri97**, **BV04**, **NN94**].

The alternating direction method of multipliers proceeds by forming an augmented lagrangian function and updating the primal and dual variables sequentially. More details can be found in [**BPC+11**, **Ber14**]. Compared to interior point methods, convergence of ADMM usually requires more iterations of a less-expensive update, whereas interior point methods converge in fewer iterations but require a more expensive update.

In this section, we compare the performance of these algorithms on fused density estimation problems. A comparison between the methods on the related problem of trend-filtering can be found in [**WSST16**], where the algorithmic preferences pertained only to the 2x2 grid graph setting. Their results favor the ADMM approach, though the regularity of this graph structure makes generalizing to general graphs difficult.

For software, we use the Operator Splitting Quadratic Program (OSQP) solver and CVXOPT. These are mature sparse QP solvers that use ADMM and interior point algorithms, respectively. They are both open source, and compare favorably to commercial solvers [**SBG+17**, **Car18**]. Our choice to use these solvers instead of custom implementations reflects that (i) these tools are representative of what is available in practice (ii) outsourcing this portion to other solvers reduces the ability for subtle differences in implementation to favor one method over the other (iii) these projects are production-quality, so their implementations are likely to be of higher quality than custom implementations.

We first compare ADMM and interior point methods on univariate fused density estimator problems. We perform 200 simulations, sampling 100 data points from each distribution. We let λ range from 0.006 to 0.1. These choices correspond to the lower bound on the λ parameter in Theorem 2.2.4 and an upper bound which selects a constant or near-constant density. We report in-solver time, in seconds, and do not include the time required to convert to the sparse formats required for each solver.

TABLE 2.1. Mean and standard deviation of run time (s) for univariate OSQP experiments

	λ		
Density	0.006	0.05	0.1
Exponential	0.0361 ± 0.1310	0.0051 ± 0.0043	0.0045 ± 0.0045
Normal	0.0209 ± 0.0912	0.0112 ± 0.0569	0.0052 ± 0.0046
Uniform	0.0269 ± 0.1077	0.0769 ± 0.0565	0.0074 ± 0.0412

TABLE 2.2. Mean and standard deviation of run time (s) for univariate CVXOPT experiments

	λ		
Density	0.006	0.05	0.1
Exponential	0.0087 ± 0.0012	0.0069 ± 0.0008	0.0078 ± 0.0011
Normal	0.0086 ± 0.0012	0.0071 ± 0.0009	0.0076 ± 0.0010
Uniform	0.0087 ± 0.0010	0.0065 ± 0.0008	0.0061 ± 0.0008

In these experiments, interior point terminated in around 10 iterations. The number of iterations in ADMM were less consistent, ranging from a few hundred to a few thousand.

For the geometric network case, we performed experiments using four examples: the San Diego and Baghdad datasets from figures 2.2 and 2.4, in addition to similar datasets in Davis, California. One of these is a fused density estimator with domain as the road network in downtown Davis, and the other is on the *entire city* of Davis–our largest example in this chapter. We choose λ in a range that progresses from overfitting to underfitting the data. By overfit, we mean that we choose λ as small as possible to make the fused density estimator problem still feasible. By underfit, we mean that the fused density estimator is a constant function. We record '-' when a solver does not run to successful completion. All experiments were run on a computer with 8 GB of memory, an intel processor with four cores at 2.50 GHz, and a 64-bit linux operating system.

TABLE 2.3. OSQP run times (s) for geometric network examples

	λ parameter		
Example	Overfit	Middle	Underfit
Baghdad	0.1086	0.0686	0.0639
San Diego	0.0920	0.0961	0.0628
Downtown Davis	0.0269	0.0769	0.0074
Davis	12.0698	0.8539	0.6052

From these experiments we see that the augmented lagrangian method outperforms interior point on the geometric network examples. The lack of regularity in the matrices D_1 and D_2, and the large-scale matrix factorizations associated with Newton limits this method in comparison to ADMM. On smaller, well-structured problems, like in the univariate examples, interior point methods are often faster. On these well-structured problems, however, the gain in performance is negligible (on the order of a tenth of a second). On the other hand, the speed and versatility of OSQP, especially in the context of large, irregular network

TABLE 2.4. CVXOPT run times (s) for geometric network examples

	λ parameter		
Example	Overfit	Middle	Underfit
Baghdad	1.5493	1.2813	1.1268
San Diego	0.5507	0.4956	0.3256
Downtown Davis	0.0812	0.5615	0.4864
Davis	-	13.4456	13.3911

structure, leads us to recommend ADMM as the method to solve the fused density estimator problem in (2.5). This supports the suggestion of using ADMM for trend-filtering in [**WSST16**], and extends their recommendation beyond the 2×2 grid graph.

2.4. Statistical Rates

In this section we prove a squared Hellinger rate of convergence for fused density estimation when the true log-density is of bounded variation. Hellinger distance is defined as

$$h^2(f, f_0) = \frac{1}{2}\int_G (\sqrt{f} - \sqrt{f_0})^2 \, dx,$$

where dx is the base measure over the edges in the geometric network G; in the univariate setting, this is just the Lebesgue measure. The factor of $\frac{1}{2}$ is a convention that ensures that the Hellinger distance is bounded above by 1. The Hellinger distance is a natural choice for quantifying rates of convergence for density estimators because it is tractable for product measures and provides bounds for rates in other metrics [**LeC73**, **LC12**, **GS02**]. The squared Hellinger risk of an estimator $\tilde{f}$ for f_0 is $\mathbb{E}[h^2(\tilde{f}, f_0)]$. The minimax squared Hellinger risk over a set of densities $\mathscr{H}$ is

$$\min_{\tilde{f}} \max_{f \in \mathscr{H}} \mathbb{E}_f[h^2(\tilde{f}, f)].$$

We find fused density estimation achieves a rate of convergence in squared Hellinger risk which matches the minimax rate over all univariate densities in $\mathscr{F}$–densities of log-bounded variation where the underlying geometric network is simply a compact interval. In this sense, univariate FDE has the best possible squared Hellinger rate of convergence over this function class. The rate we attain is $n^{-2/3}$, and the equivalence of rates is asymptotic. On an arbitrary connected geometric network, minimax rates for density estimation can

depend on the network, but our results demonstrate that FDE on a geometric network has squared Hellinger rate at most the univariate minimax rate.

We begin by establishing the minimax rate over the class $\mathscr{F}$, which gives a lower bound on the squared Hellinger rate for fused density estimation. To establish the lower bound, it is sufficient to examine the minimax rate of convergence over a set of densities contained in $\mathscr{F}$. Fixing a constant C and compact interval I, we consider the set of functions $g : I \to \mathbb{R}$

$$\mathrm{BV}(C) := \{g : \mathrm{TV}(g) \leqslant C, ||g||_\infty < C\}.$$

Because $\mathrm{BV}(C)$ is bounded below, the packing entropy of $\mathrm{BV}(C)$ and $\widetilde{\mathrm{BV}}(C) := \{\exp(g) : g \in \mathrm{BV}(C)\}$ are of the same order. From Example 6.4 in [YB99], we have that $\mathrm{BV}(C)$ has L_2 packing entropy of order $\frac{1}{\varepsilon}$. Applying Theorem 5 from [YB99] gives the minimax squared Hellinger rate over densities $\{\frac{f}{\int f} : f \in \widetilde{\mathrm{BV}}(C)\}$ as $n^{-2/3}$. In Theorem 2.4.2, we show that the FDE attains the rate of $n^{-2/3}$ over the larger class $\mathscr{F}$. Therefore, the minimax squared Hellinger rate over $\mathscr{F}$ must also equal $n^{-2/3}$, so we have proven the following theorem. For sequences a_n and b_n, we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$.

THEOREM 2.4.1. *The minimax squared Hellinger rate over $\mathscr{F}$, the set of densities f with $\log f$ of bounded variation, is $n^{-2/3}$. That is,*

$$\min_{\tilde{f}} \max_{f \in \mathscr{F}} \mathbb{E}_f[h^2(\tilde{f}, f)] \asymp n^{-2/3}.$$

To prove the FDE rate of convergence for univariate density estimation, we extend techniques developed for the theory of M-estimators, [Gee00], and locally-adaptive regression splines in Gaussian models, [MvdG+97]. A detailed proof of our main result can be found in Appendix 2.6. This rate bound for FDE is based on novel empirical process bounds for log-densities of bounded variation, and these are used in conjunction with peeling arguments to provide a uniform bound on the Hellinger error. The empirical process bounds in 2.6.3 rely on new Bernstein difference metric covering number bounds for functions of bounded variation, which can be found in Appendix 2.7. We extend the FDE rates for the univariate setting to arbitrary geometric networks in section 2.4.2; this requires embedding the geometric network onto the real line. This embedding is constructed from the depth-first search algorithm, a technique used in [PSST16] for regression over graphs, and is described in Appendix 2.6.

The subsections in this section follow this outline: In subsection 2.4.1, we provide a proof sketch of the squared Hellinger rate of convergence for the univariate FDE. In subsection 2.6.3, we detail the lemmas used to prove the main result. In subsection 2.4.2, we extend these rate results from the univariate setting to arbitrary geometric networks.

2.4.1. Upper Bounds for Rate of Univariate FDE. In this subsection we prove a squared Hellinger rate of $n^{-2/3}$ for univariate fused density estimation. Let the geometric network G be a closed interval $[a,b]$ (a single edge connecting nodes a and b). Recall the definition of $\mathscr{F}$ as the set of densities f with $\log f$ of bounded variation. Let $f_0 \in \mathscr{F}$ be a fixed density on G, so that the total variation $\mathrm{TV}(\log f_0)$ is constant as n increases.

THEOREM 2.4.2. *Let $\hat{f}_n$ be the fused density estimator of an iid sample of n points drawn from a univariate density f_0. There is a choice of λ_n such that $\lambda_n = O_P(n^{-2/3})$, the FDE is well defined, and*

$$\mathbb{E}_{f_0}[h^2(\hat{f}_n, f_0)] = O(n^{-2/3}).$$

Combined with the lower bound in Theorem 2.4.1, this gives that univariate fused density estimation attains the minimax rate over densities in $\mathscr{F}$.

Proof Sketch (Detailed proof in Appendix 2.6).

In order to control the Hellinger error for FDE, we rely on the fact that the FDE is the minimizer of (2.1). We derive an inequality involving the squared Hellinger distance, an empirical process, and fusion-penalty terms. This inequality (and in general inequalities serving this purpose; see [**Gee00**]) is referred to as a basic inequality. To reduce notation, we introduce the shorthand $\hat{h} = h(\hat{f}_n, f_0)$, $I(f) = \mathrm{TV}(\log f)$, $\hat{I} = I(\hat{f}_n)$, $I_0 = I(f_0)$, and $p_f = \frac{1}{2}\log\frac{f+f_0}{2f_0}$.

We arrive at the following basic inequality by manipulating the optimality condition, $-P_n(\log \hat{f}_n) + \lambda_n \hat{I} \leqslant -P_n(\log f_0) + \lambda_n I_0$. In fact, from the definition of the FDE we have the stronger condition $-P_n(\log \hat{f}_n) + \lambda_n \hat{I} \leqslant -P_n(\log f) + \lambda_n I(f)$ for all $f \in \mathscr{F}$, but the weaker condition will suffice.

LEMMA 2.4.3 (Basic Inequality).

$$\hat{h}^2 \leqslant 16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_0 - \hat{I}).$$

Squared Hellinger rates now follow from controlling the right hand side. We do so by considering two cases. When $\hat{h}$ is small, we show that

$$(P_n - P)(p_{\hat{f}_n}) = O_P\left(n^{-2/3}(1 + I_0 + \hat{I})\right).$$

From the basic inequality, this gives

$$\begin{aligned}\hat{h}^2 &= O_P\left(16n^{-2/3}(1 + I_0 + \hat{I}) + 4\lambda_n(I_0 - \hat{I})\right) \\ &= O_P\left(4(4n^{-2/3} - \lambda_n)\hat{I} + 4(4n^{-2/3} + \lambda_n)I_0 + 16n^{-2/3}\right). \qquad (2.6)\end{aligned}$$

Excluding details, when λ_n is chosen to dominate $4n^{-2/3}$, the first term in (2.6) is negative, so we conclude that $\hat{h}^2 = O_P\left(\max\{n^{-2/3}, \lambda_n\}\right)$.

The condition "when $\hat{h}$ is small", and the corresponding control on $(P_n - P)(p_{\hat{f}_n})$ can be formalized in the following theorem.

THEOREM 2.4.4.

$$\sup_{h(f,f_0) \leqslant n^{-1/3}(1 + I(f) + I_0)} \frac{n^{2/3}\,|(P_n - P)(p_f)|}{1 + I(f) + I_0} = O_P(1),$$

where the supremum is taken over all $f \in \mathscr{F}$.

When $\hat{h}$ is large, on the other hand, we show that $(P_n - P)(p_{\hat{f}_n}) = O_P\left(n^{-1/2} \cdot \hat{h}^{1/2} \cdot (1 + \hat{I} + I_0)^{1/2}\right)$. From the basic inequality, this gives

$$\sqrt{n}\hat{h}^2 = O_P\left(16\hat{h}^{1/2}(1 + I_0 + \hat{I})^{1/2} + 4\sqrt{n}\lambda_n(I_0 - \hat{I})\right).$$

Whence we conclude that $\hat{h}^2 = O_P\left(\max\{n^{-2/3}, \lambda_n\}\right)$. This follows from the analogue to (2.4.4) when $\hat{h}$ is large.

THEOREM 2.4.5.

$$\sup_{h > n^{-1/3}(1 + I(f) + I_0)} \frac{n^{1/2}\,|(P_n - P)(p_f)|}{h^{1/2}(f, f_0)(1 + I(f) + I_0)^{1/2}} = O_P(1),$$

where the supremum is taken over all $f \in \mathscr{F}$.

To summarize our conclusions so far: the squared Hellinger rate is $\max\{n^{2/3}, \lambda_n\}$ when λ_n balances the competing terms in (2.6). By choosing

$$\lambda_n = \max\left\{ \sup_{h(f,f_0)\leqslant n^{-1/3}(1+I(f)+I_0)} \frac{4\,|(P_n - P)(p_f)|}{1+I(f)+I_0}, n^{-2/3} \right\},$$

we have a minimal λ_n which dominates in (2.6). Furthermore, this choice of λ_n satisfies $\lambda_n = O_P(n^{-2/3})$ by Theorem 2.4.4. We have established a squared Hellinger rate of $n^{-2/3}$ for both the cases of $\hat{h}$ considered. Furthermore, this choice of λ_n satisfies the condition on λ in Theorem 2.2.4, so the FDE is well-defined.

Theorems 2.4.4 and 2.4.5 are essential components of the proof outlined above. Both of these results are new and of independent interest. Their derivation requires the following lemma.

LEMMA 2.4.6. *Let $M \in \mathbb{R}$ and $\mathscr{P}_M = \{p_f : 1 + I(f) + I_0 \leqslant M\}$. There is a constant C and choice of c_1 such that for all $C_1 \geqslant c_1$ and $\delta \geqslant \frac{M}{2} \cdot n^{-1/3}$*

$$\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, h(f,f_0)\leqslant\delta} \left|\sqrt{n}(P_n - P)(p_f)\right| \geqslant 2C_1\sqrt{M}\delta^{1/2} \right) \leqslant C\exp\left[-\frac{C_1 M \delta^{-1}}{4C^2} \right]$$

Lemma 2.4.6 can be used to prove Theorems 2.4.4 and 2.4.5 by applying the peeling device twice, once each for the parameters M and δ.

The proof of Lemma 2.4.6 requires three basic ingredients: control of the bracketing entropy of $\mathscr{P}_M$, a uniform bound on $\mathscr{P}_M$, and a relationship dictating how M scales with control of the Hellinger distance. These ingredients have the same motivation as in [**MvdG+97**], where the authors use a total variation penalty to construct adaptive estimators in the context of regression. In that work, the authors assume subgaussian errors and prove bounds on metric entropy for functions of bounded variation. The subgaussian assumption provides local error bounds, and the metric entropy condition bounds the number of sets on which we must control that error. Though similarly motivated, our context is more complicated. In order to control the M in $\mathscr{P}_M$ with the Hellinger distance, we consider coverings in the Bernstein difference metric instead of the $L_2(P)$ metric. Using the Bernstein difference allows us to achieve the results in Lemma 2.4.6, but its use requires control of generalized bracketing entropy–bracketing with the Bernstein difference–instead of the usual bracketing entropy with the $L_2(P)$ metric. In addition, the uniform bound we require is now on the Bernstein difference over $\mathscr{P}_M$.

In Appendix 2.7, we show that the bracketing entropy of $\mathscr{P}_M$, with bracketing radius δ, is of order $\frac{M}{\delta}$. This bracketing entropy results implies generalized bracketing entropy bounds, and can be proved similarly to results in monotonic shape-constrained estimation [**VDVW96**]. In order to achieve the finite sample bounds necessary to achieve these rates, Bernstein's inequality is used to provide concentration inequalities that are critical to bounding the basic inequality. With this combination of local error bounds and bracketing rates, we can apply results in the spirit of generic chaining [**Tal06**] to obtain Lemma 2.4.6.

Lastly, we translate the probabilistic results into bounds on Hellinger risk. In general, one cannot prove expected risk rates from convergence in probability because the tails may not decay quickly enough to give a finite expectation. But out of the proofs of Theorems 2.4.4 and 2.4.5, we can derive exponential tail bounds for $h^2(\hat{f}_n, f_0)$. This allows us to translate our probabilistic rates into rates on the Hellinger risk; doing so requires some care to simultaneously apply the rates in Theorems 2.4.4 and 2.4.5. These details are provided in the expanded proof in Appendix 2.6.

2.4.2. Guarantees for Connected Geometric Networks. In the previous sections we proved an $n^{-2/3}$ rate of convergence for univariate fused density estimators. In this section we extend that result to arbitrary connected geometric networks. Recall that the total-variation on a geometric network G is the sum of the total variation over the edges. In this section only, we denote the graph-induced total variation by TV_G and univariate total variation TV, because both will be used in similar contexts. Let $I_G(f) = \mathrm{TV}_G(\log f)$, $\hat{I}_G = I_G(\hat{f}_n)$, and $I_{0,G} = \mathrm{TV}_G(\log f_0)$.

The previous section derived rate of convergence results for the case of univariate data. In this section we define a depth-first embedding of a geometric network onto the real line. The idea is simple. We perform depth-first search on the geometric network, G, and lay out the edges of the network on the real line as we visit them. This map that takes the geometric network to the real line preserves the base measure on G, so inferring a density on G is equivalent to inferring the analogous one on $\mathbb{R}$. Equally important is that the univariate total variation of a function which has been mapped onto $\mathbb{R}$ can be bounded above by the graph-induced total variation of the function. We use this fact to show that squared Hellinger rate for FDE on graphs convergence is bounded above by the univariate squared Hellinger rate.

This technique is inspired by [**PSST16**], where the authors apply a similar technique to prove a rate of convergence for the fused lasso estimator on graphs. In that article, the rate of convergence and the total

variation bound immediately imply the same rate of convergence in the general case. The situation is more subtle in this setting, because the peeling arguments we use complicate the situation.

In Lemma 2.6.11, we show that for any fixed geometric network G, there is a measure-preserving embedding γ of G into $\mathbb{R}$ that preserves densities and Hellinger distance. Furthermore, for any function g on G, the (univariate) total-variation of the embedded function $g \circ \gamma^{-1}$ never exceeds twice that of the graph-values total variation. For notational convenience, we will refer to the univariate total variation $\mathrm{TV}(g \circ \gamma^{-1})$ as $\mathrm{TV}(g)$. This embedding allows us to extend the univariate squared Hellinger rates to geometric networks.

THEOREM 2.4.7. *Let $\hat{f}_n$ be the FDE of an iid sample over a connected geometric network with true density f_0. Then there exists a choice of λ_n such that $\lambda_n = O_P(n^{-2/3})$ and*

$$\mathbb{E}_{f_0}\left[h^2(\hat{f}_n, f_0)^2\right] = O(n^{-2/3}).$$

PROOF. From the previous section, we have that

$$\sup_{h(f,f_0)>n^{-1/3}(1+I(f)+I0)} \frac{n^{1/2}\,|(P_n - P)(p_f)|}{h^{1/2}(f,f_0)(1+I(f)+I_0)^{1/2}} = O_P(1) \tag{2.7}$$

and

$$\sup_{h(f,f_0)\leqslant n^{-1/3}(1+I(f)+I_0)} \frac{n^{2/3}\,|(P_n - P)(p_f)|}{1+I(f)+I_0} = O_P(1). \tag{2.8}$$

We proceed analogously to Theorem 2.4.2. Take

$$\lambda_n = \max\left\{ \sup_{h(f,f_0)\leqslant n^{-1/3}(1+I(f)+I_0)} \frac{8(P_n - P)(p_f)}{1+I(f)+I_0}, n^{-2/3} \right\}.$$

From the Basic Inequality, Lemma 2.4.3, we have

$$\hat{h}^2 \leqslant \max\left\{ \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)} \left(16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_{0,G} - \hat{I}_G)\right) \right. \tag{2.9}$$

$$\left. \mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)} \left(16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_{0,G} - \hat{I}_G)\right) \right\} \tag{2.10}$$

First, consider the case $\hat{h} > n^{-1/3}(1+\hat{I}+I_0)$. Define subsets of the probability space

$$B_L = \left\{ \sqrt{n} \left| (P_n - P)(p_{\hat{f}}) \right| > L \cdot \hat{h}^{1/2} \cdot (1+\hat{I}+I_0)^{1/2} \right\}$$

and

$$C_M = \{\hat{I} > M \geqslant 2I_{0,G}\}.$$

Because $\hat{I} \leqslant 2\hat{I}_G$ (Lemma 2.6.11), on C_M we have $I_{0,G} - \hat{I}_G < 0$ on C_M. Proceeding as in the proof of Theorem 2.4.2, the fact that $\lambda_n n^{-2/3}$ is bounded below by 1 gives that on B_L^c

$$\sqrt{n}h^{3/2} \leqslant 16 \cdot L \cdot (1 + \hat{I} + I_0)^{1/2} + \frac{4(I_{0,G} - \hat{I}_G)}{\left(16 \cdot L \cdot (1 + \hat{I} + I_0)^{1/2}\right)^{1/3}}.$$

As M gets large, this inequality and the fact that $2\hat{I}_G$ dominates $\hat{I}$ gives that $B_L^c \cap C_M = \varnothing$. So on B_L^c, for large enough M,

$$\begin{aligned}\sqrt{n}\hat{h}^2 &\leqslant 16 \cdot L \cdot \hat{h}^{1/2}(1 + I_0 + \hat{I})^{1/2} + 4\lambda_n\sqrt{n}(I_{0,G} - \hat{I}_G) \\ &\leqslant 2\max\left\{16 \cdot L \cdot \hat{h}^{1/2}(1 + I_0 + M)^{1/2}, 2\lambda_n\sqrt{n}I_{0,G}\right\}.\end{aligned}$$

This holds with probability $1 - \varepsilon$ if we choose L so that B_L holds with probability less than ε–the fact that we can do so is guaranteed by (2.7). We conclude that when $\hat{h} > n^{-1/3}(1 + \hat{I} + I_0)$, $\hat{h}^2 = O_P\left(\max\{\lambda_n, n^{-2/3}\}\right)$.

Next consider the case $\hat{h} \leqslant n^{-1/3}(1 + \hat{I} + I_0)$. Mirroring equations (2.24)-(2.27), we have

$$\begin{aligned}\hat{h}^2 &\leqslant 4(1 + I_{0,G} + \hat{I}_G)\left(\sup_{h(f,f_0)\leqslant n^{-1/3}(1+I+I_0)} \frac{4(P_n - P)(p_f)}{1 + I_{0,G} + I_G} - \lambda_n\right) + 4\lambda_n(1 + 2I_{0,G}) \\ &\leqslant 4(1 + I_{0,G} + \hat{I}_G)\left(\sup_{h(f,f_0)\leqslant n^{-1/3}(1+I+I_0)} \frac{8(P_n - P)(p_f)}{1 + I_0 + I} - \lambda_n\right) + 4\lambda_n(1 + 2I_{0,G})\end{aligned}$$

The last inequality again comes from $1 + I_0 + I \leqslant 2(1 + I_{0,G} + I_G)$. By our choice of λ_n we have that

$$\hat{h}^2 \leqslant 4\lambda_n(1 + 2I_{0,G}).$$

Because $\lambda_n = O_P(n^{-2/3})$, we have that when $\hat{h} \leqslant n^{-1/3}(1 + \hat{I} + I_0)$, $\hat{h}^2 = O_P\left(n^{-2/3}\right)$. Having established probabilistic the rate for both cases of $\hat{h}$, we must now translate these into rates for the squared Hellinger risk. In the unvariate case, we use the probabilistic bounds just derived–for the cases $\hat{h} \leqslant n^{-1/3}(1 + \hat{I} + I_0)$ and $\hat{h} > n^{-1/3}(1 + \hat{I} + I_0)$–to prove an equivalent rate in Hellinger risk. This part of the proof follows exactly as in the analogous result for the univariate case, and as such is omitted. □

This result demonstrates that univariate density estimation is the hardest type of connected geometric network for estimating log-densities of bounded variation. This is consistent with the results from [PSST16] in the regression setting. While there may be networks for which the squared Hellinger error may decrease more quickly than the $n^{-2/3}$, we reserve that study to future work.

2.5. Discussion

In this work, we introduced the fused density estimator, a nonparametric density estimator derived from total variation penalized maximum likelihood. The result is a piecewise constant density function, similar to a histogram, with bin widths that adapt to the local smoothness of the underlying density. The univariate problem formulation has a straightforward extension to geometric networks, which leads fused density estimators to have many potential applications in infrastructure networks. We have shown that the computation of fused density estimators can be reduced to a sparse QP, which makes fused density estimators tractable on large scale problems.

Our theoretical analysis provides a foundation for a more substantial understanding of the FDE. In particular, we show that univariate fused density estimation achieves the minimax squared Hellinger rate for densities of bounded variation. This serves as important validation for the method of fused density estimation. We reserve for future work the study of specific network structures and fused density estimation over higher dimensional manifolds.

Acknowledgements. Robert Bassett was supported in part by the U.S. Office of Naval Research grant N00014-17-2372. James Sharpnack acknowledges partial support from NSF grant DMS-1712996. Both authors would like to thank Ryan Tibshirani for helpful conversations, as well as Roger Wets and Matthias Köppe for their useful comments.

2.6. Appendix: Proofs

2.6.1. Proofs from Section 2.2.

Proof of Lemma 2.2.2.

Let $\hat{g}$ be any function of bounded variation. Set $\bar{g} = \hat{g} - \ln(\int_G e^{\hat{g}}\,dx$, so that $\int e^{\bar{g}}\,dx = 1$. The value of

$$-P_n(g) + \lambda\,\mathrm{TV}(g) + \int_G e^g\,dx \tag{2.11}$$

evaluated at $\hat{g}$ is

$$-P_n(\hat{g}) + \lambda \, \mathrm{TV}(\hat{g}) + \int_G e^{\hat{g}} \, dx. \tag{2.12}$$

Whereas (2.11) evaluated at $\bar{g}$ is

$$-P_n(\hat{g}) + \ln\Big(\int_G e^{\hat{g}}\Big) \, dx + \mathrm{TV}(\hat{g}) + 1. \tag{2.13}$$

Here we have used that $\mathrm{TV}(\hat{g}) = \mathrm{TV}(\bar{g})$, which follows from shift-invariance of total variation. That is, $\mathrm{TV}(g) = \mathrm{TV}(g+a)$ for any function g and constant a. Subtracting (2.13) from (2.12) gives

$$\int_G e^{\hat{g}} \, dx - \ln\Big(\int_G e^{\hat{g}} \, dx\Big) - 1.$$

Recall that $x - \ln(x) \geqslant 1$ for all $x > 0$, with equality attained if and only if $x = 1$. This proves our result, since we have shown that $\hat{g}$ cannot be a minimizer of (2.11) unless $\hat{g} = \bar{g}$, in which case $\int_G e^{\hat{g}} \, dx = 1$. □

It is worth noting that, in the proof of Lemma 2.2.2, we only require that total variation is shift invariant. A similar result holds with any shift invariant penalty substituted for TV.

Proof of Representer Theorem, Lemma 2.2.3.

Let $e \in E$. Consider any subinterval (a,b) of e which does not intersect $\{x_1, \ldots, x_n\} \cup V$. Assume, towards a contradiction, that $\hat{g}$ is a function which is not constant on (a,b). We will show that $\hat{g}$ cannot minimize (2.2).

Define

$$\bar{g} = \begin{cases} \min\{\hat{g}(a), \hat{g}(b)\} & \text{for } x \in (a,b) \\ \hat{g}(x) & \text{otherwise.} \end{cases}$$

We next consider the effect of this change on the objective in (2.2). Since no $x_i \in (a,b)$, the P_n term is unaltered by changing $\hat{g}$ to $\bar{g}$. Let $e_{[a,b]}$ be the subset of e corresponding to $[a,b]$ and $e_{\overline{[a,b]^{\complement}}}$ the closure of its complement within e. The interval (a,b) is contained in e, so we have

$$\mathrm{TV}(g|_e) = \mathrm{TV}(g|_{e_{[a,b]}}) + \mathrm{TV}(g|_{e_{\overline{[a,b]^{\complement}}}})$$

for every real-valued function g on G. From the definitions of $\hat{g}$ and $\bar{g}$,

$$\mathrm{TV}(\bar{g}|_{[a,b]}) = |\hat{g}(a) - \hat{g}(b)| \leqslant \mathrm{TV}(\hat{g}|_{e_{[a,b]}}).$$

This equality is attained if and only if $\hat{g}$ is monotonic on $[a,b]$. We have established that $\mathrm{TV}(\bar{g}) \leqslant \mathrm{TV}(\hat{g})$.

The integral term in (2.2) is less at $\bar{g}$ than its evaluation at $\tilde{g}$, since $\bar{g} \leqslant \tilde{g}$. Equality holds when $\{x : \tilde{g}(x) \neq \bar{g}(x)\}$ has measure zero on (a,b). Hence,

$$-P_n(\hat{g}) + \lambda\,\mathrm{TV}(\hat{g}) + \int_G e^{\hat{g}}\,dx \leqslant -P_n(\bar{g}) + \lambda\,\mathrm{TV}(\bar{g}) + \int_G e^{\bar{g}}\,dx.$$

We cannot have that $\hat{g}|_{[a,b]}$ is monotonic and satisfies $\{x \in (a,b) : \tilde{g}(x) \neq \bar{g}(x)\}$ has measure zero, unless $\hat{g}$ is constant on (a,b). By assumption it is not, so we conclude that $\hat{g}$ cannot satisfy (2.2) since its evaluation at the objective is strictly greater than at $\bar{g}$. Therefore, any $\hat{g}$ that satisfies (2.2) must be constant on (a,b). □

Proof of Theorem 2.2.4.

We pick up from the discussion preceding the theorem's statement. Recall that the subdifferentials of the fusion penalty terms are preserved by the exponential transformation. This allows to conclude that there is a $\hat{c},\hat{k}$ satisfying

$$0 \in \partial \left(\sum_{e\in E} \left\{ -\frac{1}{2n}\sum_{i=1}^{n_e}(c_{e,i}+c_{e,i+1}) + \left(\lambda - \frac{1}{2n}\right)\sum_{i=1}^{n_e}|c_{e,i}-c_{e,i+1}| + \sum_{i=1}^{n_e+1} s_{e,i}e^{c_{e,i}} \right\} + \lambda \sum_{v\in V}\sum_{e\in \mathrm{inc}(v)} |k_v - c_{e,v}| \right)_{(\hat{c},\hat{k})},$$

if and only if $\hat{z} = e^{\hat{c}}$ and $\hat{h} = e^{\hat{k}}$ satisfies

$$0 \in \partial \left(\sum_{e\in E} \left\{ -\frac{1}{2n}\sum_{i=1}^{n_e}(z_{e,i}+z_{e,i+1}) + \left(\lambda - \frac{1}{2n}\right)\sum_{i=1}^{n_e}|z_{e,i}-z_{e,i+1}| + \sum_{i=1}^{n_e+1} \frac{s_{e,i}}{2}\cdot z_{e,i}^2 \right\} + \lambda \sum_{v\in V}\sum_{e\in \mathrm{inc}(v)} |h_v - z_{e,v}| \right)_{(\hat{z},\hat{h})}. \tag{2.14}$$

The above subdifferentials are taken with respect to (c,k) and (z,h), respectfully. The problem that generates the optimality condition (2.14) is

$$\min_{z,h} \sum_{e\in E} \left\{ -\frac{1}{2n}\sum_{i=1}^{n_e}(z_{e,i}+z_{e,i+1}) + \left(\lambda - \frac{1}{2n}\right)\sum_{i=1}^{n_e}|z_{e,i}-z_{e,i+1}| + \sum_{i=1}^{n_e+1} \frac{s_{e,i}}{2}\cdot z_{e,i}^2 \right\} + \lambda \sum_{v\in V}\sum_{e\in \mathrm{inc}(v)} |h_v - z_{e,v}|.$$

By solving this new problem, and then applying a log-transformation, we solve the original FDE problem. Recall that the original formulation of the problem (2.1) was formulated in terms of the log-density g. Hence a solution to (2.14) gives the values of density instead of the log-density.

By construction w satisfies

$$w_{e,i} = \begin{cases} -\frac{1}{2n} & i = 1 \text{ or } i = n_e \\ -\frac{1}{n} & \text{otherwise.} \end{cases}$$

By construction, we also have that

$$||D_1 z + D_2 h||_1 = \left(\lambda - \frac{1}{2n}\right) \sum_{e \in E} \sum_{i=1}^{n_e} |z_{e,i} - z_{e,i+1}| + \lambda \sum_{v \in V} \sum_{e \in \text{inc}(v)} |h_v - z_{e,v}|.$$

Letting $S = \text{diag}(s)$, we conclude that we can solve the fused density estimator problem by solving

$$\min_{z,h} \frac{1}{2} z^\top S z + w^\top + ||D_1 z + D_2 h||_1$$

because we have translated the optimality conditions in (2.14) to the problem above. The FDE $\hat{f}$ is found by taking the piecewise constant portion of $\hat{f}_{e,i}$ to be $\hat{z}_{e,i}$ and the value of $\hat{f}$ at node v to be $\hat{h}_v$. □

THEOREM 2.6.1 (Extension of Theorem 2.2.4). *Let $x_1, ..., x_n$ be the distinct locations of observations on a geometric network G. Partition these locations into the edges they occur on and the order in which they occur, so that $x_{e,i}$ denotes the ith observation along edge e. Let $q_{e,i}$ denote the number of observations which occur at location $x_{e,i}$, and assume the penalty parameter λ satisfies $\lambda > \max_{e,i} \left\{ \frac{q_{e,i}}{n \cdot \deg(x_{e,i})} \right\}$.*

- *Let z be a vector with indices enumerating the constant portions of the fused density estimator $\hat{f}$, such that $z_{e,i}$ denotes the value of the fused density estimator on the open interval between $x_{e,i}$ and $x_{e,i-1}$, or between an observation and the end of the edge if $i = 1$ or $n_e + 1$.*
- *Let $s_{e,i}$ be the length of the segment that determines $z_{e,i}$ and $S = \text{diag}(s)$.*
- *Let h be a vector with indices enumerating the nodes in G, such that h_v denotes the value of the fused density estimator at node v.*
- *Using the convention that $q_{e,0} = 0$ and $q_{e,n_{e+1}} = 0$. Define $\bar{q}$ such that $\bar{q}_{e,i} = \frac{q_{e,i} + q_{e,i-1}}{2}$, for each $e \in E$ and $i \in \{1, ..., n_e + 1\}$.*
- *Let r be a vector whose indices enumerate the vertices of G, such that r_v denotes the number of observations that occur at node v.*

- *Let C_1 and C_2 be as in Theorem 2.2.4. That is, C_1 and C_2 are matrices with $n_1 + n_2$ rows and elements in $\{-1, 0, 1\}$. We have that $\mathrm{TV}(f) = ||C_1 z + C_2 h||_1$, and C_2 is identically zero on its first n_1 rows while having a nonzero element in each of the remaining rows. Let*

$$B = \begin{pmatrix} \mathrm{diag}(\lambda - q/2n) & 0_{n_1 \times n_2} \\ 0_{n_2 \times n_1} & \lambda I_{n_2 \times n_2} \end{pmatrix}.$$

- *Let D_1 and D_2 denote the matrices BC_1 and BC_2, respectfully.*
- *Lastly, let $u = -r/n$ and $w = -\bar{q}/n$.*

Then one can compute the fused density estimator $\hat{f}$ for this sample by solving

$$\min_{z,h} \frac{1}{2} z^\top S z + w^\top z + u^\top h + ||D_1 z + D_2 h||_1 . \tag{2.15}$$

PROOF. The proof of this theorem follows exactly as in the proof of Theorem 2.2.4, with slightly more cumbersome notation. □

The following is a more general statement of Proposition 2.2.5, and provides the dual of the more general primal problem, (2.15).

PROPOSITION 2.6.2 (Extension of Proposition 2.2.5). *The dual problem to (2.15) is*

$$\begin{aligned} \min_{y} \quad & \frac{1}{2} y^\top D_1 S^{-1} D_1^\top y + w^\top S^{-1} D_1^\top y \\ & ||y||_\infty \leqslant 1 \\ & D_2^\top y = -u. \end{aligned} \tag{2.16}$$

The primal solution $\hat{z}$ can be recovered from the dual $\hat{y}$ through the expression

$$\hat{z} = -S^{-1}(D_1^\top \hat{y} + w).$$

PROOF. Write (2.4) as

$$\begin{aligned} \min_{z,h,l} \quad & \frac{1}{2} z^\top S z + w^\top z + u^\top h + ||l||_1 \\ \text{s.t.} \quad & l = D_1 z + D_2 h \end{aligned}$$

Introducing the dual variable y, this problem has Lagrangian

$$\frac{1}{2}z^\top Sz + w^\top z + u^\top h + ||l||_1 + y^\top(D_1 z + D_2 h - l). \tag{2.17}$$

To find the dual problem, we minimize in the primal variables. This gives

$$\min_l -y^\top l + ||l||_1 = \begin{cases} 0 & \text{if } ||y||_\infty \leqslant 1 \\ -\infty & \text{otherwise.} \end{cases}. \tag{2.18}$$

In addition, we have the terms

$$\min_z \ \frac{1}{2}z^\top Sz + w^\top z + y^\top D_1 z \tag{2.19}$$

and

$$\min_h \ u^\top h + y^\top D_2 h. \tag{2.20}$$

For (2.19), we have the optimality condition

$$Sz + w + D_1^\top y = 0. \tag{2.21}$$

For (2.20), we require $D_2^\top y = -u$. Substituting (2.18)-(2.20) into (2.17), we arrive at the dual problem

$$\begin{aligned} \max_y \quad & -\frac{1}{2}y^\top D_1 S^{-1} D_1^\top y - w^\top S^{-1} D_1^\top y \\ & ||y||_\infty \leqslant 1 \\ & D_2^\top y = -u \end{aligned}$$

Translating this maximum into a minimum, and using the optimality condition in (2.21), we have the result.

□

2.6.2. Proofs from Section 2.4. *Proof of Theorem 2.4.2.*

We first show that $\hat{h}^2 = O_P(n^{-2/3})$. Fixing $\varepsilon > 0$, we want to show there are $M \in \mathbb{R}$ and $N \in \mathbb{N}$ such that $n \geqslant N$ gives $\mathbb{P}(n^{2/3}\hat{h}^2 > M) < \varepsilon$.

We will show momentarily that $\hat{h}^2 = O_P(n^{-2/3})$ in both the cases when $\hat{h} \leqslant n^{-1/3}(1+\hat{I}+I_0)$ and $\hat{h} > n^{-1/3}(1+\hat{I}+I_0)$. Once we have established that both cases are $O_P(n^{-2/3})$, there exists $M_1, M_2 \in \mathbb{R}$, $N_1, N_2 \in \mathbb{N}$ such that $n \geqslant N_1$ gives

$$\mathbb{P}\left(\left\{n^{2/3}\hat{h}^2 \geqslant M_1\right\} \bigcap \left\{\hat{h} \leqslant n^{-1/3}(1+\hat{I}+I_0)\right\}\right) < \varepsilon/2$$

and $n \geqslant N_2$ gives

$$\mathbb{P}\left(\left\{n^{2/3}\hat{h}^2 \geqslant M_2\right\} \bigcap \left\{\hat{h} > n^{-1/3}(1+\hat{I}+I_0)\right\}\right) < \varepsilon/2.$$

Therefore, for $N = \max\{N_1, N_2\}$ and $M = \max\{M_1, M_2\}$,

$$\begin{aligned}
&\mathbb{P}\left(n^{2/3}\hat{h}^2 > M\right) \\
=&\mathbb{P}\left(\left\{n^{2/3}\hat{h}^2 \geqslant M\right\} \bigcap \left\{\hat{h} \leqslant n^{-1/3}(1+\hat{I}+I_0)\right\}\right) \\
+&\mathbb{P}\left(\left\{n^{2/3}\hat{h}^2 \geqslant M\right\} \bigcap \left\{\hat{h} > n^{-1/3}(1+\hat{I}+I_0)\right\}\right) \\
<&\varepsilon/2 + \varepsilon/2 = \varepsilon.
\end{aligned}$$

This gives that $\hat{h}^2 = O_P(n^{-2/3})$.

We turn next to showing that $\hat{h}^2 = O_P(n^{-2/3})$ in both of the cases indicated. From the basic inequality, Theorem 2.4.3, we have

$$\hat{h}^2 \leqslant 16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_0 - \hat{I}).$$

Take

$$\lambda_n = \max\left\{\sup_{h(f,f_0) \leqslant n^{-1/3}(1+I(f)+I_0)} \frac{4\left|(P_n - P)(p_f)\right|}{1+I(f)+I_0}, n^{-2/3}\right\}.$$

The maximum guarantees that λ_n satisfies the assumption on λ in Theorem 2.2.4 for n large enough, so that $\hat{f}_n$ is well-defined.

We prove in Theorems 2.6.10 and 2.6.9 that

$$\sup_{h(f,f_0) > n^{-1/3}(1+I(f)+I_0)} \frac{n^{1/2}\left|(P_n - P)(p_f)\right|}{h^{1/2}(f,f_0)(1+I(f)+I_0)^{1/2}} = O_P(1) \tag{2.22}$$

and

$$\sup_{h(f,f_0) \leqslant n^{-1/3}(1+I(f)+I_0)} \frac{n^{2/3}\left|(P_n - P)(p_f)\right|}{1+I(f)+I_0} = O_P(1). \tag{2.23}$$

Equation (2.23) gives that $\lambda_n = O_P(n^{-2/3})$.

First, assume that $h(\hat{f}_n, f_0) \leqslant n^{-1/3}(1+I+I_0)$,

$$\hat{h}^2 \leqslant 16(P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(I_0 - \hat{I}) \tag{2.24}$$

$$= (P_n - P)(p_{\hat{f}_n}) + 4\lambda_n(1 + 2I_0) - 4\lambda_n(1 + I_0 + \hat{I}) \tag{2.25}$$

$$= 4(1 + I_0 + \hat{I}) \left(\frac{4(P_n - P)(p_{\hat{f}_n})}{1 + I_0 + \hat{I}} - \lambda_n \right) + 4\lambda_n(1 + 2I_0) \tag{2.26}$$

$$\leqslant 4(1 + I_0 + \hat{I}) \left(\sup_{h(f,f_0) \leqslant n^{-1/3}(1+I(f)+I_0)} \frac{4(P_n - P)(p_f)}{1 + I_0 + I(f)} - \lambda_n \right) + 4\lambda_n(1 + 2I_0) \tag{2.27}$$

Our choice of λ_n gives that the left term in this expression is less than or equal to zero. We conclude that

$$\hat{h}^2 \leqslant 4\lambda_n(1 + 2I_0). \tag{2.28}$$

And finally

$$\frac{\hat{h}}{\sqrt{\lambda_n}} \leqslant 2\sqrt{1 + 2I_0}.$$

By our choice of λ_n, this bound gives that $\hat{h}^2 = O_P(n^{-2/3})$.

Assume next that $\hat{h} > n^{-1/3}(1 + \hat{I} + I_0)$. Define subsets of the probability space

$$B_L = \left\{ \sqrt{n} \left| (P_n - P)(p_{\hat{f}}) \right| > L \cdot \hat{h}^{1/2} \cdot (1 + \hat{I} + I_0)^{1/2} \right\} \tag{2.29}$$

and

$$C_M = \{\hat{I} > M \geqslant I_0\}. \tag{2.30}$$

By (2.22), for each ε there is a corresponding L such that $\mathbb{P}(B_L) < \varepsilon$.

On $B_L^c \cap C_M$, $(I_0 - \hat{I}) < 0$, so from (2.4.3)

$$\sqrt{n}\hat{h}^2 \leqslant 16 \cdot L \cdot \hat{h}^{1/2}(1 + \hat{I} + I_0)^{1/2} + 4\lambda_n\sqrt{n}(I_0 - \hat{I}) \leqslant 16 \cdot L \cdot \hat{h}^{1/2}(1 + \hat{I} + I_0)^{1/2}.$$

Therefore,

$$\hat{h}^{1/2} \leqslant n^{-1/6} \left(16 \cdot L \cdot (1 + \hat{I} + I_0)^{1/2} \right)^{1/3}.$$

Again using (2.4.3), we have

$$\begin{aligned}\sqrt{n}\hat{h}^{3/2} &\leqslant 16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}+\frac{4\lambda_n\sqrt{n}(I_0-\hat{I})}{\hat{h}^{1/2}}\\ &\leqslant 16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}+\frac{4\lambda_n\sqrt{n}(I_0-\hat{I})}{n^{-1/6}\left(16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}\right)^{1/3}}\\ &= 16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}+\frac{4\lambda_n n^{2/3}(I_0-\hat{I})}{\left(16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}\right)^{1/3}}.\end{aligned}$$

The second inequality follows because $I_0-\hat{I}<0$ on C_M. The definition of λ_n gives that $\lambda_n n^{2/3}\geqslant 1$. Hence,

$$\leqslant 16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}+\frac{4(I_0-\hat{I})}{\left(16\cdot L\cdot(1+\hat{I}+I_0)^{1/2}\right)^{1/3}}.$$

The order of the left term is $\sqrt{\hat{I}}$, whereas the order of the right is $\hat{I}^{5/6}$. This gives that for M large enough $\sqrt{n}\hat{h}^{3/2}<0$. Of course this is not possible, so we conclude that for any fixed L, there is M large enough so that $B_L^c\cap C_M=\varnothing$.

Choose L such that $\mathbb{P}(B_L)<\varepsilon/2$. That fact that we can do so is guaranteed by (2.22). Choose M such that $B_L^c\cap C_M=\varnothing$ on this set.

We then have, on $B_L^c=B_L^c\cap C_M^c$,

$$\begin{aligned}\sqrt{n}\hat{h}^2 &\leqslant 16\cdot L\cdot\hat{h}^{1/2}(1+I_0+\hat{I})^{1/2}+4\lambda_n\sqrt{n}(I_0-\hat{I})\\ &\leqslant 16\cdot L\cdot\hat{h}^{1/2}(1+I_0+M)^{1/2}+4\lambda_n\sqrt{n}I_0\\ &\leqslant 2\max\left\{16\cdot L\cdot\hat{h}^{1/2}(1+I_0+M)^{1/2},4\lambda_n\sqrt{n}I_0\right\}.\end{aligned}\tag{2.31}$$

From this, we conclude

$$\hat{h}\leqslant\max\{n^{-1/3}\cdot(32L)^{2/3}\cdot(1+I_0+M)^{1/3},\sqrt{\lambda_n\cdot 8\cdot I_0}\}$$

on B_L^c. Choose K so that $\mathbb{P}(\lambda n^{2/3}>K)<\varepsilon/2$, which is permitted because $\lambda=O_P(n^{-2/3})$. We have

$$\hat{h}\cdot n^{1/3}\leqslant\max\{(32L)^{2/3}(1+I_0+M)^{1/3},\sqrt{8\cdot K\cdot I_0}\}.\tag{2.32}$$

The right hand side is constant, depending on the choice of ε. The set on which this bound does not hold has probability less than ε, by the choice of B_L and K.

Having examined probabilistic rates for both cases $\hat{h} \leqslant n^{-1/3}(1+\hat{I}+I_0)$ and $\hat{h} > n^{-1/3}(1+\hat{I}+I_0)$, we turn next to proving the same rate for squared Hellinger risk. This requires a more refined application of Theorems 2.6.10 and 2.6.9.

We will show that there exist $n_0 \in \mathbb{N}$ and $c \geqslant 0$ such that $n \geqslant n_0$ implies $\mathbb{E}_{f_0}[\hat{h}^2 n^{2/3}] \leqslant c$. We have

$$\mathbb{E}_{f_0}[\hat{h}^2 n^{2/3}] = \mathbb{E}_{f_0}[\hat{h}^2 n^{2/3}(\mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)} + \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)})]. \tag{2.33}$$

We will consider both terms in this summand individually. First, we have

$$\mathbb{E}_{f_0}[\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)}] = \int_0^\infty \mathbb{P}(\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)} \geqslant u)\, du.$$

From (2.28),

$$\int_0^\infty \mathbb{P}(\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)} \geqslant u)\, du \leqslant \int_0^\infty \mathbb{P}(4\lambda_n(1+2I_0)n^{2/3} \geqslant u)\, du. \tag{2.34}$$

From the definition of λ_n and Theorem 2.6.10,

$$\mathbb{P}(4\lambda_n(1+2I_0)n^{2/3} \geqslant u) \leqslant c_0 \exp\left[-\frac{u}{4(1+2I_0)c_0^2}\right]$$

for n and u large. This allows us to integrate the right-hand side of (2.34), which gives that $\mathbb{E}_{f_0}\left[\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}\leqslant n^{-1/3}(1+\hat{I}+I_0)}\right]$ is finite.

On the other hand, consider the second expectation $\mathbb{E}\left[\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)}\right]$. Again we have

$$\mathbb{E}_{f_0}\left[\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)}\right] = \int_0^\infty \mathbb{P}(\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)} \geqslant u)\, du. \tag{2.35}$$

Denote by A_u the event that $\{\hat{h}^2 n^{2/3} \mathbb{1}_{\hat{h}>n^{-1/3}(1+\hat{I}+I_0)} \geqslant u\}$. Let B_L and C_M be as in (2.29)-(2.30), and denote by Λ_K the event $\{\lambda_n n^{2/3} > K\}$. Choosing $L = \left(\frac{u^3}{3\cdot 32^2}\right)^{1/7}$, $M = L^5$, and $K = \frac{u^2}{8I_0}$ gives (2.31) for large enough u. Furthermore, both of the arguments in the maximum of (2.32) are less than u. Recalling that

$B_L^c \cap C_M = B_L^c$, this gives

$$\begin{aligned}\mathbb{P}(A_u) &\leqslant \mathbb{P}(A_u \cap B_L) + \mathbb{P}(A_u \cap B_L^c \cap \Lambda_K^c) + \mathbb{P}(A_u \cap B_L^c \cap \Lambda_K) \\ &\leqslant \mathbb{P}(B_L) + 0 + \mathbb{P}(\Lambda_K) \\ &\leqslant c \exp\left[-\frac{L}{c^2}\right] + c_0 \exp\left[-\frac{K}{c_0^2}\right].\end{aligned}$$

This last inequality follows from Theorems 2.6.10 and 2.6.9. The fact that $\mathbb{P}(A_u \cap B_L^c \cap \Lambda_K)$ equals zero follows from (2.32) and our choice of L, M, and K. Therefore the expectation in (2.35) is finite. Since we have shown that both of the expectations in (2.33) are bounded by constants for n_0 large enough, the result is proven. □

Proof of the Basic Inequality, Lemma 2.4.3.

We have

$$\begin{aligned}4P_n(p_{\hat{f}_n}) - \lambda_n \hat{I} &= 2\int \log\left(\frac{\hat{f}_n + f_0}{2f_0}\right) dP_n - \lambda_n I_0 \\ &\geqslant \int \log\left(\frac{\hat{f}_n}{f_0}\right) dP_n - \lambda_n \hat{I} \\ &\geqslant -\lambda I_0\end{aligned}$$

The first inequality comes from the concavity of log. The second is from the definition of $\hat{f}_n$ as the minimizer of $\int \log f \, dP_n + \lambda_n I(f)$, which implies $\int \log \hat{f}_n \, dP_n + \lambda_n \hat{I} \leqslant \int \log f_0 \, dP_n + \lambda_n I_0$.

We also have that

$$\begin{aligned}-16\int p_{\hat{f}_n} dP &\geqslant 16h^2\left(\frac{\hat{f}_n + f_0}{2}, f_0\right) \\ &\geqslant h^2(\hat{f}_n, f_0),\end{aligned}$$

by Lemmas 4.1 and 4.2 in [**Gee00**].

Therefore,

$$\begin{aligned} 16\int p_{\hat{f}_n}\, d(P_n - P) - 4\lambda_n \hat{I} &\geqslant -16\int p_{\hat{f}_n}\, dP - 4\lambda_n I_0 \\ &\geqslant 16h^2\left(\frac{\hat{f}_n + f_0}{2}, f_0\right) - 4\lambda_n I_0 \\ &\geqslant h^2(\hat{f}_n, f_0) - 4\lambda_n I_0 \end{aligned}$$

This proves the result. □

2.6.3. Empirical Process Results. The goal of this section is to prove the following statements, Theorems 2.4.4 and 2.4.5, which were used in the proof of Theorem 2.4.2.

$$\sup_{h(f,f_0)\leqslant n^{-1/3}(1+I(f)+I_0)} \frac{n^{2/3}\,|(P_n - P)(p_f)|}{1 + I(f) + I_0} = O_P(1) \tag{2.36}$$

$$\sup_{h(f,f_0)>n^{-1/3}(1+I(f)+I_0)} \frac{n^{1/2}\,|(P_n - P)(p_f)|}{h^{1/2}(f,f_0)(1 + I(f) + I_0)^{1/2}} = O_P(1) \tag{2.37}$$

These are the simplifications of the results of Theorems 2.6.10 and 2.6.9, respectively. We begin by introducing notation and relevant definitions.

The *Bernstein Difference* for a parameter $K \in \mathbb{N}$, is given by ρ_K, where

$$\rho_K^2(g) = 2K^2 \int \left(e^{|g|/K} - 1 - |g|/K\right) dP$$

Generalized entropy with bracketing, denoted $\mathscr{H}_{B,K}$ is entropy with bracketing, where the $L_2(P)$ metric is replaced by the Bernstein difference ρ_K. That is, $\mathscr{H}_{B,K}(\varepsilon, \mathscr{G}, P)$ is the logarithm of minimal number of ε-brackets needed to cover G. The bracket for a pair of functions (g_l, g_u) is the set of functions g with $g_l \leqslant g \leqslant g_u$. An ε-bracket (with respect to ρ_K) adds the further condition that $\rho_K(g_l, g_u) < \varepsilon$. A collection of ε-bracket covers $\mathscr{G}$ if each $g \in \mathscr{G}$ belongs to one of the ε-brackets. We denote by H_B the typical entropy with bracketing; that is, with brackets formed in the $L_2(P)$ metric.

The following theorem is an important tool at our disposal.

THEOREM 2.6.3 ([Gee00], 5.11). *Let $\mathscr{G}$ be a function class which satisfies*

$$\sup_{g\in\mathscr{G}} \rho_K(g) \leqslant R.$$

Then there is a universal constant C such that for any a, C_0, C_1 which satisfy

$$a \leqslant C_1\sqrt{n}R^2/K, \tag{2.38}$$

$$a \geqslant C_0\left(\max\left\{\int_0^R \mathscr{H}_{B,K}^{1/2}(u,\mathscr{G},P)\,du,\ R\right\}\right), \tag{2.39}$$

$$C_0^2 \geqslant C^2(C_1+1), \tag{2.40}$$

we have

$$\mathbb{P}\left(\sup_{g\in\mathscr{G}}\left|\sqrt{n}(P_n-P)(g)\right| \geqslant a\right) \leqslant C\exp\left[-\frac{a^2}{C^2(C_1+1)R^2}\right].$$

Our statement of Theorem 2.6.3 is a simplification of the full statement in the listed reference. Because we only work with bracketing entropy integrals which are convergent, we simplify according to the author's comments following the theorem, omitting the second condition in the full statement and taking the lower bound in the bracketing entropy integral in (2.39) to be zero.

The following lemmas will also be required.

LEMMA 2.6.4 ([Gee00], 5.8). *Suppose that*

$$||g||_\infty \leqslant K$$

and

$$||g||_2 \leqslant R.$$

Then

$$\rho_{2K}(g) \leqslant \sqrt{2}R.$$

LEMMA 2.6.5 ([Gee00], 5.10). *Suppose $\mathscr{G}$ is a set of functions such that*

$$\sup_{g\in\mathscr{G}}||g||_\infty \leqslant K.$$

Then

$$\mathscr{H}_{B,4K}(\sqrt{2}\delta,\mathscr{G},P) \leqslant H_B(\delta,\mathscr{G},P) \text{ for all } \delta > 0.$$

LEMMA 2.6.6 ([Gee00], 7.2 & 4.2). *Let p_f be of the form $p_f = \frac{1}{2}\log\frac{f+f_0}{2f_0}$, as occurred in Lemma 2.4.3. Then*

$$\rho_1(p_f) \leqslant 4h\left(\frac{f+f_0}{2}, f_0\right) \leqslant \frac{4h(f,f_0)}{\sqrt{2}}$$

LEMMA 2.6.7. *Let L and K be natural numbers such that $L > K$. Then for any function g, $\rho_K(g) \geqslant \rho_L(g)$.*

PROOF. From the Taylor series expansion of e^x,

$$\begin{aligned}
\rho_K^2(g) &= 2K^2 \int \left(e^{|g|/K} - 1 - |g|/K\right) dP \\
&= 2\int K^2 \sum_{m=2}^{\infty} \frac{|g|^m}{m!\cdot K^m}\, dP \\
&= 2\int \sum_{m=2}^{\infty} \frac{|g|^m}{m!\cdot K^{m-2}}\, dP \\
&\geqslant 2\int \sum_{m=2}^{\infty} \frac{|g|^m}{m!\cdot L^{m-2}}\, dP \\
&= \rho_L^2(g)
\end{aligned}$$

□

This last lemma is a culmination of new results on bracketing entropy. Its proof can be found in Appendix 2.7, along with other contributions on bracketing entropy of function classes with uniformly bounded variation. We denote the quantity $1 + I(f) + I_0$ by $J(f)$.

LEMMA 2.6.8. *The set of functions $\mathscr{P}_M = \{p_f : J(f) \leqslant M\}$ satisfies, for some constant A,*

$$H_B(\delta, \mathscr{P}_M, P) \leqslant A \cdot \frac{M}{\delta}, \quad \forall \delta > 0.$$

Furthermore, $p_f \in \mathscr{P}_M$ implies $||p_f||_\infty < M$.

With these lemmas in hand, we are ready to state and prove our main results. We will prove a sequence of constrained results, and then use a peeling device to obtain the concentration inequalities. The method of proof, and particularly our use of the peeling device, is interesting in its own right. Our first result is

Lemma 2.4.6, which establishes bounds for the supremum of the empirical process indexed by $\{f : J(f) \leqslant M \text{ and } h(f, f_0) \leqslant \delta\}$ for constants δ and M.

Proof of Lemma 2.4.6.

By Lemma 2.6.6, $h(f, f_0) \leqslant \delta$ gives that $\rho_1(p_f) \leqslant \frac{4}{\sqrt{2}}\delta = 2^{3/2}\delta$.

By Lemma 2.6.7, $\rho_1(p_f) \leqslant 2^{3/2}\delta$ gives that $\rho_{4M}(p_f) \leqslant 2^{3/2}\delta$ for all $M \geqslant 1$.

By Lemma 2.6.8, $p_f \in \mathscr{P}_M$ gives that $||p_f||_\infty \leqslant M$. From Lemmas 2.6.5 and 2.6.8.

$$\mathscr{H}_{B,4M}(\delta, \mathscr{P}_M, P) \leqslant H_B(\delta/\sqrt{2}, \mathscr{P}_M, P) \leqslant \frac{A\sqrt{2}M}{\delta}.$$

Collecting these facts, we seek to apply Theorem 2.6.3. We have $\rho_{4M}(p_f) \leqslant 2^{3/2}\delta$. From the conditions in the theorem (with $R = 2^{3/2}\delta$, $K = 4M$ and $a = 2^{-1/2}C_1\sqrt{M}\delta^{1/2}$), it suffices to choose δ, C_0, C_1 such that

$$a \leqslant C_1\sqrt{n}\frac{2^3\delta^2}{4M} = \frac{2C_1\sqrt{n}\delta^2}{M} \tag{2.41}$$

$$a \geqslant C_0 \int_0^R H_B^{1/2}(u/\sqrt{2}, \mathscr{P}_M, P) = 2C_0\sqrt{AM\delta} \tag{2.42}$$

$$C_0^2 \geqslant C^2(C_1 + 1) \tag{2.43}$$

Choose $C_1 = 2C_0\sqrt{2A}$. Then (2.41) is satisfied for $\delta \geqslant \frac{M}{2} \cdot n^{-1/3}$, (2.42) is satisfied by the choice of a, and (2.43) is satisfied for large enough C_0. By Theorem 2.6.3 we have for all $\delta \geqslant \frac{M}{2} \cdot n^{-1/3}$ (if $C_1 \geqslant 1$)

$$\begin{aligned}\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, h(f,f_0) \leqslant \delta} \left|\sqrt{n}(P_n - P)(p_f)\right| \geqslant 2C_1\sqrt{M}\delta^{1/2}\right) &\leqslant C\exp\left[-\frac{4C_1^2 M\delta}{C^2(C_1+1)2^3\delta^2}\right] \\ &\leqslant C\exp\left[-\frac{C_1 M\delta^{-1}}{4C^2}\right]\end{aligned}$$

□

THEOREM 2.6.9. *There are constants c, n_0 and t_0 so that when $n \geqslant n_0$ and $T \geqslant t_0$*

$$\mathbb{P}\left(\sup_{p_f \in P, h(f,f_0) > n^{-1/3}J(f)} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f, f_0)J^{1/2}(f)} \geqslant T\right) \leqslant c\exp\left[-\frac{T}{c^2}\right].$$

PROOF. We first prove the following: there are constants n_0, t_0 and c_0 such that for all $n \geqslant n_0$, $T \geqslant t_0$, and $M \geqslant 1$

$$\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, h(f,f_0) > \frac{M}{2} n^{-1/3}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f,f_0)} \geqslant T\sqrt{\frac{M}{2}} \right) \leqslant c_1 \exp\left[-\frac{TM}{c_1^2} \right]. \tag{2.44}$$

The proof of this claim is an application of the peeling device [**Gee00**][Section 5.3] to Lemma 2.4.6. Let $S = \min\{s \in \mathbb{N} : 2^{-s} < \frac{M}{2} n^{-1/3}\}$. We will form a union bound by partitioning into sets with $\{2^{-s-1} < h(f,f_0) \leqslant 2^{-s}\}$ for integer-valued s. Because Hellinger distance is bounded above by 1, we need not consider negative values of s. Let $T = 4C_1$. Applying this union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, h(f,f_0) > \frac{M}{2} n^{-1/3}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f,f_0)} \geqslant T\sqrt{\frac{M}{2}} \right) \\
&\leqslant \sum_{s=1}^{S} \mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, 2^{-s} < h(f,f_0) \leqslant 2^{-s+1}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f,f_0)} \geqslant T\sqrt{\frac{M}{2}} \right) \\
&\leqslant \sum_{s=1}^{S} \mathbb{P}\left(\sup_{p_f \in \mathscr{P}_M, h(f,f_0) \leqslant 2^{-s+1}} \left|\sqrt{n}(P_n - P)(p_f)\right| \geqslant 2^{\frac{-s+1}{2}} \cdot 2C_1\sqrt{M} \right)
\end{aligned}$$

We have $2^{-s+1} \geqslant \frac{M}{2} n^{-1/3}$ for $s \leqslant S$, so applying Lemma 2.4.6 gives the further bound

$$\begin{aligned}
&\leqslant \sum_{s=1}^{S} C \cdot \exp\left[-\frac{C_1 \cdot M \cdot (2^{-s+1})^{-1}}{4C^2} \right] \\
&= \sum_{s=1}^{S} C \cdot \exp\left[-\frac{C_1 \cdot M \cdot 2^{s-1}}{4C^2} \right] \\
&\leqslant \sum_{s=1}^{S} C \cdot \exp\left[-\frac{C_1 M}{8C^2} - \frac{2^{s-2}}{4C^2} \right] \\
&\leqslant \exp\left[-\frac{C_1 M}{8C^2} \right] \sum_{s=1}^{S} C \exp\left[-\frac{2^{s-2}}{4C^2} \right] \\
&= c_1 \exp\left[-\frac{TM}{c_1^2} \right].
\end{aligned} \tag{2.45}$$

Here, c_1 is some constant, since the final summation is convergent as S approaches infinity. The third inequality in this chain follows from $C_1 M 2^{s-1} \geqslant \frac{MC_1}{2} + MC_1 2^{s-2}$, so that when $M \geqslant 1$ and $C_1 \geqslant 1$,

$$C_1 M 2^{s-1} \geqslant \frac{MC_1}{2} + 2^{s-2}.$$

Of course, it suffices to consider $M \geqslant 1$ because $J(f) \geqslant 1$. This proves the claim.

We use the claim to prove the result by again applying the peeling device, but this time with respect to $J(f)$. Because $J(f) \geqslant 1$, we need only peel in sets $\{2^s \leqslant J(f) \leqslant 2^{s+1}\}$ for $s \geqslant 0$. This gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{p_f \in \mathscr{P}, h(f,f_0) > n^{-1/3} J(f)} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f,f_0) J^{1/2}(f)} \geqslant T \right) \\
&\leqslant \sum_{s=0}^{\infty} \mathbb{P}\left(\sup_{p_f \in \mathscr{P}, J(f) \leqslant 2^{s+1}, h(f,f_0) > n^{-1/3} 2^s} \frac{|\sqrt{n}(P_n - P)(p_f)|}{h^{1/2}(f,f_0)} \geqslant T 2^{s/2} \right).
\end{aligned}$$

Applying the claim and manipulating as in (2.45), there is a constant c which permits the following bound.

$$\begin{aligned}
&\leqslant \sum_{s=0}^{\infty} c_1 \exp\left[-\frac{T 2^{s+1}}{c_1^2} \right] \\
&\leqslant \exp\left[-\frac{T}{2c_1^2} \right] \sum_{s=0}^{\infty} c_1 \exp\left[-\frac{2^s}{c_1^2} \right] \\
&\leqslant c \exp\left[-\frac{T}{c^2} \right].
\end{aligned}$$

□

THEOREM 2.6.10. *There are constants n_0, t_0, and c such that for all $n \geqslant n_0$ and $T \geqslant t_0$*

$$\mathbb{P}\left(\sup_{p_f \in \mathscr{P}, h(f,f_0) \leqslant n^{-1/3} J(f)} \frac{\left| n^{2/3}(P_n - P)(p_f) \right|}{J(f)} \geqslant T \right) \leqslant c_0 \exp\left[-\frac{T}{c_0^2} \right]$$

PROOF. First we apply the peeling device to the quantity $J(f)$. We partition into sets with $2^s < J(f) \leqslant 2^{s+1}$. Since $J(f) \geqslant 1$, it suffices to take $s \geqslant 0$. We have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{p_f \in \mathscr{P}, h(f,f_0) \leqslant n^{-1/3} J(f)} \frac{|\sqrt{n}(P_n - P)(p_f)|}{J(f) n^{-1/6}} \geqslant T \right) \\
&= \mathbb{P}\left(\sup_{p_f \in \mathscr{P}, h(f,f_0) \leqslant n^{-1/3} J(f)} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)} \sqrt{J(f) n^{-1/3}}} \geqslant T \right) \\
&\leqslant \sum_{s=0}^{S} \mathbb{P}\left(\sup_{p_f \in \mathscr{P}, h(f,f_0) \leqslant n^{-1/3} J(f), 2^s \leqslant J(f) \leqslant 2^{s+1}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)} \sqrt{J(f) n^{-1/3}}} \geqslant T \right)
\end{aligned}$$

We peel this expression in $h(f,f_0)$. For $s \in \mathbb{N}$, let $R_s = \max\{r \in \mathbb{N} : 2^{-r} \geqslant n^{-1/3} 2^{s+1}\}$. Let

$$\mathscr{N}_{s,r} = \{p_f \in \mathscr{P}, 2^{-r-1} < h(f,f_0) \leqslant 2^{-r}, 2^s \leqslant J(f) \leqslant 2^{s+1}, h(f,f_0) \leqslant n^{-1/3} J(f)\}$$

for $r = 0, \dots, R_s - 1$ and

$$\mathscr{N}_{s,R_s} = \{p_f \in \mathscr{P}, h(f,f_0) \leqslant 2^{s+1} n^{-1/3} \leqslant 2^{-R_s}, 2^s \leqslant J(f) \leqslant 2^{s+1}, h(f,f_0) \leqslant n^{-1/3} J(f)\}.$$

Applying the peeling device gives the further bound.

$$\leqslant \sum_{s=0}^{\infty} \left\{ \sum_{r=0}^{R_s-1} \mathbb{P}\left(\sup_{\mathscr{N}_{s,r}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)}\sqrt{J(f)n^{-1/3}}} \geqslant T \right) + \mathbb{P}\left(\sup_{\mathscr{N}_{s,R_s}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)}\sqrt{J(f)n^{-1/3}}} \geqslant T \right) \right\}. \tag{2.46}$$

In this last term, $J(f)n^{-1/3}$ and $J(f)$ can be bounded below on $\mathscr{N}_{s,R_s}$. Indeed, $J(f) > 2^s$ and

$$J(f)n^{-1/3} > 2^s n^{-1/3} > 2^{-R_s-2}.$$

Inserting these bounds gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathscr{N}_{s,R_s}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)}\sqrt{J(f)n^{-1/3}}} \geqslant T \right) \\
\leqslant &\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_{2^{s+1}}, h(f,f_0) \leqslant 2^{-R_s}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{2^{s/2} 2^{-(R_s+2)/2}} \geqslant T \right).
\end{aligned}$$

Applying Lemma 2.4.6 (with $T = 4\sqrt{2}C_1$) bounds this term by an expression of the form $c_1 \exp\left[-\frac{T 2^{s+1} 2^{R_s}}{c_1^2}\right]$, for some constant c_1. For $r < R_s$, we have the following chain of inequalities

$$\begin{aligned}
&\mathbb{P}\left(\sup_{N_{s,r}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{\sqrt{J(f)}\sqrt{J(f)n^{-1/3}}} \geqslant T \right) \\
\leqslant &\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_{2^{s+1}}, 2^{-r-1} < h(f,f_0) \leqslant 2^{-r}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{2^{s/2} h^{1/2}(f,f_0)} \geqslant T \right) \\
\leqslant &\mathbb{P}\left(\sup_{p_f \in \mathscr{P}_{2^{s+1}}, h(f,f_0) \leqslant 2^{-r}} \frac{|\sqrt{n}(P_n - P)(p_f)|}{2^{s/2} 2^{-(r+1)/2}} \geqslant T \right).
\end{aligned}$$

According to the definition of R_s, $2^{-r} \geqslant 2^{s+1} n^{-1/3}$, so Lemma 2.4.6 (with $T = 4C_1$) allows us to bound this probability by $c_2 \exp\left[-\frac{T 2^{s+1} 2^r}{c_2^2}\right]$.

The double summand (2.46) is thus bounded by

$$\sum_{s=0}^{\infty} \left\{ \sum_{r=0}^{R_s-1} c_1 \exp\left[-\frac{T 2^{s+1} 2^r}{c_1^2}\right] + c_2 \exp\left[-\frac{T 2^{s+1} 2^{R_s}}{c_2^2}\right] \right\}.$$

Reducing twice according to the manipulation in (2.45), this expression can be bounded by a term of the form $c_0 \exp\left[-\frac{T}{c_0^2}\right]$. □

2.6.4. Depth-First Embedding a Geometric Network into $\mathbb{R}$. The goal of this section is to define an embedding γ, of a fixed geometric network G into $\mathbb{R}$, which approximately preserves total variation. Let g be a function of bounded variation on G. On each edge $e = [a_e, b_e]$,

$$\mathrm{TV}_G(g|_e) = \left|g(a_e) - \lim_{x \searrow a_e} g(x)\right| + \mathrm{TV}(g|_{(a_e,b_e)}) + \left|g(b_e) - \lim_{x \nearrow b_e} g(x)\right|. \tag{2.47}$$

Define

$$\tilde{g}(x) = \begin{cases} \lim_{z \searrow a_e} g(x) & \text{if } x = a_e \\ g(x) & \text{if } x \in (a_e, b_e) \\ \lim_{z \nearrow b_e} g(x) & \text{if } x = b_e \end{cases}.$$

The fact that g is of bounded variation gives that these limits exist. Furthermore,

$$\mathrm{TV}_G(g|_{(a_e,b_e)}) = \mathrm{TV}_G(\tilde{g}|_{[a_e,b_e]}).$$

We can therefore rewrite (2.47) as

$$\mathrm{TV}_G(g|_e) = |g(a_e) - \tilde{g}(a_e)| + \mathrm{TV}(\tilde{g}|_e) + |g(b_e) - \tilde{g}(b_e)|.$$

From this equation we conclude

$$\mathrm{TV}_G = \sum_{e \in E} |g(a_e) - \tilde{g}(a_e)| + \mathrm{TV}(\tilde{g}|_e) + |g(b_e) - \tilde{g}(b_e)|. \tag{2.48}$$

Equation (2.48) gives us the following insight: the total variation of a function on a network can be decomposed into jumps at nodes and total variation along an open intervals. By separating limit nodes from the true value at the node, we can create an expanded network that represents this decomposition. For each node, and each edge incident to that node, define a limit node as the limit approaching the node along the incident edge. Similarly, define a value node as the true value at a node. The expanded graph is the defined on the limit and value nodes, with the inherited connectivity. Open intervals corresponding to the original edges in the geometric graph are edges between two limit nodes, and value nodes are only connected to limit

nodes. We perform this expansion in order to guarantee that each edge in the original network is traversed by a depth-first search.

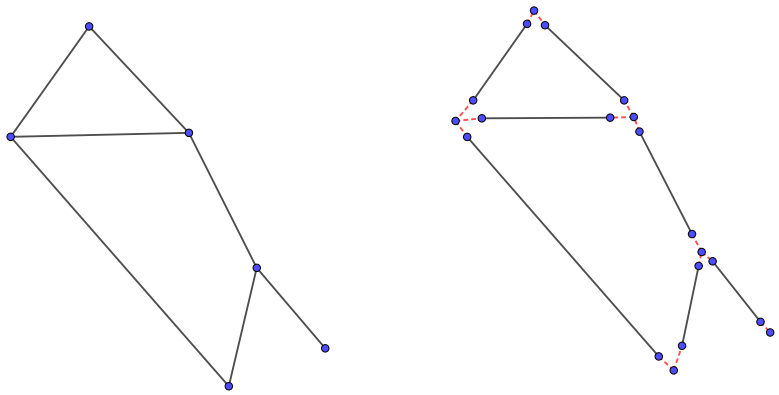


FIGURE 2.7. A geometric network on the left, and its expansion on the right. In the expansion, red edges represent edges between limit nodes and value nodes. Black edges correspond to open intervals in the original geometric network.

In order to perform the embedding of G into $\mathbb{R}$, we apply a slight modification of the technique in depth-first search fused lasso [**PSST16**]. The idea is this: traverse the nodes of the expanded network according to depth-first search, starting at some arbitrary root node. Glue edges together according to the order in which they are visited in the depth-first search. Each of the intervals will be traversed, according to depth-first search. For any function the total variation of the resulting univariate embedding never exceeds twice that of the graph-induced total variation. We formalize this result in the following theorem.

THEOREM 2.6.11. *Let G be a connected geometric network, and $\gamma : G \to \mathbb{R}$ be the embedding of G into $\mathbb{R}$ according to depth-first search. Then*

(i) Each edge of the original (non-expanded) graph is traversed.

(ii) $\mathrm{TV}(g \circ \gamma^{-1}) \leqslant 2\,\mathrm{TV}_G(g)$ *for all* $g : G \to \mathbb{R}$.

(iii) Because we use them simultaneously, we dnote μ and μ_G denote the Lebesgue and base measure on $\mathbb{R}$ and G, respectively. For any function $f : G \to \mathbb{R}$ and set $A \subseteq \mathbb{R}$,

$$\int_{\gamma^{-1}(A)} f \, d\mu_G = \int_A f \circ \gamma^{-1} \, d\mu.$$

It follows that for a random variable X on G with density f_0, $\gamma(X)$ has density $f \circ \gamma^{-1}$. Furthermore, for any functions f and f_0 on G, $h(f, f_0) = h(f \circ \gamma^{-1}, f_0 \circ \gamma^{-1})$.

PROOF. For (i), assume for contradiction that there is an open interval of the network G which is not traversed in the depth-first search of the expanded network. Because the degree of a limit node is two, a limit node must have been a leaf of the DFS spanning tree. But this cannot be. Indeed, one of the limit nodes of the open interval must have been reached first in the depth-first search. Because limit nodes have degree two, when DFS reached that limit node it would proceed across the edge, contradicting that the open interval was not traversed.

For (ii), consider two nodes visited consecutively in DFS of the expanded graph: $\tau(i)$ and $\tau(i+1)$, the ith and $i+1$th nodes visited, respectfully. There are two cases to consider. First, assume that $\tau(i)$ is not a leaf of the DFS tree. This implies there is an edge e such that $\mathrm{TV}(g|_{\tau(i)}^{\tau(i+1)}) = \mathrm{TV}(g|_e)$. For the other case, assume that $\tau(i)$ is a leaf of the DFS tree. From (i) we know that $\tau(i)$ is not a limit node. And because every limit node has degree two, we have that $\tau(i+1)$ is a limit node. Hence the univariate total variation between $\tau(i)$ and $\tau(i+1)$ is $|g(\tau(i+1)) - g(\tau(i))|$. Furthermore, there is a path π, traversed by DFS, such that π starts at $\tau(i)$ and ends at $\tau(i+1)$. This requires that the network G be connected, so that the path π is a subset of the graph G. According to the triangle inequality,

$$\mathrm{TV}(g|_{\tau(i)}^{\tau(i+1)}) \leqslant \mathrm{TV}_G(g|_\pi).$$

We next use the following fundamental property of DFS (see for example, [**Cor09**]): DFS visits each edge exactly twice. In other words, each edge in G can occur as a member of π at most twice. This gives that

$$\mathrm{TV}(g) = \sum_i \mathrm{TV}(g|_{\tau(i)}^{\tau(i+1)}) \leqslant \sum_i \mathrm{TV}_G(g|_{pi}) \leqslant 2\sum_{e\in E} \mathrm{TV}_G(g|_e).$$

For (iii), let $f : G \to \mathbb{R}$, and $A \subseteq \gamma(G) \subset \mathbb{R}$. Then

$$\int_{\gamma^{-1}(A)} f\, d\mu_G = \sum_{e\in E} \int_{\gamma^{-1}(A)\cap e} f\, d\mu_G.$$

On each edge e, γ is the identity and $\mu_G = \mu$. Therefore,

$$\sum_{e\in E} \int_{\gamma^{-1}(A)\cap e} f\, d\mu_G = \sum_{e\in E} \int_{A\cap\gamma(e)} f\circ\gamma^{-1}\, d\mu = \int_A f\circ\gamma^{-1}\, d\mu.$$

The remaining claims follow from this result. For any random variable x on G,

$$\mathbb{P}(\gamma(x) \in A) = \mathbb{P}(x \in \gamma^{-1}(A)) = \int_{\gamma^{-1}(A)} f \, d\mu_G = \int_A f \circ \gamma^{-1} \, d\mu.$$

Therefore $f \circ \gamma^{-1}$ is the density of $\gamma(x)$. Similarly, have that $h(f, f_0) = h(f \circ \gamma^{-1}, f_0 \circ \gamma^{-1})$ because

$$\int_G (\sqrt{f} - \sqrt{f_0})^2 \, d\mu_G = \int_{\gamma^{-1}(\gamma(G))} (\sqrt{f} - \sqrt{f_0})^2 \, d\mu_G = \int_{\gamma(G)} (\sqrt{f \circ \gamma^{-1}} - \sqrt{f_0 \circ \gamma^{-1}})^2 \, d\mu.$$

□

2.7. Appendix: Bracketing Entropy Results

The primary result in this appendix is the following.

THEOREM 2.7.1. *Let $\mathscr{P}_M$ be the set of functions $\{p_f : f \in \mathscr{F}, J(f) \leqslant M\}$. For some constant A, the bracketing entropy of $\mathscr{P}$ satisfies*

$$H_B(\delta, \mathscr{P}_M, P) \leqslant A \cdot \frac{M}{\delta}, \quad \forall \delta > 0.$$

The proof of this result is decomposed into the following lemmas. In Lemma 2.7.2 we show that $\mathscr{P}_M$ is uniformly bounded, has nonnegative and nonpositive values, and has uniformly bounded total variation. In Lemma 2.7.6, we show that any set of functions satisfying these properties is sufficient for the conclusion in Theorem 2.7.1. This gives the result for $\mathscr{P}_M$.

LEMMA 2.7.2. *The set of functions $\{p_f : J(f) \leqslant M\}$ has total variation uniformly bounded by $M/2$, and each function in the set takes nonnegative and nonpositive values. Furthermore, $J(f) \leqslant M$ gives that $||p_f||_\infty \leqslant M/2$.*

PROOF. The assertion about total variation follows from Lemma 2.7.3. We also have that each function takes both nonpositive and nonnegative values. Indeed, consider p_f. From its definition

$$p_f = \frac{1}{2} \log \left(\frac{f + f_0}{2 f_0} \right).$$

For each f, the fact that both f and f_0 integrate to 1 give that for some point $\underline{x} \in \mathscr{X}$ $f(\underline{x}) \leqslant f_0(\underline{x})$. We then have

$$p_f(\underline{x}) = \frac{1}{2} \log \left(\frac{f(\underline{x}) + f_0(\underline{x})}{2 f_0(\underline{x})} \right) \leqslant 0.$$

Similarly, there exists $\bar{x} \in \mathscr{X}$ such that $f(\bar{x}) \geqslant f_0(\bar{x})$. We then have that $p_f(\bar{x}) \geqslant 0$.

The last claim follows by combining both of the above: a function which takes both nonnegative and nonpositive values and has total variation bounded by $M/2$, is bounded by $M/2$ itself. □

LEMMA 2.7.3. *We have the following results.*

(1) For any constant $a \geqslant 0$ and any function f, $\mathrm{TV}(\log(f(x)+a)) \leqslant \mathrm{TV}(\log(f(x)))$.

(2) $J(p_f) \leqslant M$ gives that $\mathrm{TV}(p_f) \leqslant \frac{M}{2}$.

PROOF. The first claim is intuitive because the derivative of log is strictly decreasing. For the proof, consider any two points x_1, x_2 in a compact interval I. Let f be a real-valued function on I. Consider $|\log(f(x_2)+a) - \log(f(x_1)+a)|$ for some $a \geqslant 0$. Without loss of generality, assume that $f(x_2) \geqslant f(x_1)$. Then

$$\begin{aligned} &|\log(f(x_2)+a) - \log(f(x_1)+a)| = \log(f(x_2)+a) - \log(f(x_1)+a) \\ &= \log\left(\frac{f(x_2)+a}{f(x_1)+a}\right) \\ &\leqslant \log\left(\frac{f(x_2)}{f(x_1)}\right). \end{aligned} \tag{2.49}$$

Total variation is defined as the supremum over all point partitions P, in the interval I, of the following sum

$$\mathrm{TV}(g) = \sup_P \sum_{x \in P} |g(x_{i+1}) - g(x_i)|.$$

In computing $\mathrm{TV}(\log(f(x)+a))$, we bound each of the terms in the summand with (2.49), to conclude that $\mathrm{TV}(\log(f(x)+a)) \leqslant \mathrm{TV}(\log(f(x)))$. This gives us the first claim.

For the second claim, we use the following facts about total variation: $\mathrm{TV}(f+g) \leqslant \mathrm{TV}(f) + \mathrm{TV}(g)$, $\mathrm{TV}(-f) = \mathrm{TV}(f)$, and $\mathrm{TV}(c) = 0$ for any constant c. Using these and the first claim, we have

$$\begin{aligned}
\mathrm{TV}(p_f) &= \mathrm{TV}\left(\frac{1}{2}\log\left(\frac{f+f_0}{2f_0}\right)\right) \\
&= \mathrm{TV}\left(\frac{1}{2}\log\left(\frac{f}{2f_0}+\frac{1}{2}\right)\right) \quad \text{so by the first claim} \\
&\leqslant \mathrm{TV}\left(\frac{1}{2}\log\left(\frac{f}{2f_0}\right)\right) \\
&\mathrm{TV}(\frac{1}{2}\log(f) - \frac{1}{2}\log(2f_0)) \\
&\leqslant \frac{1}{2}\mathrm{TV}(\log(f)) + \frac{1}{2}\mathrm{TV}(\log(2f_0)) \\
&\leqslant \frac{1}{2}\mathrm{TV}(\log(f)) + \frac{1}{2}\mathrm{TV}(\log(f_0) + \frac{1}{2}\mathrm{TV}(\log(2)) \\
&\leqslant \frac{J(f)}{2}.
\end{aligned}$$

This gives the second claim. □

We next have a lemma for the bracketing entropy of monotone classes of functions, which we will relate to functions of bounded variation. Denote the *bracketing number* of the function class $\mathscr{F}$ with bracketing width ε and metric $d : \mathscr{F} \times \mathscr{F} \to \mathbb{R}$ by $N_{[]}(\varepsilon, \mathscr{F}, d)$.

LEMMA 2.7.4 ([**VDVW96**], Theorem 2.7.5). *For every probability measure Q, there exists a constant A such that the bracketing of monotone functions $f : \mathbb{R} \to [0,1]$ satisfies*

$$\log N_{[]}(\varepsilon, \mathscr{F}, L_2(Q)) \leqslant K\left(\frac{1}{\varepsilon}\right).$$

LEMMA 2.7.5. *Let $f : [a,b] \to \mathbb{R}$ be a function such that $\mathrm{TV}(f) \leqslant k$, and there are $\bar{x}$ and $\underline{x}$ in $[a,b]$ such that $f(\bar{x}) \geqslant 0$ and $f(\underline{x}) \leqslant 0$. Then f can be represented as the difference of two nondecreasing functions g, h with $\mathrm{TV}(g)$ and $\mathrm{TV}(h)$ bounded by k. Furthermore, for all $x \in [a,b]$,*

$$-k \leqslant g(x) \leqslant k \text{ and } -k \leqslant h(x) \leqslant k.$$

PROOF. Denote by $\mathrm{TV}_{x'}^{x''}(f)$ the total variation of f on $[x', x'']$. Define

$$g(x) := \frac{f(x) + \mathrm{TV}_a^x(f)}{2}, \quad h(x) := \frac{\mathrm{TV}_a^x(f) - f(x)}{2}.$$

Of course, $f = g - h$.

Let $x'' > x'$. Then

$$g(x'') - g(x') = \frac{f(x'') - f(x') + \mathrm{TV}_a^{x''}(f) - \mathrm{TV}_a^{x'}(f)}{2} \tag{2.50}$$

and

$$h(x'') - h(x') = \frac{\mathrm{TV}_a^{x''}(f) - \mathrm{TV}_a^{x'}(f) - (f(x'') - f(x'))}{2}. \tag{2.51}$$

We have that

$$\mathrm{TV}_a^{x''}(f) - \mathrm{TV}_a^{x'}(f) = \mathrm{TV}_{x'}^{x''}(f) \geqslant \left| f(x'') - f(x') \right|$$

which allows us to conclude that (2.50) and (2.51) are positive. Hence g and h are nondecreasing. Lastly,

$$\mathrm{TV}_a^b(g) = \frac{f(b) + \mathrm{TV}_a^b(f) - f(a) + 0}{2} = \frac{f(b) - f(a) + \mathrm{TV}_a^b(f)}{2} \leqslant \mathrm{TV}_a^b(f) \leqslant k$$

and

$$\mathrm{TV}_a^b(h) = \frac{\mathrm{TV}_a^b(f) - f(b) + (0 - f(a))}{2} = \frac{\mathrm{TV}_a^b(f) + f(a) - f(b)}{2} \leqslant \mathrm{TV}_a^b(f) \leqslant k$$

We have shown the total variation bounds.

The inequality in the statement of the lemma follows from the nondecreasing nature of these functions. From this property, we have

$$\frac{f(a)}{2} = g(a) \leqslant g(x) \leqslant g(b) = \frac{k - f(b)}{2} \tag{2.52}$$

and

$$\frac{-f(a)}{2} = h(a) \leqslant h(x) \leqslant h(b) = \frac{k - f(b)}{2}. \tag{2.53}$$

Because of the fact the assumptions on $\bar{x}$ and $\underline{x}$, $|f(a)| \leqslant \max\{|f(a) - f(\underline{x})|, |f(a) - f(\bar{x})|\} \leqslant \mathrm{TV}_a^b(f) = k$. The same is true of $f(b)$. The conclusion follows by substituting these inequalities into (2.52) and (2.53). □

LEMMA 2.7.6. *Let $\mathscr{G}$ be a set of functions each of which has nonnegative and nonpositive values and have total variation bounded by M. Let Q be a probability measure. The bracketing entropy of $\mathscr{G}$ grows like*

$\frac{1}{\varepsilon}$. *That is, for some constant K not depending on Q,*

$$\log N_{[]}(\varepsilon, \mathscr{G}, L_2(Q)) \leqslant K(\frac{M}{\varepsilon})$$

PROOF. Consider the set of functions $\bar{\mathscr{G}} = \frac{1}{M}\mathscr{G}$. From Lemma 2.7.2, the set $\bar{\mathscr{G}}$ maps from $\mathbb{R}$ to $[-1,1]$, and has total variation bounded by 1. By Lemma 2.7.5, $\bar{\mathscr{G}} \subseteq \mathscr{H} - \mathscr{F}$, where each $\mathscr{H}$ and $\mathscr{F}$ contain monotone functions which map $\mathbb{R} \to [-1,1]$. By Lemma 2.7.4, the classes $\bar{\mathscr{H}} := \frac{1}{2}\mathscr{H} + \frac{1}{2}$ and $\bar{\mathscr{F}} := \frac{1}{2}\mathscr{F} + \frac{1}{2}$ each have bracketing numbers of the form $L^{C_1/\varepsilon}$ and $L^{C_2/\varepsilon}$ for constants L, C_1, and C_2. We can form an ε-bracket of $\mathscr{G}$ from all pairs of $\varepsilon/4M$ brackets of $\bar{\mathscr{H}}$ and $\bar{\mathscr{F}}$.

$$g = M\bar{g} = M(h-f) = M2(\bar{h} - \frac{1}{2} - (\bar{f} - \frac{1}{2})) = 2M(\bar{h} - \bar{f}).$$

Access to an $\varepsilon/2M$ bracketing cover of $\bar{\mathscr{H}}$ and $\bar{\mathscr{F}}$ gives functions l, u, a, b such that

$$l \leqslant h \leqslant u, \quad a \leqslant f \leqslant b$$

and both $||l-u||_{L_2(Q)}$ and $||b-a||_{L_2(Q)}$ are less than $\varepsilon/4M$. We then have

$$2M(l-b) \leqslant g \leqslant 2M(u-a).$$

We have formed a bracket in $\mathscr{G}$ of the form $(2M(l-b), 2M(u-a))$. These form an ε bracket because

$$||2M(l-b) - 2M(u-a)||_{L_2(Q)} \leqslant 2M\,||l-u|| + 2M\,||a-b|| < \varepsilon.$$

There are $L^{4MC_1/\varepsilon} \times L^{4MC_2/\varepsilon}$ such brackets, so the bracketing entropy satisfies

$$\log N_{[]}(\varepsilon, \mathscr{G}, L_2(Q)) \leqslant 4(C_1 + C_2)\log(L)\frac{M}{\varepsilon}.$$

This gives the result. □

CHAPTER 3

Maximum a Posteriori Estimators as a Limit of Bayes Estimators

3.1. Introduction

The purpose of this chapter is to relate two point estimates in Bayesian estimation: the maximum a posteriori (MAP) estimator and Bayes estimator. Both the MAP and Bayes estimator are defined in terms of optimization problems, so that any connection between MAP and Bayes estimators can be extended to a connection between corresponding optimization problems. It is commonly accepted ([**Rob07**, §4.1.2] [**Gew05**, Thm.2.4.3], [**Lee12**, §7.6.5]) that *MAP estimation is the limit of Bayes estimation*. This relationship is appealing from a theoretical perspective because it allows MAP estimators to be subsumed by the statistical analysis and intuition of Bayes estimators. However, this assertion must be carefully studied, because it is not true in the general setting proposed in much of the literature. We apply the theory of variational analysis, a common tool used in the approximation of optimization problems, to investigate the relationship between MAP and Bayes estimators.

This chapter revises the relationship between MAP and Bayes estimators, placing the theory on a solid mathematical foundation. First, we provide a counterexample to the commonly accepted notion of MAP estimators as a limit of Bayes estimators having 0-1 loss. We then provide additional conditions and resurrect the limiting relationship between the estimators by relying on the theory of variational analysis. Because each of the estimators is defined as the maximizer of a certain optimization problem, a natural setting for the analysis of this problem is the space of upper semi-continuous functions, where we use the appropriate topology for convergence of maximizers, namely *hypo-convergence*. In general, the approach in this chapter is applicable to any estimator defined in terms of an optimization problem.

This chapter is not the first to apply variational analytic techniques to statistical estimation. One of the earliest examples is the notion of epi-convergence in distribution, which was introduced in [**SW86**]

This chapter is based on joint work with Julio Deride [**BD18**].

and developed further in [Pfl91] and [Kni99]. The majority of applications of epi-convergence in distribution [Pfl95, Kni01, KF00], and general applications of variational analytic theory to statistical estimation problems [KW91, AW94, Sha91, DW88, KR93, Gey94], have focused on asymptotic consistency of estimators: the notion of approximating a *true* optimization problem by one constructed from a finite sample, as the size of the sample tends to infinity.

This chapter differs from the literature in a number of ways. First, we consider the limit of optimization problems defined over the same measure space. Instead of contrasting the empirical solution to the true, we instead assume that the function to be optimized is changing, but that the "true" underlying measure is known a priori. In this sense, we focus more on the approximation of loss functions than on the measure over which their expected value is taken. We also focus on almost-sure convergence results, as opposed to the weaker notion of distributional convergence. Lastly, the convergence in this chapter deals explicitly with the Bayesian framework and the relationship between MAP and Bayes estimators.

The rest of this chapter is organized as follows. In section 3.2, a mathematical formulation of the Bayesian point estimation framework is reviewed, including the commonly-accepted argument of the relationship between MAP and Bayes estimators. Section 3.3 gives variational analytic background, introducing the notion of an upper semi-continuous density and the main convergence results for the hypo-convergence topology. Section 3.4 provides an example where a sequence of Bayes estimators corresponding to 0-1 loss does not converge to a MAP estimator. In light of the counterexample, some condition is required for MAP estimation to be a limiting case of Bayesian estimation. The informal arguments in the literature give the misleading impression that no condition is needed. We provide a necessary condition in section 3.5, and use it to prove positive results relating MAP and Bayes Estimators.

We conclude this section with a comment on our notation. Preserving the notation in [RW09], we use Greek letters ν and η to denote sequence indices. We use superscript indexing, so that x^ν is an index of the sequence x.

3.2. Bayesian Background

In this section we review necessary Bayesian preliminaries. In point estimation, we wish to infer the value of an unknown parameter θ from an observation x.

A *parametric model* is a family of distributions $p(\cdot|\theta) : \mathscr{X} \to \mathbb{R}$ indexed by a set $\Theta \subseteq \mathbb{R}^n$. The set Θ is referred to as the parameter space, and the measurable space $\mathscr{X}$ is the called the sample space. We assume that each of these distributions is absolutely continuous with respect to Lebesgue measure; hence each $p(\cdot|\theta)$ is a probability density function.

A *Bayesian model* is a parametric model and a distribution π on Θ. The distribution π is called a *prior* on the parameter space Θ. The Bayesian point estimation problem is the following: Given an observation x of a random variable X on $\mathscr{X}$, find $\hat{\theta} \in \Theta$ such that $p(\cdot|\hat{\theta})$ is a "best choice" for the distribution of X among distributions in the parametric model. In this sense, a Bayesian point estimate is a function $\hat{\theta} : \mathscr{X} \to \Theta$ which takes observations to parameter values. Because we only refer to Bayesian point estimates in this chapter, we simply refer to them as estimators.

Given a Bayesian model and an observation x, we define a posterior distribution on the parameter space Θ through Bayes' rule

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_{z\in\Theta} p(x|z)\pi(z)\,dz} \tag{3.1}$$

We assume that for each $x \in \mathscr{X}$, $\int_{z\in\Theta} p(x|z)\pi(z)\,dz$ is finite and nonzero, so that (3.1) defines a density. By taking $\pi(\theta|x) = 0$ outside of Θ, we extend the posterior density so that it is defined on all of $\mathbb{R}^n$ for each x. Hence, without loss of generality, we assume that $\Theta = \mathbb{R}^n$.

A common method of point estimation is through Bayes estimators. Given a loss function $L : \Theta \times \Theta \to [0,\infty)$ which quantifies cost for discrepancies in the true and estimated parameter value, a *Bayes estimator* is an estimator $\hat{\theta}_B$ which minimizes posterior expected loss.

$$\hat{\theta}_B(x) \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}^{z|x}\left[L(\theta,z)\right] = \operatorname*{argmin}_{\theta\in\Theta} \int_{z\in\Theta} L(\theta,z)\pi(z|x)\,dz. \tag{3.2}$$

The flexibility and ubiquity of (3.2) should not be understated. With different choices of loss functions, one can define the posterior mean, median and other quantiles [**Rob07**, 2.5], as well as a variety of robust variants through expected loss minimization [**FHT**, 10.6]. For our purposes, we will focus on one particular family of loss functions, the 0-1 loss functions. The 0-1 loss function L^c is defined for any $c > 0$ as

$$L^c(\theta,z) = \begin{cases} 0 & \text{for } ||\theta - z|| < \frac{1}{c} \\ 1 & \text{otherwise.} \end{cases} \tag{3.3}$$

In the above and what follows, $||\cdot||$ denotes the standard Euclidean norm. Because of the topological nature of the arguments of that follow, any equivalent norm could be used instead. We focus on standard Euclidean for ease of exposition.

The rest of this chapter will deal almost exclusively with the 0-1 loss, so we emphasize our notation: superscript L denotes the 0-1 loss function. We also denote by $\hat{\theta}_B^c$ the Bayes estimator associated with the 0-1 loss function L^c.

Another popular estimation procedure maximizes the posterior density directly. This defines a *maximum a posteriori estimator*, $\hat{\theta}_{MAP}$, which is given by the set of modes of the posterior distribution

$$\hat{\theta}_{MAP}(x) \in \operatorname*{argmax}_{\theta} \pi(\theta|x).$$

This estimator can be interpreted as an analog of maximum likelihood for Bayesian estimation, where the distribution has become a posterior. A number of sources ([**Rob07**, §4.1.2] [**Gew05**, Thm.2.4.3], [**Lee12**, §7.6.5], [**HLD+13**], [**Fig04**]) claim that maximum a posteriori estimators are limits of Bayes estimators, in the following sense. Consider the sequence of 0-1 loss functions, $\{L^\nu : \boldsymbol{R}^n \times \boldsymbol{R}^n \to [0,+\infty)\}_{\nu \in \boldsymbol{N}}$, defined as

$$L^\nu(\theta,z) = \begin{cases} 0 & \text{for } \|\theta - z\| < \frac{1}{\nu} \\ 1 & \text{otherwise} \end{cases}, \tag{3.4}$$

and define $\hat{\theta}_B^\nu$ as the Bayes estimator associated to the loss function L^ν, for each ν, i.e.,

$$\hat{\theta}_B^\nu(x) \in \operatorname*{argmin}_{\theta \in \Theta} \boldsymbol{E}^{z|x}\left[L^\nu(\theta,z)\right].$$

First we translate the Bayes estimator to a maximization problem.

$$\begin{aligned}\hat{\theta}_B^{\nu}(x) &\in \operatorname*{argmin}_{\theta\in\Theta} \int_{z\in\Theta} L^{\nu}(\theta,z)\pi(z|x)\,dz \\ &= \operatorname*{argmin}_{\theta\in\Theta} \left(1 - \int_{\|\theta-z\|<\frac{1}{\nu}} \pi(z|x)\,dz\right) \\ &= \operatorname*{argmax}_{\theta\in\Theta} \int_{\|\theta-z\|<\frac{1}{\nu}} \pi(z|x)\,dz. \end{aligned} \tag{3.5}$$

The claim is that the sequence $\hat{\theta}_B^{\nu}$ converges to $\hat{\theta}_{MAP}$. When the justification is provided, it proceeds as follows. Taking the limit as $\nu \to \infty$, we have

$$\begin{aligned}\lim_{\nu\to\infty} \hat{\theta}_B^{\nu}(x) &\in \lim_{\nu\to\infty} \operatorname*{argmax}_{\theta\in\Theta} \int_{\|\theta-z\|<\frac{1}{\nu}} \pi(z|x)\,dz \\ &= \operatorname*{argmax}_{\theta\in\Theta} \pi(\theta|x) \\ &= \hat{\theta}_{MAP}(x). \end{aligned} \tag{3.6}$$

This justification does not hold in general, and in fact MAP estimators are not necessarily a limiting case of Bayes estimators under the 0-1 loss indicated above. In section 3.4, we exhibit a continuous and unimodal posterior density which is a counterexample to the claim. The problem lies in the limiting argument in (3.6)– the limit of maximizers is not a maximizer without additional conditions, which we establish in section 3.5.

It is worthwhile to note what is correct about the argument in (3.6). Denoting by s_n the volume of the unit ball in $\mathbb{R}^n$, we have that

$$\lim_{\nu\to\infty} s_n \cdot \nu^n \cdot \int_{||\theta-z||<\frac{1}{\nu}} \pi(z|x)\,dz = \pi(\theta|x)$$

Θ-almost everywhere by the Lebesgue differentiation theorem. This gives that a scaled version of $\int_{||\theta-z||<\frac{1}{\nu}} \pi(z|x)\,dz$ converges pointwise a.e. to $\pi(\theta|x)$. But pointwise convergence does not guarantee convergence of maximizers. This will require the notion of hypo-convergence of functions, which we introduce in the next section.

3.3. Convergence of maximizers for Non-Random Functions

3.3.1. General setting. This section summarizes the main results for convergence of optimization problems. The more common approach in the optimization literature is that of minimization, rather than maximization, where the theory of epi-convergence is developed for extended real-valued functions. Here, an adaptation to the maximization setting is presented, which is more natural in our estimation setting.

A function $f : \boldsymbol{R}^n \to \overline{R}$ is said to be *proper* if f is not constantly $-\infty$ and never takes the value ∞. The *effective domain* of the function f is the set

$$\operatorname{dom} f = \{x \in \boldsymbol{R}^n | \, f(x) > -\infty\},$$

and its *hypograph* is the set in $\boldsymbol{R}^{n+1}$

$$\operatorname{hypo} f = \{(x, \lambda) | \, \lambda \leqslant f(x)\}.$$

The function f is called *upper semi-continuous* (usc) if its hypograph is a closed subset of $\boldsymbol{R}^{n+1}$. An equivalent condition is that for each $\alpha \in \mathbb{R}$ the upper level set of f

$$\operatorname{lev}_{\geqslant \alpha} f = \{x \in \boldsymbol{R}^n | \, f(x) \geqslant \alpha\}$$

is closed. A sequential definition of upper semi-continuity can also be stated: for each $x \in \boldsymbol{R}^n$, and each sequence x^ν converging to x,

$$\limsup_{x^\nu \to x} f(x^\nu) \leqslant f(x).$$

We say that a sequence of functions f^ν *hypo-converges* to a function f if both of the following hold for each $x \in \mathscr{X}$.

$$\liminf_\nu f^\nu(x^\nu) \geqslant f(x) \text{ for some } x^\nu \to x$$

$$\limsup_\nu f^\nu(x^\nu) \leqslant f(x) \text{ for every } x^\nu \to x,$$

The notion of hypo-convergence is well-developed because of its importance in proving properties about sequences of optimization problems. An equivalent definition of hypo-convergence follows by identifying each function with its hypograph, and applying the notion of set convergence à-la Painlevé-Kuratowski. Or,

one can characterize hypo-convergence via the Attouch-Wets topology on hypographs. We refer the reader to [**RW09**, Ch.7] for details on the former and [**Bee93**, Ch.3] on the latter.

The following alternative characterization of hypo-convergence will be useful in the sections that follow.

PROPOSITION 3.3.1. [**RW09**, 7.29] *$f^\nu \xrightarrow{hypo} f$ with f usc if and only if the following two conditions hold:*

$$\limsup_\nu \left(\sup_B f^\nu \right) \leqslant \sup_B f \text{ for every compact set } B \subseteq \mathbb{R}^n$$
$$\liminf_\nu \left(\sup_O f^\nu \right) \geqslant \sup_O f \text{ for every open set } O \subseteq \mathbb{R}^n$$

A sequence of functions f^ν is said to be *eventually level-bounded* if for each $\alpha \in \mathbb{R}$ the sequence of upper level sets $\text{lev}_{\geqslant \alpha} f^\nu$ is eventually bounded. That is, there exists a bounded set M, $\alpha \in \mathbb{R}$ and $m \in \mathbb{N}$ such that for all $\nu \geqslant m$, $\text{lev}_{\geqslant \alpha} f^\nu \subseteq M$.

The wide-spread adoption of hypo-convergence in optimization is largely due to the following theorem and its related extensions

THEOREM 3.3.2. [**RW09**, 7.33] *Assume $f^\nu \xrightarrow{hypo} f$, where f^ν is eventually level-bounded and $\{f^\nu, f\}$ are usc and proper. Then*

$$\sup f^\nu \to \sup f$$

and any point which is a limit of maximizers of f^ν maximizes f.

This theorem is attractive because it effectively gives the convergence of maximizers for approximations to optimization problems. In our application to Bayesian point estimation, we wish to establish a similar theorem providing convergence of estimators for approximations to MAP and Bayes estimation. We do so in section 3.5.

3.3.2. Upper Semi-Continuous Densities. Before proceeding, we list the primary assumption that we use throughout this chapter.

ASSUMPTION 3.3.3. For each $x \in \mathscr{X}$, the random vector θ has a density $\pi(\theta|x) : \mathbb{R}^n \to \overline{\mathbb{R}}$ (with respect to Lebesgue measure μ) which is continuous almost everywhere. In other words, for each x there is a set $C \subseteq \mathbb{R}^n$ such that $\pi(\theta|x)$ is continuous at each point in C and $\mu\left(\mathbb{R}^n \backslash C\right) = 0$.

In this framework, we allow a density $\pi(\theta|x)$ to take ∞ as a value, so as not to rule out common densities like the gamma distribution. Obviously a density cannot take ∞ as a value on a set of positive measure.

When dealing with arbitrary measurable functions, the notion of pointwise value is inherently ambiguous because measurable functions are only unique up to alteration on sets of measure zero. The added structure of continuity allows us to define the notion of pointwise value, by referring to the pointwise value of the unique continuous representative of the measurable function. We would like to generalize this to a broader class of functions.

Assumption 3.3.3 provides some minimalist structure for which we can also define an unambiguous notion of pointwise value of a function. This is necessary because without some method of defining a pointwise value of a density, the notion of maximizing the density (as required in MAP estimation) is meaningless.

For each x, take $\pi(\theta|x)$ to be its continuous version on C. On $\mathbb{R}^n \backslash C$, take $\pi(\theta|x)$ to be the upper semi-continuous envelope of $\pi(\theta|x)|_C$. That is, for any $\theta \in \mathbb{R}^n$

$$\pi(\theta|x) = \sup\{\limsup \pi(\theta^\nu|x) : \theta^\nu \to \theta, \theta^\nu \subset C\}. \tag{3.7}$$

The set on the right side of (3.7) is nonempty because the complement of a measure zero set is dense. Furthermore, because $\mathbb{R}^n \backslash C$ has measure zero, the integral of $\pi(\theta|x)$ over $\mathbb{R}^n$ is unchanged by taking the upper semi-continuous envelope. We refer to densities which satisfy assumption 3.3.3 and (3.7) as *upper semi-continuous (usc) densities*. In the remaining sections, densities are assumed to be upper semi-continuous.

Upper semi-continuous densities are a natural class of functions to consider because they provide a large amount of flexibility while still retaining the notion of pointwise value. They contain continuous densities as a special case, but also include intuitive concepts like histograms. Many nonparametric density estimation procedures also produce functions in this family, e.g. [**RW13b**], [**CDSS14**], and kernel density estimates with piecewise continuous kernels.

3.4. Counterexample

This section provides a counterexample to the following claim: Any limit of Bayes estimators $\hat{\theta}_B^{c^\nu}$ having 0-1 loss L^{c^ν} with $c^\nu \to \infty$ is a MAP estimator. First, we provide the Bayesian model. Let $\Theta = \mathbb{R}$ and $\mathscr{X} = \mathbb{R}$. Take $p(x|\theta)$ to be a standard normal distribution. In other words, the parameter θ has no influence

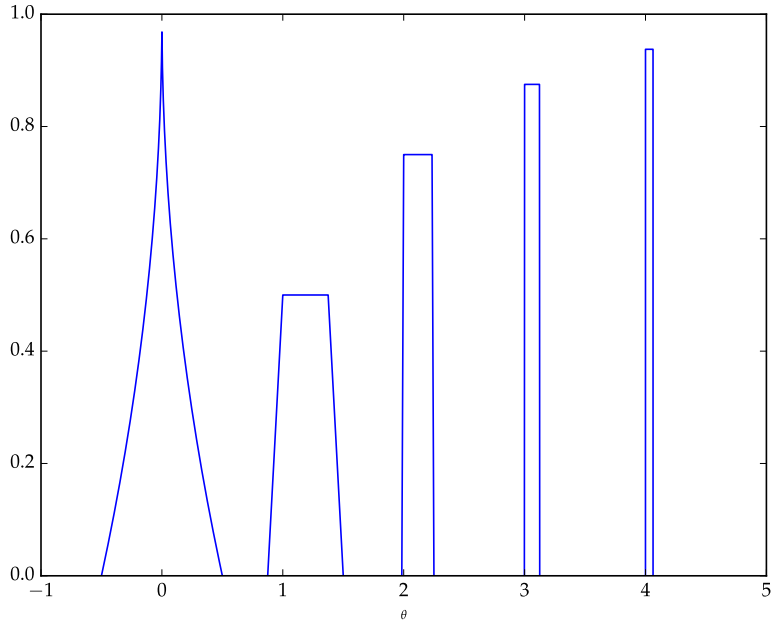


FIGURE 3.1. Posterior Density π

on the distribution of X. While trivial in the fact that θ does not influence X, this example will facilitate computation and illustrate the issue with the argument in (3.6). Consider a prior density $\pi : R \to R$ given by

$$\pi(\theta) = \begin{cases} 1 - \sqrt{|2\theta|} & \theta \in \left(\frac{-1}{2}, \frac{1}{2}\right) \\ (2^n - 1)4^n(\theta + \frac{1}{8^n} - n) & \theta \in \left[n - \frac{1}{8^n}, n\right],\ n \in \mathbb{N}_+ \\ 1 - \frac{1}{2^n} & \theta \in \left[n, n + \frac{1}{2^n} - \frac{1}{8^n}\right],\ n \in \mathbb{N}_+ \\ (1 - 2^n)4^n\left(\theta - (n + \frac{1}{2^n})\right) & \theta \in \left[n + \frac{1}{2^n} - \frac{1}{8^n}, n + \frac{1}{2^n}\right]\ n \in \mathbb{N}_+ \\ 0 & \text{otherwise} \end{cases}.$$

This density is depicted in Figure 3.1. One can easily check that π is indeed a continuous density function. Because θ and X are independent in this Bayesian model, equation (3.1) gives that the posterior is equal to the prior. Thus, for the remainder of this section, we will refer to π as the posterior distribution, and drop the dependence on x in the notation. We also comment that any choice of parametric model where altering θ does not change $p(x|\theta)$ will give rise to the same posterior distribution–the standard normal assumption is just for concreteness.

Note that the posterior density π has a unique maximum, or *posterior mode*, at $\theta = 0$, where it takes the value 1. Therefore, the MAP estimator of this Bayesian model is $\hat{\theta}_{MAP} = 0$. On the other hand, considering the 0-1 loss function L^c, we recall the equivalent definition of the associated Bayes estimator from equation (3.5)

$$\hat{\theta}_B^c \in \operatorname*{argmax}_{\theta} \int_{||z-\theta||<\frac{1}{c}} \pi(z)\, dz. \tag{3.8}$$

We next analyze the limiting behavior of the sequence of Bayes Estimators $\hat{\theta}_B^c$ when $c \to \infty$.

Consider the sequence of scalars $\{c^\nu = 2 \cdot 4^\nu : \nu \in \mathbb{N}_+\}$, and the associated family of loss functions L^{c^ν}. We will prove that for the Bayesian model that we have constructed, $\hat{\theta}_B^{c^\nu} \not\to \hat{\theta}_{MAP}$. This will show that a limit of Bayes estimators with 0-1 loss is not necessarily a MAP estimator. In order to find the maximum in (3.8), we can consider the maximum on each nonzero piece of π. For the non-zero portion around the origin, the integral in (3.8) is obviously maximized at 0. Furthermore, for each "bump" of π, the integral is maximized (perhaps non-uniquely) at the midpoint of the interval where the density is constant. In order to show that $\hat{\theta}_B^{c^\nu} \not\to \hat{\theta}_{MAP}$, it suffices to show that for each ν there is a $\theta \notin (-1/2, 1/2)$ such that the evaluation of (3.8) at θ is strictly greater than the evaluation at zero. This gives that $\hat{\theta}_B^{c^\nu}$ cannot have a limit point in $(-1/2, 1/2)$, and hence cannot have 0, the MAP estimator, as a limit point. We now perform the required calculations.

i. Evaluation of (3.8) at 0.

$$\int_{|z|<\frac{1}{2\cdot 4^\nu}} 1 - \sqrt{|2z|}\, dz = 2\int_0^{\frac{1}{2\cdot 4^\nu}} 1 - \sqrt{2z}\, dz = \frac{1}{4^\nu} - \frac{4\sqrt{2}}{3\cdot 2^{\frac{3}{2}}} \cdot \frac{1}{8^\nu} \tag{3.9}$$

ii. For each $\nu \in \mathbb{N}$, evaluating (3.8) at $\theta = 2\nu + \frac{1}{2^{2\nu+1}}$ gives

$$\int_{\left|z-\left(2\nu+\frac{1}{2^{2\nu+1}}\right)\right|<\frac{1}{2\cdot 4^\nu}} \pi(z)\, dz = \int_{[2^\nu, 2^\nu+\frac{1}{2^{2\nu}}-\frac{1}{8^{2\nu}}]} \pi(z)\, dz + \int_{[2^\nu+\frac{1}{2^\nu}-\frac{1}{8^{2\nu}}, 2^\nu+\frac{1}{2^{2\nu}}]} \pi(z)\, dz,$$

The first part of this sum is an integral over a constant piece of π. The second is a linear piece of π. Bounding the sum of the integrals below by the value of just the integral over the constant piece, we have that

$$\int_{\left|z-\left(2\nu+\frac{1}{2^{2\nu+1}}\right)\right|<\frac{1}{2\cdot 4^\nu}} \pi(z)\, dz \geqslant \left(1 - \frac{1}{2^{2\nu}}\right)\left(\frac{1}{2^{2\nu}} - \frac{1}{8^{2\nu}}\right) = \frac{1}{4^\nu} - \left(2 - \frac{1}{16^\nu}\right)\frac{1}{16^\nu}. \tag{3.10}$$

Since (3.10) is strictly greater than (3.8) for all $\nu \geqslant 1$, we conclude that $\hat{\theta}_B^{c^\nu} \notin (\frac{-1}{2}, \frac{1}{2})$ for all $\nu \geqslant 1$. Hence the sequence of Bayes estimators $\hat{\theta}_B^{c^\nu}$ has $c^\nu \to \infty$, but does not have a MAP estimator as a limit point. This concludes the counterexample and shows that the claim is false.

3.5. Convergence results

In this section we provide conditions on posterior distributions which guarantee that a sequence of Bayes Estimators with 0-1 loss has a MAP estimator as a limit point. In addition, we provide a partial converse by showing that each MAP estimator is a limit of approximate Bayes estimators.

Define an estimator $\hat{\theta}$ of θ to be an *ε-approximate Bayes estimator* with respect to loss L if for each $x \in \mathscr{X}$.

$$\int_{z\in\Theta} L(\hat{\theta}(x), z)\pi(z|x)\,dz \geqslant \sup_{\theta\in\Theta} \int_{z\in\Theta} L(\theta, z)\pi(z|x)\,dz - \varepsilon(x).$$

Here, ε is a function from $\mathscr{X}$ to $\mathbb{R}$. We say that $\hat{\theta}$ is a *limit of approximate Bayes estimators* if there are sequences V^ν, $\hat{\theta}^\nu$ and ε^ν such that $\hat{\theta}^\nu$ converges almost surely to $\hat{\theta}$, $\hat{\theta}^\nu$ is an ε^ν-approximate Bayes estimator with respect to loss V^ν for each ν, and ε^ν converges $\mathscr{X}$-almost surely to 0.

As discussed in the introduction, we assume that all densities are upper semi-continuous and satisfy assumption 3.3.3.

We begin with a deterministic result. The proof of the next lemma, which is related to the epi-convergence of mollifying approximates from [**RW09**], is included in the Appendix 3.6.

LEMMA 3.5.1. *Assume $f : \Theta \to \overline{\mathbb{R}}$ is an upper semi-continuous density. Let s_n denote the volume of the unit ball in $\mathbb{R}^n$. Define*

$$f^\nu(\theta) := \nu^n \cdot s_n \cdot \int_{||\theta - z|| < \frac{1}{\nu}} f(z)dz.$$

Then f^ν hypo-converges to f.

We note that even though upper semi-continuous densities may take ∞ as a value, and hence need not be proper, f^ν must be proper because f integrates to one.

This lemma allows us to prove the following result.

THEOREM 3.5.2. *If $\pi(\theta|x)$ is proper $\mathscr{X}$ almost surely, then any MAP estimator is a limit of approximate Bayes estimators.*

PROOF. Let $\hat{\theta}_{MAP}$ be a maximum a posteriori estimator of θ. By definition, $\hat{\theta}_{MAP}(x) \in \operatorname{argmax} \pi(\theta|x)$. Let x be any element in $\mathscr{X}$. For ease of notation, we will drop the explicit dependence on x by writing $\hat{\theta}_{MAP} := \hat{\theta}_{MAP}(x)$ and letting $f(\theta) := \pi(\theta|x)$. By lemma 3.5.1, $f^\nu \xrightarrow{hypo} f$. From the definition of hypo-convergence, there is a sequence $\theta^\nu(x) \to \hat{\theta}_{MAP}(x)$ such that $\limsup_\nu f^\nu(\theta^\nu) \leqslant f(\hat{\theta}_{MAP})$. Also directly from the definition, $\liminf_\nu f^\nu(\theta^\nu) \geqslant f(\hat{\theta}_{MAP})$. Hence $\lim_\nu f^\nu(\theta^\nu) = f(\hat{\theta}_{MAP})$.

Repeating this construction pointwise for each $x \in \mathscr{X}$, we define a sequence of estimators with $\hat{\theta}^\nu(x)$ the θ^ν sequence above. Define $\varepsilon^\nu(x)$ as $\sup_{\theta\in\Theta} f^\nu(\theta) - f^\nu(\theta^\nu)$. We claim that L^ν, $\hat{\theta}^\nu$, and ε^ν satisfy the conditions so that $\hat{\theta}_{MAP}$ is an approximate Bayes estimator. We must verify the three conditions in the definition. The first two, that $\hat{\theta}^\nu$ converges almost surely to $\hat{\theta}_{MAP}$ and that $\hat{\theta}^\nu$ is an ε^ν-approximate Bayes estimator, are true by construction. Lastly, we must show that $\varepsilon^\nu \to 0$ almost surely. By monotonicity of the integral, we know that

$$f^\nu(\theta^\nu) \leqslant \sup_\theta f^\nu(\theta) \leqslant f(\hat{\theta}_{MAP}).$$

For each $x \in \mathscr{X}$ with $\pi(\theta|x)$ proper, we combine this inequality with the fact that $\lim_\nu f^\nu(\theta^\nu) = f(\hat{\theta}_{MAP})$ to arrive at the following

$$0 \leqslant \varepsilon^\nu(x) = \sup f^\nu - f^\nu(\theta^\nu) \leqslant f(\hat{\theta}_{MAP}) - f^\nu(\theta^\nu) \to 0$$

Note that $\pi(\theta|x)$ must be proper for this to hold: it guarantees that $f(\hat{\theta}_{MAP}) < \infty$, and hence $f(\hat{\theta}_{MAP}) = \lim_\nu f^\nu(\theta^\nu)$ gives $f(\hat{\theta}_{MAP}) - \lim_\nu f^\nu(\theta^\nu) = 0$.

We conclude that $\varepsilon^\nu(x) \to 0$ almost everywhere, since $\pi(\theta|x)$ is proper $\mathscr{X}$ almost everywhere. This shows the third condition in the definition, so that $\hat{\theta}_{MAP}$ is a limit of approximate Bayes estimators. □ □

The notion of an approximate Bayes estimator captures the idea that Bayes and MAP estimators are close in terms of evaluation of the objective. They need not be close in distance to each other in Θ. This point is subtle, and underlies the confusion in the incorrect claim in (3.6). In fact, as the counterexample in the previous section shows, Theorem 3.5.2 cannot be strengthened from approximate to true Bayes estimators without additional conditions.

We turn now to providing those conditions, in a sort of converse to Theorem 3.5.2. We turn our focus to characterizing when a sequence of Bayes estimators converges to a MAP estimator. We again begin with a deterministic result.

THEOREM 3.5.3. *Assume there exists α such that* $\text{lev}_{\geqslant\alpha} f$ *is bounded and has nonempty interior. Then* $\operatorname{argmax} f^\nu$ *is eventually nonempty, and any sequence maximizers of f^ν as $\nu \to \infty$ has a maximizer of f as a limit point.*

PROOF. Let f be an usc density. Assume there is an $\alpha \in \mathbb{R}$ such that the upper level set $\text{lev}_{\geqslant\alpha} f$ is bounded and has nonempty interior. Note that $\text{lev}_{\geqslant\alpha}$ is compact by the Heine-Borel Theorem because upper level sets are closed when f is upper semicontinuous.

First, we show that there is an index ν_0 such that $\operatorname{argmax} f^\nu$ is nonempty for each $\nu > \nu_0$. Then we show that any sequence of maximizers of f^ν, has as a limit point a maximizer of f.

To show the existence of a maximizer, we first show that f^ν is upper semi-continuous with a bounded, nonempty upper level set. That f^ν attains its maximum then follows because an upper semi-continuous function attains its maximum on a compact set. We recall from section 3.3 that our framework allows that this maximum could be ∞.

Fix $\nu \in \mathbb{N}$ and $\theta \in \mathbb{R}^n$. Let $\{\theta^\eta\}_{\eta\in\mathbb{N}}$ be any sequence that converges to θ. To show upper semicontinuity we must have

$$\limsup_\eta f^\nu(\theta^\eta) \leqslant f^\nu(\theta),$$

but this follows from the following chain of inequalities

$$\begin{aligned}
\limsup_\eta f^\nu(\theta^\eta) &= \limsup_\eta s_n \cdot \nu^n \cdot \int_{||\theta^\eta - z|| < \frac{1}{\nu}} f(z)\,dz \\
&= \limsup_\eta s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} f(z+\theta^\eta)\,dz \\
&\leqslant s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} \limsup_\eta f(z+\theta^\eta)\,dz \\
&\leqslant s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} f(z+\theta)\,dz \\
&= f^\nu(\theta).
\end{aligned}$$

The first inequality follows from Fatou's lemma, and the second from the upper semicontinuity of f. Hence f^ν is upper semicontinuous.

The family of functions f^ν has a bounded upper level set because $\text{lev}_{\geqslant\alpha} f$ bounded with constant M implies that $\text{lev}_{\geqslant\alpha} f^\nu$ is bounded with constant $M + \frac{1}{\nu}$. Lastly, we show that the upper level set $\text{lev}_{\geqslant\alpha} f^\nu$ is

nonempty. Let $\theta \in \operatorname{int} \operatorname{lev}_{\geqslant \alpha} f^\nu$, and denote by δ radius such that $\mathbb{B}(\theta, \delta) \subseteq \operatorname{int} \operatorname{lev}_{\geqslant \alpha} f^\nu$. Choose $\nu_0 \geqslant \frac{1}{\delta}$. Then for any $\nu > \nu_0$

$$f^\nu(\theta) = s_n \cdot \nu^n \int_{||\theta - z|| < \frac{1}{\nu}} f(z)\, dz \geqslant s_n \cdot \nu^n \cdot \alpha \cdot \frac{1}{\nu^n} \cdot s_n = \alpha.$$

So $\theta \in \operatorname{lev}_{\geqslant \alpha} f^\nu$. We conclude that this level set is nonempty and bounded, so that f^ν attains its maximum.

Now let $\hat{\theta}^\nu$ be any sequence of maximizers of f^ν. For $\nu > \nu_0$, $\hat{\theta}^\nu \in \operatorname{lev}_{\geqslant \alpha} f^\nu$. From Lemma 3.5.1 and Proposition 3.3.1

$$\liminf_\nu f^\nu(\hat{\theta}_B^\nu) = \liminf_\nu \sup_{\mathbb{R}^n} f^\nu \geqslant \sup_{\mathbb{R}^n} f^\nu$$

and

$$\begin{aligned}
&\limsup_\nu f^\nu(\hat{\theta}^\nu) \\
&= \limsup_\nu (\sup_{\mathbb{R}^n} f^\nu) \\
&= \limsup_\nu \left(\sup_{\overline{\mathbb{B}}(0, M+1)} f^\nu \right) \\
&\leqslant \sup_{\overline{\mathbb{B}}(0, M+1)} f \\
&= \sup_{\mathbb{R}^n} f.
\end{aligned}$$

So $\lim f^\nu(\hat{\theta}_B^\nu) = \sup_{\mathbb{R}^n} f$.

Since $\hat{\theta}^\nu$ is eventually in $\overline{\mathbb{B}}(0, M+1)$, which is compact, it has a convergent subsequence $\hat{\theta}^{\nu_k} \to \tilde{\theta}$. By upper semi-continuity

$$f(\tilde{\theta}) \geqslant \limsup f(\hat{\theta}^{\nu_k}) = \lim f(\hat{\theta}^\nu).$$

Hence $\tilde{\theta}$ maximizes f. Furthermore, by upper semi-continuity any limit point of $\hat{\theta}^\nu$ maximizes f. This proves the theorem. □ □

We now prove our main result about the relationship between estimators.

THEOREM 3.5.4. *Suppose that for each $x \in \mathcal{X}$, $\pi(\theta|x)$ satisfies the following property: There exists α such that*

$$\{\theta : \pi(\theta|x) > \alpha\}$$

is bounded and has nonempty interior. Then for any of sequence of Bayes estimators $\hat{\theta}_B^\nu$ with 0-1 *loss L^ν*

(i) There is a MAP estimator $\hat{\theta}_{MAP}$ such that $\hat{\theta}_{MAP}(x)$ is a limit point of $\hat{\theta}_B^\nu(x)$ for each x.

(ii) Every limit point of $\hat{\theta}_B^\nu(x)$ defines a MAP estimator, in the sense that there is a MAP estimator $\hat{\theta}_{MAP}$ with $\hat{\theta}_{MAP}(x)$ equal to that limit point.

PROOF. Fix $x \in \mathscr{X}$. By assumption, $\pi(\theta|x)$ has a bounded level set with nonempty interior. From the deterministic result, Theorem 3.5.3, the sequence $\hat{\theta}_B^\nu(x)$ has a limit point $\tilde{\theta}(x)$, and this limit point maximizes $\pi(\theta|x)$. Define a MAP estimator $\hat{\theta}_{MAP}(x) := \tilde{\theta}(x)$ pointwise for each $x \in \mathscr{X}$. This proves (i).

For (ii), the result follows since the method of defining a MAP estimator in the previous paragraph is valid for every limit point of $\hat{\theta}_B^\nu(x)$. □ □

As a consequence of the proof, we note the following: if $\hat{\theta}_B^\nu$ converges to some estimator $\hat{\theta}$ almost surely, then $\hat{\theta}$ is a MAP estimator.

We next establish related results for various shape constrained densities. Recall that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is quasiconcave if for each $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \geqslant \min\{f(x), f(y)\}.$$

For a quasiconcave function, local maximality implies global maximality. This shape constraint captures the notion of a probability distribution having a "single peak". It is also sometimes referred to as unimodality, but we avoid this terminology because it has multiple meanings in the literature.

THEOREM 3.5.5. *Assume that for each $x \in \mathscr{X}$, $\pi(\theta|x)$ is quasiconcave. Then the conclusion of Theorem 3.5.4 holds.*

PROOF. For each x, since $\int \pi(z|x)\,dz = 1$, there is an $\alpha > 0$ such that $\mu(\{\theta : \pi(\theta|x) \geqslant \alpha\})$ has positive measure. By quasiconvexity, $\text{lev}_{\geqslant\alpha}\,\pi(\theta|x) = \{\theta|\pi(\theta|x) \geqslant \alpha\}$ is convex.

Note $\text{lev}_{\geqslant\alpha}$ having an interior point is equivalent to the set containing n affinely independent points. Suppose, towards a contradiction, that $\text{lev}_{\geqslant\alpha}\,\pi(\theta|x)$ does not contain any interior points, then its affine hull lies in an $n-1$ dimensional plane. This contradicts the set having positive measure. Hence $\{\theta : \pi(\theta|x)\}$ must have an interior point.

In order to apply Theorem 3.5.4 we must also show that the level set $\text{lev}_{\geqslant\alpha}\,\pi(\theta|x)$ is bounded. Fix B to be any n-dimensional ball in $\text{lev}_{\geqslant\alpha}\,\pi(\theta|x)$. If θ^ν were a sequence in $\text{lev}_{\geqslant\alpha}\,\pi(\theta|x)$ such that $||\theta^\nu|| \to \infty$,

then

$$\int_{z\in\Theta} \pi(z|x)\,dz \geqslant \int_{z\in \text{lev}_{\geqslant\alpha}\,\pi(\theta|x)} \pi(z|x)\,dz \geqslant \mu(\{\text{conv}\,(B\cup\{\theta^v\})\})\cdot\alpha$$

Here, μ denotes Lebesgue measure. One can easily show that $\mu(\text{conv}\,(B\cup\{\theta^v\}))\to\infty$ when $\theta^v\to\infty$. This contradicts that $\int\pi(z|x)dz = 1$. Hence there cannot exist such a sequence, so the level set is bounded and the result is proven. □ □

The following corollary about log-concave densities follows immediately. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is log-concave if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) \geqslant f(x)^\lambda f(y)^{1-\lambda}$$

Log-concave densities have appeared in recent work of [**Ruf07**], [**DR+09**] due to their attractive computational and theoretical properties in nonparametric density estimation.

COROLLARY 3.5.6. *Assume that for each $x\in\mathscr{X}$, $\pi(\theta|x)$ is log-concave. Then the conclusion of theorem 3.5.4 holds.*

PROOF. Log-concavity implies quasiconcavity, and the result follows immediately from the previous theorem. □ □

3.6. Proof of Lemma 1

In this appendix we provide the proof of Lemma 3.5.1.

PROOF. Let $\theta\in\mathbb{R}^n$. To show hypo-convergence, we must show that for each sequence $\theta^v\to\theta$, $\limsup_v f^v(\theta^v)\leqslant f(x)$ and that there exists a sequence $\theta^v\to\theta$ with $\liminf_v f^v(\theta^v)\geqslant f(\theta)$.

Fix $\varepsilon>0$. Since f is upper semi-continuous at θ, there is a $\delta>0$ such that $||z-\theta||<2\delta$ gives $f(z)-f(\theta)<\varepsilon$.

Consider any sequence $\theta^v\to\theta$. We have that

$$f^v(\theta^v)-f(\theta) = s_n\cdot v^n\cdot\int_{||\theta^v-z||>\frac{1}{v}}(f(z)-f(\theta))dz = s_n\cdot v^n\cdot\int_{||z||<\frac{1}{v}}(f(z+\theta^v)-f(\theta))\,dz.$$

Choose $\nu_0 \in \mathbb{N}$ so that $||\theta - \theta^\nu|| < \delta$ and $\frac{1}{\nu} < \delta$ for each $\nu > \nu_0$. Then for any $\nu > \nu_0$,

$$s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} (f(z+\theta^\nu) - f(\theta))\, dz \leqslant s_n \cdot \nu^n \cdot \varepsilon \cdot \int_{||z|| < \frac{1}{\nu}} dz = \varepsilon.$$

Thus $\limsup_\nu f^\nu(\theta^\nu) \leqslant f(\theta)$.

To establish the second part of the hypo-convergence definition, we focus our attention on constructing a sequence that satisfies the required inequality.

Consider any $\eta \in \mathbb{N}$. Recall that f^ν is an upper semi-continuous density. Let C be the set where f is continuous. Because C is dense, for each $\nu \in \mathbb{N}$, there is a $y^\nu \in C$ such that $||y^\nu - x|| < \frac{1}{\nu}$. Furthermore, $y^\nu \in C$ means that there is a $\delta(y^\nu, \eta) > 0$ such that any $z \in \Theta$ which satisfies $||y^\nu - z|| < \delta(y^\nu, \eta)$ also has

$$|f(y^\nu) - f(z)| < \frac{1}{\eta}.$$

Here we use function notation for δ to emphasize that δ depends on both y^ν and η.

For each η, define a sequence such that

$$z^{\nu,\eta} = \begin{cases} 0 & \text{when } \frac{1}{\nu} > \delta(y^1, \eta) \\ y^1 & \text{when } \delta(y^2, \eta) \leqslant \frac{1}{\nu} < \delta(y^1, \eta) \\ y^2 & \text{when } \delta(y^3, \eta) \leqslant \frac{1}{\nu} < \min\{\delta(y^2, \eta), \delta(y^1, \eta)\} \\ y^3 & \text{when } \delta(y^4, \eta) \leqslant \frac{1}{\nu} < \min_{i \leqslant 4} \delta(y^i, \eta) \\ \vdots & \vdots \end{cases}$$

Extracting a diagonal subsequence from the sequences generated according to this procedure gives a sequence θ^ν such that $\theta^\nu \to \theta$ and $\frac{1}{\nu} < \delta(\theta^\nu, \nu)$. In particular, $|f(\theta^\nu) - f(z)| < \frac{1}{\nu}$ for z with $||\theta^\nu - z|| < \frac{1}{\nu}$.

Hence, for any $\varepsilon > 0$, choosing $\nu > \frac{2}{\varepsilon}$ gives

$$\begin{aligned} |f^\nu(\theta^\nu) - f(\theta)| &\leqslant |f^\nu(\theta^\nu) - f(\theta^\nu)| + |f(\theta^\nu) - f(\theta)| \\ &\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

We conclude that $\lim_\nu f^\nu(\theta^\nu) = f(\theta)$, so the result is proven. □

□

CHAPTER 4

Log-Concave Duality in Estimation and Control

4.1. Introduction

We consider the problem of estimating the state of a noisy dynamical system based only on noisy measurements of the system. In this chapter, we assume linear dynamics, so that the progression of state variables is

$$\mathbf{X}_{t+1} = F_t(\mathbf{X}_t) + \mathbf{W}_{t+1}, \quad t = 0, ..., T-1 \tag{4.1}$$

$$\mathbf{X}_0 = \mathbf{W}_0 \tag{4.2}$$

$\mathbf{X}_t$ is the state variable–a random vector indexed by a discrete time-step t which ranges from 0 to some final time T. All of the results in this chapter still hold in the case that the dimension of $\mathbf{X}_t$ is time-dependent, but for notational convenience we will assume that $\mathbf{X}_t \in \mathbb{R}^{n_x}$ for $t = 0, ..., T$. F_t is then a $n_x \times n_x$ real-valued matrix that, though it may vary with time, is known a priori. The $\mathbf{W}_t$ term is a random vector in $\mathbb{R}^{n_x}$ that represents noise in the system dynamics. Note that, in this formulation, the random vectors $\mathbf{W}_t$ are *primitive*, in the sense that they generate all of the randomness associated with the problem. The state variables $\mathbf{X}_t$ are secondary, being derived from applying dynamic equations to $\mathbf{W}_t$ terms.

In addition to the dynamics that govern the state progression, we also have a measurement process which dictates the observable information at time t. We assume that the measurement process is linear.

$$\mathbf{Z}_t = H_t(\mathbf{X}_t) + \mathbf{V}_t \tag{4.3}$$

The vector $\mathbf{Z}_t$ is a (secondary) random vector of dimension n_z. Again, we can consider the case that the dimension of $\mathbf{Z}_t$ changes with time, but for notational convenience we will assume that the measurements

This chapter is based on joint work with Michael Casey and Roger J-B Wets [Wal09].

have a fixed dimension. This gives that H_t is an $n_z \times n_x$ matrix, which similar to F_t may vary with time but is known in advance. $\mathbf{V}_t$ is a primitive random vector of dimension $\mathbb{R}^{n_z}$ that represents measurement noise.

Different information structures in this setup correspond to different types of estimation problem. In this chapter we consider the smoothing problem, which consists of estimating of $\mathbf{X_0}, ..., \mathbf{X}_T$ after all measurement variables $\mathbf{Z}_0, ...\mathbf{Z}_T$ have been observed. In this sense, the information associated with the problem is constant–the set of measurements which we use to estimate $\mathbf{X}_0$ is the same as the measurements with which we estimate $\mathbf{X}_T$. This differs from the filtering problem, which is one of sequential state estimation. In the filtering problem, the set of available measurements depends on the time of the state being estimated. Of course, the difference between these problems can be formulated in terms of measurability with respect to certain filtrations, but we avoid this language because we show momentarily that the main problem of interest is deterministic.

Kalman, in his seminal paper [**Kal60**], assumed that there was no measurement noise associated with the system, so that $\mathbf{V}_t \equiv 0$. Motivated by minimizing mean-squared error, he sought to find the conditional expectation of the states given the measurement. Under the assumption that the dynamic and measurement noise are Gaussian, conditional expectation reduces to a deterministic maximum a posteriori (MAP) problem, in which the optimal estimate is the mode of the conditional density

$$f_X(\mathbf{X}_0, ..., \mathbf{X}_T | (\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T))$$

Assuming that $\mathbf{W}_t \sim \mathcal{N}(0, Q_t)$ and $\mathbf{V}_t \equiv 0$, the problem can be derived explicitly, as in [**Cox64**]. We are left with the following optimization problem:

$$(\mathscr{P}_{Kal}) \qquad \min_{x_0,...,x_T} \sum_{t=0}^{T} \frac{1}{2} w_t' Q_t^{-1} w_t$$

$$\text{subject to } x_{t+1} = F_t x_t + w_{t+1}, \ \ t = 0, ..., T-1$$

$$x_0 = w_0$$

$$z_t = H_t x_t, \ \ t = 0, ..., T$$

In this formulation, the variables x_t and w_t are estimates of the random variables $\mathbf{X}_t$ and $\mathbf{W}_t$, respectively. In this sense, w_t is also a decision variable, but we prefer to write the problem in this reduced formulation where a decision $x_0, \ldots, x_T$ generates the variables $w_0, \ldots, w_T$.

Kalman observed that the Linear Quadratic Regulator problem

$$(D_{Kal}) \qquad \min_{u_0,\ldots,u_T} \sum_{t=0}^{T} \frac{1}{2} y_t' Q_t y_t$$

$$\text{subject to } y_{t+1} = F_t' y_t + H_t' u_{t+1}$$

$$y_0 = H_0' u_0$$

over controls $\{u_t\}_{t=0}^T$ and states $\{y_t\}_{t=0}^T$, is *dual* to the estimation problem above. Kalman defined this duality in terms of the equations that characterize their solutions: the algebraic Riccati equation which characterizes the value function of (D_{Kal}) is the same equation that governs the propagation of the variance of the estimate in ($\mathscr{P}_{Kal}$), with a time reversal. Since the Linear Quadratic Regulator problem is one of optimal control, the relationship between the problems is described as duality between estimation and control, in the Linear-Quadratic Gaussian setting.

Compared to the equation-correspondence duality which is typical in the engineering literature [**Tod08**], [**Dav77**] , we take a different approach by using the duality theory of convex programming. The allows us to extend the duality of estimation and control to the more general setting where noise and measurement noise have *log-concave* densities. This includes the Linear-Quadratic Gaussian framework, but by viewing the duality in a convex-analytic framework we gain more insight into the relationship between estimation and control. Previous literature has focused almost exclusively on either equation-correspondence duality or convex analytic duality between estimation and control problems. We will focus this work on duality in the convex-analytic sense.

The rest of this chapter is organized as follows. In section 2 we state and prove the main result of the chapter: a duality result between estimation and optimal control when the noise terms in (4.3), (4.1) are log-concave. Section 3 applies this result to the case where the noise terms have densities which are exponentiated monitoring functions, so that no constraint qualification is required for strong duality. Section 4 contains a practical example, where the solution to the optimal control problem is used to generate an optimal state estimate.

We conclude this section by establishing some definitions and notations that we will use throughout the rest of this chapter. Recall that a function is called *lower semi-continuous* (lsc) if for every x in its domain,

$$\liminf_{x^v \to x} f(x^v) \geqslant f(x)$$

for every sequence $x^v \to x$. In addition, recall that a convex function which takes extended real-values is called *proper* if it is not identically ∞ and never takes the value $-\infty$ [**RWW10**]. In keeping with convex-analytic literature, we refer to the domain of an extended-valued convex function as the set where it assumes a finite value.

Lastly, given a convex function $f : \mathbb{R}^n \to (-\infty, \infty]$, we denote by f^* the convex conjugate of f, which is defined as

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y \rangle - f(x)\}.$$

Conjugation is ubiquitous in convex-analytic duality theory, and this chaper is no exception. For more details and background on conjugation and its relation to duality, the reader should consult any of [**Roc74**] [**Roc97**] [**RWW10**].

4.2. Estimation with Convex Penalties

In this section we consider the case where the the random vectors $\mathbf{W}_t$ and $\mathbf{V}_t$ in (4.1) and (4.3) have log-concave density functions. Recall that a function $\phi : \mathbb{R}^d \to \mathbb{R}$ is *log-concave* if

$$\phi(x) = \exp -f(x)$$

where $f : \mathbb{R}^d \to (-\infty, \infty]$ is a convex function. By convention, we adopt that $e^{-\infty} = 0$.

The collection of random vectors with log-concave densities is broad enough to include many commonly used distributions, such as the normal, Laplace, and exponential [**BB05**]. Moreover, it is closed with respect to taking marginals, convolutions, and forming product measures [**DJD88**]. These characteristics make MAP estimation in the presence of log-concave noise much more amenable to computation than the more general unimodal class, because they guarantee that conditional expectations, sums of random variables, and joint densities formed by independent log-concave random variables remain log-concave. These are exactly the operations performed when considering MAP estimation in the presence of linear dynamics

and measurements. The broader class of unimodal distributions, on the other hand, does not enjoy these properties, making them much more difficult to work with in the context of discrete-time state estimation.

Nonparametric density estimation within the class of log-concave random vectors also has attractive theoretical and computational properties. We will not review those results here (see, for example, [**DR**, **PWM07**, **Ruf**]), but we do comment that the results in the later sections find rich applications in nonparametric log-concave density estimation, particularly because of their non-smooth nature.

The maximum a posteriori estimation of the states, given measurements $z_0, ..., z_T$ can be derived similarly to the Gaussian case.

PROPOSITION 4.2.1. *Assume that $\mathbf{W}_t$ and $\mathbf{V}_t$ are independent and have log-concave densities e^{-f_t} and e^{-g_t}, respectively, for $t = 0, ..., T$. Then the maximum a posterior estimate of the states $\mathbf{X}_0, ..., \mathbf{X}_T$ given $(\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T)$ is given by the solution to the problem*

$$(\mathscr{P}) \qquad \min_{x_0,...,x_T} \sum_{t=0}^{T} f_t(w_t) + \sum_{t=0}^{T} g_t(z_t - H_t x_t)$$

$$\text{subject to } x_{t+1} = F_t x_t + w_{t+1}, \ \ t = 0, ..., T-1$$

$$x_0 = w_0$$

Equivalently, one can use an extended formulation, minimizing over w_t and x_t, or simply minimizing in w_t.

PROOF. In maximum a posteriori estimation, we seek to maximize the density

$$p_{\mathbf{X}}(\mathbf{X}_0, ..., \mathbf{X}_T | (\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_n)).$$

By Bayes' Theorem

$$p_{\mathbf{X}}((\mathbf{X}_0, ..., \mathbf{X}_T) = (x_0, ..., x_T) | (\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T))$$
$$= \frac{p_{\mathbf{Z}}((\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T) | (\mathbf{X}_0, ..., \mathbf{X}_T) = (x_0, ..., x_T)) p_{\mathbf{X}}(\mathbf{X}_0, ..., \mathbf{X}_T) = (x_0, ..., x_T))}{p_{\mathbf{Z}}((\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T))}.$$

By the independence of measurement noise,

$$p_{\mathbf{Z}}((\mathbf{Z}_0, ..., \mathbf{Z}_T) = (z_0, ..., z_T) | (\mathbf{X}_0, ..., \mathbf{X}_T) = (x_0, ..., x_T)) = \prod_{t=0}^{T} p_{\mathbf{V}_t}(z_t - H_t x_t).$$

Furthermore, since the process is Markov (by independence of dynamic noise)

$$p_{\mathbf{X}}((\mathbf{X}_0,...,\mathbf{X}_T)=(x_0,...,x_T))=p_{\mathbf{X}_0}(x_0)\cdot p_{\mathbf{X}_1}(x_1|x_0)\cdot...\cdot p_{\mathbf{X}_T}(x_T|x_{T-1}).$$

Our posterior becomes

$$\frac{\prod_{t=0}^{T}p_{\mathbf{V}_t}(z_t-H_tx_t)\cdot p_{\mathbf{X}_0}(x_0)\cdot\prod_{t=1}^{T}p_{\mathbf{X}_t}(x_t|x_{t-1})}{p_{\mathbf{Z}}((\mathbf{Z}_0,...,\mathbf{Z}_T)=(z_0,...,z_T))}.$$

By the assumptions on the distributions of $\mathbf{V}_t$ and $\mathbf{W}_t$, this is

$$C(z_0,...,z_T)\cdot\exp\left\{-f_0(x_0)-\sum_{t=0}^{T-1}f_{t+1}(x_{t+1}-F_t(x_t))-\sum_{t=0}^{T}g_t(z_t-H_tx_t)\right\}$$

where $C(z_0,...,z_n)$ is some term not depending on $(x_0,...,x_T)$. Maximizing this expression in $(x_0,...,x_T)$ is then equivalent to minimizing

$$f_0(x_0)+\sum_{t=0}^{t-1}f_{t+1}(x_{t+1}-F_t(x_t))+\sum_{t=0}^{T}g_t(z_t-H_t(x_t)).$$

This gives us the problem in the statement of the proposition. The different formulations follow because each choice of $(x_0,,,x_T)$ generates a unique $(w_0,...,w_T)$, according to the dynamics, and vice versa. □

The extension from the Gaussian noise to log-concave random vectors is signficant. The fact that the functions f_t and g_t in $\mathscr{P}$ can take the value ∞ permits a choice of densities which do not have full support. Correspondingly, the MAP problem then becomes one of traditional convex optimization [**Roc97**], [**RWW10**], where constraints are built in to the objective function by allowing that function to take infinite values.

The next lemma provides information about the function f used to define a log-concave density.

LEMMA 4.2.2. *Assume that $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a convex function which defines the density of a random variable $\mathbf{X}\sim e^{-f(x)}$. Then*

(a) If $\mathrm{cl}(f)$ is the lower-semicontinuous hull of f, then $e^{\mathrm{cl}(f)}$ is also a density function for $\mathbf{X}$.

(b) f is proper

(c) $dom(f)$ is full-dimensional

(d) f is level-bounded, so that the minimum of f over $\mathbb{R}^n$ is attained.

PROOF. First we prove (a). Since a convex function is continuous on the interior of its domain [**Roc97**][10.1], the only points where f may fail to be lower semicontinuous is on the boundary of its domain. The domain of a convex function is obviously convex, and since the boundary of a convex set has Lebesgue measure zero [**Jar15**][Lemma 1.8.1], $\mathrm{cl}(f)$ and f are equal almost everywhere. Hence $e^{\mathrm{cl}(f)}$ is also a density function for X, since it differs from the given density on a set of measure zero. This results allows us to refer to pointwise values of f, by which we mean the values of the unique lower-semicontinuous extension $\mathrm{cl}(f)$.

(b) follows from (a) and the fact that $\int e^{-f(x)}dx = 1$. Because an improper lower-semicontinuous convex function can have no finite values [**Roc97**][Cor 7.2.1], f must be proper in order for $e^{-f(x)}$ to integrate to one.

For (c), if $dom(f)$ were not full dimensional then it is a subset of a proper affine subspace of $\mathbb{R}^n$. This set has measure zero, which violates the condition that $e^{-f(x)}$ integrates to one.

Lastly, we prove (d). In order that $\int e^{-f(x)}dx = 1$, we must have $f(x) \to \infty$ as $|x| \to \infty$. This means that f is level bounded, which combined with the fact that we can without loss of generality take f to be lower-semicontinuous, gives that f attains its minimum [**RWW10**][Thm 1.9] □

To simplify calculations in the results that follow, we will rewrite the problem $\mathscr{P}$ in a more compact form. We borrow from Rockafellar [**Roc99**] the notion of a supervector, which is simply a concatenated vector consisting of a variable at all time steps. Let $w = (w_0, ..., w_T)'$, $z = (z_0, ..., z_T)'$, $x = (x_0, ..., x_T)'$ be the supervectors corresponding MAP estimates of the dynamical noise, measurements, and states, respectively. Define

$$f(w) = \sum_{t=0}^{T} f_t(w_t),$$

$$g(z) = \sum_{t=0}^{T} g_t(z_t).$$

Note that each of these functions is separable with respect to the components of their respective supervectors. Hence infimums and supremums can be performed with respect to each component.

Define

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -F_0 & 1 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & -F_{T-1} & 1 \end{pmatrix}$$

so that the dynamical system constraint in $\mathscr{P}$ can be represented as

$$Ax - w = 0. \tag{4.4}$$

Similarly, let

$$H = \begin{pmatrix} H_0 & 0 & \cdots & 0 \\ 0 & H_1 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_T \end{pmatrix}.$$

Then the measurement constraint can be rewritten in supervector notation as well, allowing us to rewrite problem $\mathscr{P}$

$$\begin{aligned} \min_{x,w} \quad & f(w) + g(z - Hx) \\ \text{s.t. } & Ax - w = 0. \end{aligned} \tag{4.5}$$

We now turn our attention to convex-analytic duality. For concreteness, we assume that $C \subseteq \mathbb{R}^n$ and $D \subseteq \mathbb{R}^m$. Recall that a convex problem $\min_{x \in C} h(x)$ is *dual* to a concave problem $\max_{y \in D} k(y)$ if there is a convex-concave function $L : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ such that

$$h(x) = \sup_{y \in D} L(x;y) \quad \text{and} \quad k(y) = \inf_{x \in C} L(x;y).$$

This definition, from [**Roc74**], is equivalent to the notion of duality in which one perturbs constraints in order to generate a saddle function L. Indeed, a perturbation function F can be generated from the equation

$$F(x,u) = \sup_{y \in D} \left\{ K(x,y) - u'y \right\}.$$

Of course, this duality framework subsumes the familar Lagrangian duality, Fenchel duality, and various other duality schemes. We refer the reader to [**Roc74**] for details and examples, and focus on applying this theory to the problem at hand.

THEOREM 4.2.3. *When f and g are convex, problem ($\mathscr{P}$) is the primal problem associated with the saddle function*

$$L(w,x;u,y) = f(w) + z'u - g^*(u) - u'Hx + y'(Ax - w)$$

on $C := \mathbb{R}^{(T+1)\times n_x} \times \mathbb{R}^{(T+1)\times n_x}$, $D := \mathbb{R}^{(T+1)\times n_z} \times \mathbb{R}^{(T+1)\times n_x}$

PROOF. To prove that ($\mathscr{P}$) is the primal problem for the saddle-function L, we will show that

$$(\mathscr{P}) = \min_{(w,x)\in C} \sup_{(u,y)\in D} L(w,x;u,y)$$

which satisfies the definition in [**Roc74**]. For ease of notation, in what follows we omit the sets over which we take the infimums and supremums. By the separability of L,

$$\sup_{u,y} L(w,x;u,y) = f(w) + \sup_{u} \left\{u'(z - Hx) - g^*(u)\right\} + \sup_{y} \left\{y'(Ax - w)\right\}$$

Obviously the right-most supremum is 0 when $Ax - w = 0$ and ∞ otherwise. Furthermore, Lemma 4.2.2 gives that without loss of generality g is proper and lsc. Therefore the Fenchel-Moreau Theorem [**RWW10**, Th 11.1] gives that

$$\sup_{u} \left\{u'(z - Hx) - g^*(u)\right\} = g^{**}(z - Hx) = g(z - Hx).$$

Thus

$$\begin{aligned} &\min_{w,x} \sup_{u,y} L(w,x;u,y) \\ &= \min_{w,x} f(w) + g(z - Hx) \\ &\text{s.t. } Ax - w = 0 \end{aligned}$$

□

THEOREM 4.2.4. *The dual problem associated with L on* $\mathbb{R}^{(T+1)\times n_x} \times \mathbb{R}^{(T+1)\times n_x}$, $\mathbb{R}^{(T+1)\times n_z} \times \mathbb{R}^{(T+1)\times n_x}$ *is*

$$\sup_{y,u} z'u - f^*(y) - g^*(u) \tag{4.6}$$
$$s.t.\ A'y - H'u = 0$$

Furthermore, this problem has an equivalent reduced formulation where the supremum is taken over u.

PROOF. The dual problem associated with the triple L, $\mathbb{R}^{(T+1)\times n_x} \times \mathbb{R}^{(T+1)\times n_x}$, $\mathbb{R}^{(T+1)\times n_z} \times \mathbb{R}^{(T+1)\times n_x}$ is

$$\begin{aligned}
&\sup_{u,y} \inf_{w,x} L(w,x;v,y) \\
&= \sup_{u,y} \inf_{w,x} f(w) + z'u - g^*(u) - u'Hx + y'(Ax - w) \\
&= \sup_{u,y} \inf_{w,x} f(w) - w'y + x'(A'y - H'u) + z'u - g^*(u) \\
&= \sup_{u,y} z'u - g^*(u) + \inf_{w} \{f(w) - w'y\} + \inf_{x} \{x'(A'y - H'u)\} \\
&= \sup_{u,y} z'u - g^*(u) - f^*(y) \\
&\text{s.t. } A'y - H'u = 0
\end{aligned}$$

Lastly, the equivalent reduced formulation follows because the each u generates a unique y vector according to the constraints. □

Appealing to the separability of f and g, and expanding the matrices A' and H', we have established the following main result.

THEOREM 4.2.5. *The dual problem associated with the estimation problem ($\mathscr{P}$) is the optimal control problem*

$$(\mathscr{D}) \qquad \sup_{u_0,\dots,u_T} \sum_{t=0}^{T} f_t^*(y_t) + g_t^*(u_t) - z_t' u_t$$

$$s.t.\ y_t = F_t' y_{t+1} + H_t' u_t,\ \ t = 0,\dots,T-1$$

$$y_T = H' u_T$$

The next theorem provides a condition, known as a constraint qualification, for strong duality to hold between the estimation problem $\mathscr{P}$ and control 4.6 problems above.

THEOREM 4.2.6. *[**Roc97**, Theorem 28.2] Assume that the problem $\mathscr{P}$ is strictly feasible, so that there exists a pair (x,w) satisfying (4.4), $w \in int(dom(f))$, $z - Hx \in int(dom(g))$. Then a strong duality relationship exists between the problems $\mathscr{P}$ and $\mathscr{D}$. In other words, the supremum in $\mathscr{D}$ equals the optimal value in $\mathscr{P}$. Furthermore, this supremum is attained.*

PROOF. This follows directly from the strong duality theorem in [**Roc97**, Theorem 28.2]. Note that the typical formulation of this constraint qualification requires only a *relative interior* point, when the domains are considered as subsets of their affine hulls, instead of an interior point. However, since by Lemma 4.2.2 the domains of f and g are full-dimensional, the notions are equivalent. □

4.3. The Piecewise Linear Quadratic Case

In this section we investigate structural constraints on the functions f_t and g_t in the densities of $\mathbf{W}_t$ and $\mathbf{V}_t$ that allow us to remove the constraint qualification condition in 4.2.6.

Recall that a function is linear-quadratic if it is polynomial of degree at most two, so that constant and linear functions are included in this family.

THEOREM 4.3.1. *Assume that f_t and g_t are convex and piecewise linear-quadratic. If either $\mathscr{P}$ or $\mathscr{D}$ are feasible, then strong duality holds between these estimation and optimal control problems, so that their optimal objective values are equal. Furthermore, both problems attain their optimal objective values.*

PROOF. If f_t and g_t are piecewise linear quadratic, then the reformulated problem 4.5 is a piecewise linear-quadratic program. Lemma 4.2.2 gives us that each of f_t and g_t are proper, and hence each of their

conjugates is as well. Combined with the assumption that one of $\mathscr{P}$ and $\mathscr{D}$ is feasible, we know that the optimal objective value of this problem is finite. Strong duality and the attaining of optimal values then follow directly from [**RWW10**][Thm 11.42] □

Theorem 4.3.1 removes the contraint qualification of Theorem 4.2.5 by imposing extra structure on the functions f_t and g_t. By assuming that f_t and g_t are piecewise linear-quadratic, strong duality becomes automatic.

When f_t and g_t are arbitrary convex functions, computing closed form expressions for the conjugates that occur in the dual problem may be difficult. The conjugate function, Lagrangian, and dual problem $\mathscr{D}$ are especially easy to compute in the special case that f_t and g_t are monitoring functions, which includes many problems of practical interest.

A *monitoring function* is a function $\rho_{U,M} : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$\rho_{U,M}(x) = \sup_{u \in U} \{x'u - \frac{1}{2}u'Mu\}$$

where $U \subseteq \mathbb{R}^n$ is a nonempty polyhedral set and M is an $n \times n$ positive semidefinite matrix.

Monitoring functions are flexible tools for modeling penalties. They are proper, convex, and piecewise linear-quadratic [**RWW10**][Ex 11.18], and can be used to model a variety of linear and quadratic penalties in addition to polyhedral constraints. A probabilistic interpretation of the use of monitoring functions in robust smoothing problems can be found in [**ABP13**]. The authors detail the construction of many commonly used penalties in robust optimization, and provide remarks for constructing others.

The incentive for considering the case that f_t and g_t are monitoring functions is two-fold. The first is that the MAP problem with Gaussian noise is contained in this case. The second is that the framework provides enough structure to make the computation of the conjugates and the dual control problem $\mathscr{D}$ straight-forward.

We'll be aided in this by the following lemma.

LEMMA 4.3.2. *If $\rho_{U,M}$ is a monitoring function, then the conjugate $\rho^*_{U,M}$ is given by*

$$\rho^*_{U,M}(y) = \begin{cases} \frac{1}{2}y'My & \text{when } y \in U \\ \infty & \text{otherwise} \end{cases}$$

PROOF.

$$\begin{aligned}\rho^*_{U,M}(y) &= \sup_{x\in\mathbb{R}^n}\left\{x'y-\rho_{U,M}(x)\right\}\\ &= \sup_{x\in\mathbb{R}^n}\left\{x'y-\sup_{u\in U}\left\{x'u-\frac{1}{2}u'Mu\right\}\right\}\\ &= \sup_{x\in\mathbb{R}^n}\inf_{u\in U}\left\{x'(y-u)+\frac{1}{2}u'Mu\right\} \qquad (4.7)\end{aligned}$$

This can be interpreted as the dual to the problem

$$\inf_{u\in U}\frac{1}{2}u'Mu$$

$$\text{s.t. } u=y$$

In the same spirit as 4.3.1, we apply [**RWW10**][11.42] to see that, when U is nonempty, 4.7 equals

$$\begin{aligned}&\inf_{u\in U}\sup_{x\in\mathbb{R}^n}\left\{x'(y-u)+\frac{1}{2}u'Mu\right\}\\ &\qquad=\begin{cases}\frac{1}{2}y'My & \text{when } y\in U\\ \infty & \text{otherwise}\end{cases}\end{aligned}$$

□

We've arrived at a precise formulation of the dual control problem $\mathscr{D}$.

COROLLARY 4.3.3. *If* $\mathbf{W}_t$ *and* $\mathbf{V}_t$ *have a PLQ density*

$$\mathbf{W}_t\propto e^{-\rho_{W_t,M_t}}\quad \mathbf{V}_t\propto e^{-\rho_{V_t,N_t}}$$

with $M_t, N_t \succeq 0$ and and W_t, V_t nonempty and polyhedral, then the MAP problem $\mathscr{P}$ is dual to the control problem

$$\sup_{u_0,\dots,u_T} \sum_{t=0}^{T} \frac{1}{2} y_t' M_t y_t + \frac{1}{2} u_t' N_t u_t - z_t' u_t \tag{4.8}$$

$$s.t.\ y_t = F_t' y_{t+1} + H_t' u_t,\ \ t = 0,\dots,T-1$$

$$y_T = H' u_T$$

$$y_t \in W_t, \quad t = 0,\dots,T \tag{4.9}$$

$$u_t \in V_t, \quad t = 0,\dots,T \tag{4.10}$$

Optimal values are attained and strong duality holds when either problem is feasible.

PROOF. Follows directly from 4.2.5 and 4.3.1. □

As an application of Theorem 4.3.3, we verify the strong duality between estimation and control in the linear-quadratic Gaussian setting. By taking $\mathbf{W}_t \sim \mathscr{N}(0, Q_t)$ and $\mathbf{V}_t \sim \mathscr{N}(0, R_t)$, we can represent the density functions as in Theorem 4.3.3 by taking

$$W_t = \mathbb{R}^{n_x},\ M_t = Q_t, \quad V_t = \mathbb{R}^{n_z}, N_t = R_t$$

Because $\rho_{t,\mathbb{R}^{n_x},Q_t}$ and $\rho_{t,\mathbb{R}^{n_z},R_t}$ have domains $\mathbb{R}^{n_x}$ and $\mathbb{R}_{n_z}$, respectively, the MAP problem is feasible for any measurements $(z_0,\dots,z_T)$. Strong duality then follows directly from Theorem 4.3.3.

Moreover, by taking $\mathbf{W}_t \sim \mathscr{N}(0, Q_t)$ and $\mathbf{V}_t \equiv 0$, we recover the duality of the two problems considered by Kalman from the introduction. The time reversal is in fact an artifact of taking the convex analytic dual, though recovery of the Riccati-covariance propagation equivalence requires considering each problem as sequential, which is not the appoach that we've taken here.

4.4. Applications: Reconstructing an Estimator from Optimal Controls

In this section we apply the results of the previous sections to construct the solution to an optimal estimation problem from the solution to its dual problem of optimal control. We focus on a nonsmooth problem of practical interest. First, we formulate an estimation problem where the density of the measurement noise is generated via log-concave maximum likelihood estimation from a sample of measurement noise. We then

use the result 4.2.5 to construct a corresponding dual control problem. Finally, we use the solution to this control problem to construct an optimal estimator for the original problem.

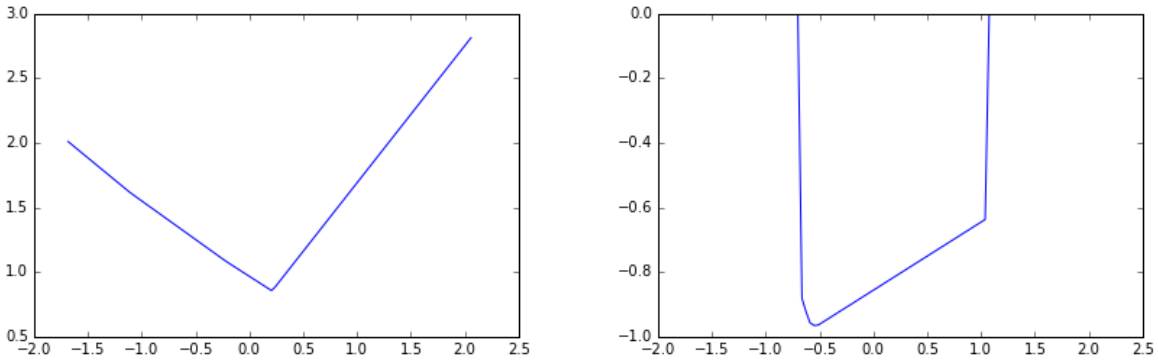
The set up for this problem motivated by the following scenario. A practioner aims to estimate the current and previous states of a dynamical system from a set of noisy measurements. Through calibration of a sensor or observation, the practioner has the ability to generate sample data for the measurement noise. How can one use this data to formulate an estimation problem which reflects the tendencies of the sensor? We would like to capitalize on the ability to generate measurement noise, instead of defaulting to a Gaussian noise assumption.

The following theorem, due to [**Ruf**] and [**PWM07**], as well as the computational results in [**DR**], allow us to form a nonparametric density estimate of a log-concave density based on a set of sample data.

THEOREM 4.4.1. *If $X_1,...,X_n$ are i.i.d. observations from a univariate log-concave density, then the nonparametric MLE exists, is unique, and is of the form $\hat{\phi}_n = \exp \hat{f}_n$. The function $\hat{f}_n$ is piecewise linear on $[X_{(1)},X_{(n)}]$, with the set of knots contained in $\{X_1,...,X_n\}$. Outside of $[X_{(1)},X_{(n)}]$, $\hat{f}$ takes the value ∞.*

For the generation of our example problem, we use 10 time steps and a two dimensional state space, motivated by components representing position and velocity. The dynamics matrices F_t corresponds to the physical dynamics that would occur in such a situation. We produce a "true" sequence of states to be estimated by generating dynamical system noise according to a $\mathcal{N}(0,I)$ distribution. We take the measurement operators H_t be the sum of the components. Lastly, we construct measurements from a sample of Laplace$(0,1)$ measurement noise.

For the formulation of the MAP problem, we assume that the dynamical system noise was generated according to $\mathcal{N}(0,I)$ distribution. For the measurement noise, we construct a log-concave MLE estimator $e^{-\hat{g}}$ of the density from a sample of size 100 generated from a Laplace(0,1) distribution. $\hat{g}$ and its convex conjugate $\hat{g}^*$ are illustrated in figure 4.1.



(a) $\hat{g}$ (b) $\hat{g}^*$

FIGURE 4.1. $\hat{g}$ and $\hat{g}^*$, where the MLE density is $e^{-g(x)}$

The MAP problem $\mathscr{P}$ is then

$$(\mathscr{P}_{ex}) \qquad \min_{x_0,\dots,x_{10}} \sum_{t=0}^{10} \frac{1}{2}||w_t||^2 + \hat{g}(z_t - \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot x_t)$$

$$\text{s.t. } x_{t+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} x_t + w_{t+1}, \quad t = 0,\dots,T-1$$

$$x_0 = w_0$$

According to 4.2.5 this problem has as its dual the control problem

$$(\mathscr{D}_{ex}) \qquad \max_{u_0,\dots,u_{10}} \sum_{t=0}^{10} \frac{1}{2}||y_t||^2 + \hat{g}^*(u_t) - u_t' z_t$$

$$\text{s.t. } y_t = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} y_{t+1} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u_t, \quad t = 0,\dots,T-1$$

$$y_T = \begin{pmatrix} 1 \\ 1 \end{pmatrix} u_T$$

Because our dynamical system noise in $\mathscr{P}_{ex}$ has full support, Theorem 4.2.6 guarantees that strong duality holds between the problems, and that the dual control problem attains its solution. Assume that we have solved this control problem and have a corresponding optimal control (u^*, y^*). Since an optimal estimate (w^*, x^*) gives a saddle point $(w^*, x^*; u^*, y^*)$ to the Lagrangian L in Theorem 4.2.4, it follows from

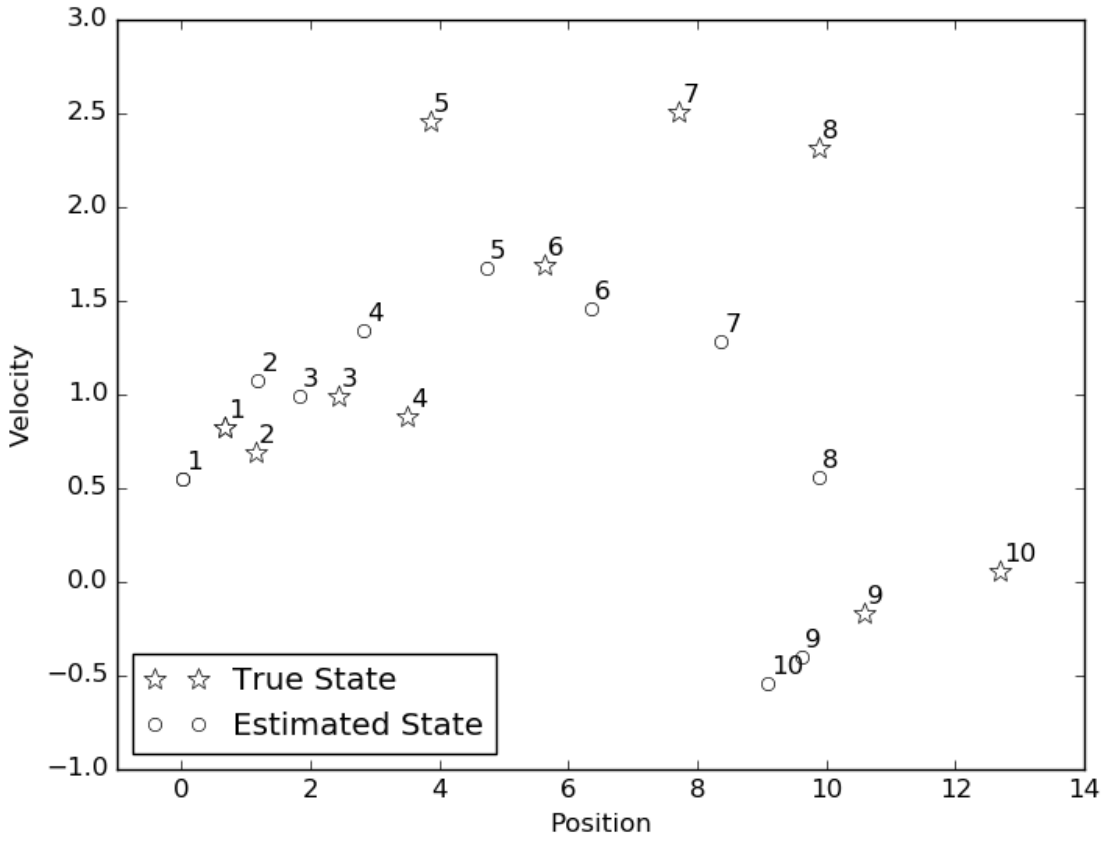


FIGURE 4.2. Dual Reconstruction of State Estimate

the proof of this theorem that w^* minimizes $f(w) - w'y^*$. In our problem f, when $f(w) = \frac{1}{2}||w||^2$, which yields the relationship $w^* = y^*$. This is similar to the relationship between primal and dual solutions in the Fenchel Duality framework. See [**Ber09**][Prop. 5.3.8] for further details. Note that this allows us to reconstruct a primal solution from a dual solution and vice versa. In particular, the relationship between y^* and w^* is linear when the dynamical system noise is assumed to be Gaussian.

Solving $\mathscr{D}_{ex}$ to optimality gives y^*, from which we generate w^* and then an optimal estimate x^*. Figure 4.2 contains a plot of the estimated and true state.

Though we have demonstrated a convenient technique to generate solutions for an estimation problem from the solution to its dual control problem, in this example we have no reason to believe that solving $\mathscr{D}_{ex}$ is any easier than solving the original problem $\mathscr{P}_{ex}$. Nevertheless, the results in this and previous sections

provide motivation for further investigation into applying control algorithms to solve estimation problems and vice versa.

CHAPTER 5

A Duality Result for Portfolio Optimization in Set-Valued Conic Market Models

5.1. Introduction

Portfolio optimization problems have a long and rich history. Traditionally, portfolio optimization took place in models without transaction costs, as in [Mer71]. Portfolio optimization problems in financial market models which include proportional transaction costs first appeared in the work of Magill and Constantinides [MC76]. The two asset model was solved rigorously and improved upon shortly thereafter [DN90] [SS94]. Since then, much more focus has been placed on deriving results in markets with transaction costs, as researchers tried to develop results analogous to the classical case.

In this chapter, we consider the conic market model initially developed by Kabanov [Kab99] for modeling foreign transaction markets. The conic market model expresses portfolios in terms of the number of physical assets they contain, as opposed to their value in currency. This allows the formulation of wealth processes without the explicit use of stochastic integration. Though this quality may be seem surprising to portfolio optimization veterans, it is attractive in terms of its simplicity and intuitive nature. In order to compare portfolios, we consider them as assets of vectors under a partial ordering, and invoke the theory of set optimization to formulate the portfolio problem.

Set optimization is primarily motivated by the desire to optimize with respect to a non-total order relation. Vector-valued optimization fits directly into this framework, with component-wise comparison. However, we prefer to work in a set-optimization framework as opposed to the less general vector-optimization because of its succinct theory [HHL$^+$ng]. Following the work of [Ham09], we introduce an ordering of sets which generates a complete lattice. This allows us to define corresponding notion of infimums and supremums of sets–a fundamental step in the formulation of the portfolio optimization problem. We then

This chapter is based on joint work with Khoa Le [BL16].

use the tools in [HL14] to formulate a set-valued dual to the portfolio optimization problem. Because we consider the multistage problem, our results are generalize those in [Wan11].

Constructing a primal-dual pair of problems for set-valued portfolio optimization provides insight into the relationship between traditional portfolio optimization theory and the proportional transaction cost case. But it is also our hope that the results contained here would be of more than just theoretical interest. Recent work [LRU14] [HLR14] has been investigating computational techniques for solving set-valued optimization problems. In particular, [LRU14] uses both the primal and dual formulation of a set optimization problem to work towards computing a solution. In this sense, the results in this chapter provide a valuable relationship which will help bring the portfolio optimization problem considered closer to practitioners, as computational techniques progress.

The rest of this chapter is organized as follows. In the first section, we introduce the material which we use to formulate the set-valued portfolio optimization problem. This includes a review of the conic market model in addition to a summary of the set-optimization tools and techniques the we will use in our problem formulation. The next section explicitly formulates the multi-period utility maximization problem. In the third section, we discuss duality in the set optimization framework. The fourth section is devoted to our main results, the formulation of a dual problem and a proof that a strong duality relationship holds. The last section applies the main results to an example utility maximization problem.

5.2. Preliminaries

5.2.1. Conical Market Model. In this section, we recall the framework of the conic market model with transaction costs introduced in [Kab99], though we primarily follow the development in [Sch04].

Consider a financial market which consists of d traded assets. In classical models, we assume that at some terminal time T all assets are liquidated, i.e. converted to some numeraire. In certain applications, this is unrealistic. For example, an agent with a portfolio consisting of assets in both US and European markets should not need to choose between liquidation into Euro or USD to establish its relative value. For this reason, we use a numeraire-free approach by considering vector-valued portfolios. In particular, we express portfolios in terms of the number of physical units of each asset, instead of the value of those assets with respect to a numeraire. This approach is especially interesting when liquidation into some numeraire has an

associated transaction cost. In this case conversion to a unified currency is irreversible, and different choices of numeraire could result in different relative values of portfolios.

We consider a market in which transaction costs are proportional to the number of units exchanged. To model these costs, we introduce the notion of a bid-ask matrix.

DEFINITION 5.2.1. A *bid-ask matrix* is a $d \times d$ matrix Π such that its entries π^{ij} satisfy

(1) $\pi^{ij} > 0$, for $1 \leqslant i, j \leqslant d$

(2) $\pi^{ii} = 1$, for $1 \leqslant i \leqslant d$.

(3) $\pi^{ij} \leqslant \pi^{ik}\pi^{kj}$, for $i \leqslant i, j, k \leqslant d$.

The terms of trade in the market are given by the bid-ask matrix, in the sense that the entry π^{ij} gives the number of units of asset i which can be exchanged to one unit of asset j. Thus the pair $\{\frac{1}{\pi^{ji}}, \pi^{ij}\}$ denotes the bid and ask prices of the asset j in terms of the asset i. A financial interpretation of the first and second properties of a bid-ask matrix is straightforward, with the third condition ensuring that an agent cannot achieve a better exchange rate through a series of exchanges than exchanging directly.

Next, we consider the notions of solvency and available portfolios. Recall that, given a set $C \subseteq \mathbb{R}^d$, the *convex cone generated by C* is the set

$$\text{cone}(C) = \{\sum_{i=1}^{n} \lambda_i y_i : y_i \in C, \lambda_i \geqslant 0, 1 \leqslant i \leqslant n, n \in \mathbb{N}\}.$$

DEFINITION 5.2.2. For a given bid-ask matrix Π, the *solvency cone* $K(\Pi)$ is the convex cone in $\mathbb{R}^d$ generated by the unit vectors e^i and $\pi^{ij}e^i - e^j$, $1 \leqslant i, j \leqslant d$.

Solvent positions in vector valued portfolios are those which can be traded to the zero portfolio. The vector $\pi^{ij}e^i - e^j$, which consists of π^{ij} long position in asset i and one short in asset j, is solvent because the terms of trade given by Π allow exchanging $\pi^{ij}e^i$ to e^j. It follows that any non-negative linear combination of $\pi^{ij}e^i - e^j$ is also solvent. We also allow an agent to discard non-negative quantities of an asset in order to trade to the zero portfolio, which justifies including the e^i vectors in the solvency cone definition.

What is that set of portfolios that can be obtained from the zero portfolio, according to the terms of trade governed by Π? Similar to the definition of the solvency cone, it consists of vectors $e^j - \pi^{ij}e^i$, which correspond to trades at the exchange rate given by Π. Again permitting trades where agents discard resources, we see that the set of portfolios available at price zero is the cone $-K(\Pi)$.

Given a cone $K \subseteq \mathbb{R}^d$, we denote by K^+ the positive polar cone of K, i.e.,

$$K^+ = \{w \in \mathbb{R}^d : \langle v, w \rangle \geqslant 0 \text{ for } v \in K(\Pi)\}.$$

Recall that the interior of K^+ is the set

$$\operatorname{int} K^+ = \{w : \langle w, x \rangle > 0, \forall x \in K, x \neq 0\}.$$

DEFINITION 5.2.3. A nonzero element $w \in \mathbb{R}^d$ is a *consistent price system* for the bid-ask matrix Π if w is in the positive polar cone of $K(\Pi)$, so that

$$w \in K^+(\Pi) = \{w \in \mathbb{R}^d : \langle v, w \rangle \geqslant 0 \text{ for } v \in K(\Pi)\}.$$

The set of all consistent price systems for a bid-ask matrix is then simply $K^+(\Pi)\backslash\{0\}$.

The notion of a consistent price system has an important financial interpretation. A price system w gives a non-negative price w^i for each asset i. One interpretation of the definition of a consistent pricing system is that, if we fix some numeraire asset i, then w satisfies the condition that the frictionless exchange rate $\frac{w^j}{w^i}$ for asset j is less than π^{ij}. Allowing for arbitrary choice of numeraire i and asset j, this is equivalent to

$$\{w \in \mathbb{R}^d_+ \backslash \{0\} : \frac{w^j}{w^i} \geqslant \pi^{ij} \text{ for } 1 \leqslant i, j \leqslant d\}.$$

One can easily show that this set is in fact equal to $K^+(\Pi)\backslash\{0\}$, the set of all price systems consistent with Π.

Fixing a filtered probability space $(\Omega, (\mathscr{F})_{t=0}^T, \mathbb{P})$, we model a financial market by $(\Pi_t)_{t=0}^T$, an $\mathscr{F}$ adapted process taking values in the set of bid-ask matrices. Such a process will be called a bid-ask process. We make the following simplifying assumptions.

ASSUMPTION 5.2.4. $(\Omega, (\mathscr{F})_{t=0}^T, \mathbb{P})$ satisfies

- $\mathscr{F}_0 = \{\varnothing, \Omega\}$ is trivial.
- The model is in discrete time with $t = 0, ..., T$
- The probability space Ω is finite, with $|\Omega| = N$
- Each element in Ω has nonzero probability, i.e. $\mathbb{P}[\omega_n] = p_n > 0$, where $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$ and $n = 1, .., N$.

The assumption that Ω is finite means that we can identify the space of all d-dimensional random variables with $\mathbb{R}^{d\times N}$ and inner product $\mathbb{E}\langle x,y\rangle$, where x and y are random vectors. As a result, the different topologies $L^\infty(\Omega,\mathscr{F},\mathbb{P})$, $L^1(\Omega,\mathscr{F},\mathbb{P})$, $L^0(\Omega,\mathscr{F},\mathbb{P})$, etc. on the set of all $\mathbb{R}^d$-valued random variables $X:\Omega\to\mathbb{R}^d$ are isomorphic. We will refer to this topology simply as $L^p(\Omega,\mathscr{F},\mathbb{P})$ for some $p\in(1,\infty)$, in order to make clear when we are referring to the dual space $L^q(\Omega,\mathscr{F},\mathbb{P})$ where $\frac{1}{p}+\frac{1}{q}=1$. For ease of notation, we will denote these spaces $L^p(\mathscr{F};\mathbb{R}^d)$ and $L^q(\mathscr{F};\mathbb{R}^d)$. For the sake of notation, we will denote the components of a vector $x\in\mathbb{R}^{d\times N}$ by $x_i(\omega)$ for $\omega\in\Omega$, $1\leqslant i\leqslant N$.

Again exploiting the finiteness of Ω, we know that any cone in $L^p(\mathscr{F}_t,\mathbb{R}^d)$ generated by a finite set of random vectors $\{x_i\}_{i=1}^m$ is generated by $\{x_i\mathbb{1}_{\Gamma_j}\}$ in $\mathbb{R}^{d\times N}$, where $\{\Gamma_j\}_{j\in J}$ is the set of atoms of $\mathscr{F}_t$.

Let $(\Pi_t)_{t=0}^T$ be a bid-ask process. This generates a cone-valued process $(K_t)_{t=0}^T$ where each K_t is an associated solvency cone. We denote by $L^p(\mathscr{F}_t,K_t)$ the set

$$\{X\in L^p(\mathscr{F}_t;\mathbb{R}^d):X_t(\omega)\in K_t(\omega)\text{ for each }\omega\in\Omega\}.$$

for each $\omega\in\Omega$.

We can now define the notion of a self-financing portfolio.

DEFINITION 5.2.5. An $\mathbb{R}^d$-valued adapted process $\vartheta=(\vartheta_t)_{t=0}^T$ is called a self-financing portfolio process if the increments

$$\xi_t(\omega):=\vartheta_t(\omega)-\vartheta_{t-1}(\omega)$$

belong to the cone $-K_t(\omega)$ of portfolios available at price zero, for all time $t=0,\dots,T$. We also put $\vartheta_{-1}=0$ by convention.

For each $t=0,\dots,T$, we denote by A_t the convex cone in $L^p(\mathscr{F}_t;\mathbb{R}^d)$ formed by the random variables ϑ_t, where $\vartheta=(\vartheta_i)_{i=0}^T$ runs through the self-financing portfolio processes. We always assume that the initial portfolio ϑ_0 is deterministic. A_T may be then interpreted as the set of positions available at time T from an initial endowment $0\in\mathbb{R}^d$. More precisely, if we denote by $\mathbb{1}\in L^p(\mathscr{F}_0,\mathbb{R}^d)$ the constant random variable that assumes the value 1, we have the following result:

PROPOSITION 5.2.6. *For each $t=1,...,T$,*

$$A_t=-K_0\mathbb{1}-L^p(\mathscr{F}_1;K_1)-\cdots-L^p(\mathscr{F}_t;K_t)\,.$$

PROOF. Assume $x \in A_t$. Then x is a convex combination of random variables

$$x = \sum_{j=1}^{n} \lambda^j \vartheta_t^j$$

where for each j, $(\vartheta_i^j)_{i=0}^T$ is a self-financing portfolio and $\lambda^j \geqslant 0$. We rewrite x as

$$x = \sum_{j=1}^{n} \lambda^j \left(\vartheta_t^j - \vartheta_{t-1}^j + \vartheta_{t-1}^j - \ldots - \vartheta_0^j + \vartheta_0^j \right).$$

Expanding this sum, each $\sum_{j=1}^n \lambda^j (\vartheta_i^j - \vartheta_{i-i}^j) \in -K(\Pi_i)$ since $-K(\Pi_i)$ is a cone and $\vartheta_i^j - \vartheta_{i-1}^j \in -K(\Pi_i)$ for every j. We have established that

$$A_t \subseteq -K_0 \mathbb{1} - L^p(\mathscr{F}_1; K_1) - \cdots - L^p(\mathscr{F}_t; K_t).$$

The reverse containment follows by a symmetric argument and is omitted. □

Similarly, if one starts with an initial endowment $x_0 \in \mathbb{R}^d$, then the collection of all random portfolios available at time T is given by $A_T(x_0) = x_0 \mathbb{1} + A_T$. Explicitly, we have

$$A_T(x_0) = x_0 \mathbb{1} - K_0 \mathbb{1} - L^p(\mathscr{F}_1; K_1) - \cdots - L^p(\mathscr{F}_T; K_T), \tag{5.1}$$

with convention $A_T(0) = A_T$.

Another important concept in a financial market model is the concept of arbitrage. In the conic market model framework, the bid-ask process $(\Pi_t)_{t=0}^T$ is said to satisfy the no arbitrage property if

$$A_T \cap L^p(\mathscr{F}_T; \mathbb{R}^d_+) = \{0\}. \tag{5.2}$$

We will assume that our market model satisfies the no arbitrage property.

In classical financial market models, no arbitrage is intimately connected to the existence of an equivalent martingale measure. The corresponding notion in the conic market model is a consistent pricing process.

DEFINITION 5.2.7. An adapted $\mathbb{R}^d_+$-valued process $(Z_t)_{t=0}^T$ is called a consistent pricing process for the bid-ask process $(\Pi_t)_{t=0}^T$ if Z is martingale and $Z_t(\omega)$ lies in $K_t(\omega)^+ \backslash \{0\}$ for each $t = 0, \ldots, T$.

The following extension of the Fundamental Theorem on Asset Pricing, due to Kabanov and Stricker, [**KS01**], establishes the connection between no arbitrage and consistent pricing processes.

THEOREM 5.2.8. *Let Ω be finite. The bid-ask process $(\Pi_t)_{t=0}^T$ satisfies the no arbitrage condition if and only if there is a consistent pricing system $Z = (Z_t)_{t=0}^T$ for $(\Pi_t)_{t=0}^T$.*

This theorem is a fundamental component in the proof of our main duality result, Theorem 5.5.2.

We will make use of this result in the form of the following lemma

LEMMA 5.2.9. *A bid-ask process satisfies the no arbitrage condition if and only if*

$$(-A_T^+) \cap \operatorname{int}(L^q(\mathscr{F}_T, \mathbb{R}_+^d)) \neq \varnothing.$$

PROOF. For the forward direction, assume a bid-ask process satisfies the no arbitrage condition. Then

$$A_T \cap L^p(\mathscr{F}_T, \mathbb{R}_+^d) = \{0\}.$$

Since Ω is finite, A_T is the sum of finitely generated closed convex cones, so it is a finitely generated convex cone, and hence closed. Let $C := \operatorname{conv}(\{e(\omega)_i, 1 \leqslant i \leqslant d, \omega \in \Omega\})$, where each $e(\omega)_i$ is a unit vector in $\mathbb{R}^{d\times N}$. Since C is the convex hull of a finite set of points in $L^p(\mathscr{F}_T, \mathbb{R}_+^d)$, it is compact. Obviously $0 \notin C$. By the separation theorem (in the case that one set is closed and the other compact) there is a nonzero $z \in \mathbb{R}^n$ that strictly separates C and A_T. That is,

$$\sup_{x\in A_T} \langle z, x\rangle < \inf_{y\in C} \langle z, y\rangle.$$

C is compact, so the expression on the right-hand side of the inequality is finite. Since A_T is a cone, the left-hand side of the inequality must then be zero, so $z \in -A_T^+$. Furthermore, $\langle z, y\rangle > 0$ for each $y \in C$, so that $\langle z, \lambda y\rangle > 0$ for all $y \in C$ and $\lambda \neq 0$. Since C generates $L^p(\mathscr{F}_T)$, we have that $\langle z, \lambda y\rangle > 0$ for all $y \in L^p(\mathscr{F}_T, \mathbb{R}_+^d)$ with $y \neq 0$. It follows that

$$z \in \operatorname{int}\left((L^p(\mathscr{F}_T, \mathbb{R}_+^d))^+\right) = \operatorname{int}\left(L^q(\mathscr{F}_T, \mathbb{R}_+^d)\right).$$

For the reverse direction, assume that

$$(-A_T^*) \cap \operatorname{int}(L^q(\mathscr{F}_T, \mathbb{R}_+^d)) \neq \varnothing.$$

Then there is a z such that

$$\langle z,x\rangle \leqslant 0 \text{ for } x \in A_T \text{ and } \langle z,y\rangle > 0 \text{ for } y \in K\backslash\{0\}.$$

This obviously implies that A_T and $K\backslash\{0\}$ are disjoint. □

5.2.2. Set Optimization. In this section, we review the components of set-valued optimization that will be necessary to introduce the portfolio optimization problem. For a more detailed exposition of the set-valued optimization framework and the corresponding duality theory, see [**HHL$^+$ng**, **Ham09**, **HL14**].

We begin by constructing a suitable notion of "order" for sets. Let Z be a non-trivial real linear space. Given a convex cone $C \subsetneq Z$ with $0 \in C$, we have a preordering of Z, denoted by $\leqslant_C$, which is defined as

$$z_1 \leqslant_C z_2 \iff z_2 - z_1 \in C$$

for any $z_1, z_2 \in Z$. The following are equivalent to $z_1 \leqslant_C z_2$,

$$z_1 \leqslant_C z_2 \iff z_2 - z_1 \in C \iff z_2 \in z_1 + C \iff z_1 \in z_2 - C.$$

These last two expressions can be used to extend $\leqslant_C$ from Z to $\mathscr{P}(Z)$, the power set of Z. Given $A, B \in \mathscr{P}(Z)$, we define two possible extensions.

$$A \leqslant_C B \iff B \subseteq A + C$$

$$A \leqslant_C B \iff A \subseteq B - C.$$

We use $+$ to denote Minkowski addition of sets, with set convention that $A + \varnothing = \varnothing + A = \varnothing$ for all $A \in \mathscr{P}(Z)$.

In what follows we will exclusively discuss the relation $\leqslant_C$, which is appropriate for set-valued minimization. Each of the results we include has a corresponding result for $\leqslant_C$ in the maximization context, which we omit. For further details see [**HHL$^+$ng**].

In addition, we assume that Z is equipped with a Hausdorff, locally convex topology. We consider the space

$$\mathscr{G}(Z,C) = \{A \subseteq Z : A = \operatorname{cl}\operatorname{conv}(A + C)\}$$

where cl conv is the closure of the convex hull. We abbreviate $\mathscr{G}(Z,C)$ to $\mathscr{G}(C)$, when Z is clear from the context. We define an associative and commutative binary operation $\oplus : \mathscr{G}(C) \times \mathscr{G}(C) \to \mathscr{G}(C)$ by

$$A \oplus B = \mathrm{cl}\,(A + B)$$

Observe that in $\mathscr{G}(C)$, the relation $\preccurlyeq_C$ reduces to containment. For any $A, B \in \mathscr{G}(C)$,

$$A \preccurlyeq_C B \iff A \subseteq B.$$

As shown in [**Ham09**], the pair $(\mathscr{G}(C), \supseteq)$ is a complete lattice, meaning that $\supseteq$ yields a partial order on $\mathscr{G}(C)$, and that each subset of $\mathscr{G}(C)$ has an infimum and supremum with respect to $\supseteq$ in $\mathscr{G}(C)$. Given $\varnothing \neq \mathscr{A} \subseteq \mathscr{G}(C)$, the infimal and supremal elements in $\mathscr{G}(C)$ are

$$\inf_{(\mathscr{G}(C),\supseteq)} \mathscr{A} = \mathrm{cl}\,\mathrm{conv} \bigcup_{A \in \mathscr{A}} A, \qquad \sup_{(\mathscr{G}(C),\supseteq)} \mathscr{A} = \bigcap_{A \in \mathscr{A}} A. \tag{5.3}$$

In order to preserve intuition, it is useful to recall how this framework relates to the familiar complete lattice of the extended real numbers $\mathbb{R} \cup \{\pm\infty\}$ with the $\leqslant$ order. The extended real-numbers translate into the set-valued framework described above by using the ordering cone $C = \mathbb{R}_+$ and identifying each point $z \in \mathbb{R}$ with the set $\{z\} + \mathbb{R}_+$ in $(\mathscr{G}(\mathbb{R}, \mathbb{R}_+), \supseteq)$. Moreover, $+\infty$ and $-\infty$ in the usual framework are replaced by $\varnothing$ and $\mathbb{R}$, respectively, in the set-valued case.

Next, assume that X, Y are two locally convex spaces, and that $D \subseteq Y$ is a convex cone with $0 \in D$. Let $f : X \to \mathscr{G}(C)$ and $g : X \to \mathscr{G}(D)$ be two set-valued functions. We consider optimization problems of the form

$$\min_{x \in X} f(x) \text{ subject to } 0 \in g(x).$$

Where the minimum refers to the set-valued ordering previously discussed. In other words, we want to find the set

$$\inf_{(\mathscr{G}(C),\supseteq)} \{f(x) | x \in X, 0 \in g(x)\} = \mathrm{cl}\,\mathrm{conv} \bigcup \{f(x) | x \in X, 0 \in g(x)\}.$$

This is the minimum of our optimization problem.

Extending the notion of a minimizer to the set-valued case is slightly more subtle. Given $f : X \to \mathscr{G}(C)$ and $M \subseteq X$, we denote the set of all values of f on M by

$$f[M] = \{f(x) | x \in M\}.$$

The minimal elements of $f[M]$ are defined by

$$\text{Min } f[M] := \{f(x) | f(x) \in f[M] \text{ and } \forall f(y) \in f[M] \text{ with } f(y) \supseteq f(x), f(y) = f(x)\}.$$

Similarly, an element $\bar{x}$ is a minimizer of f on M if $f(\bar{x}) \in \text{Min } f[M]$.

In addition to a minimality condition, we also expect a solution to attain the infimum of a problem. We say that the infimum of a problem

$$\min f(x) \text{ subject to } x \in X$$

is *attained* at a set $\overline{X} \subseteq X$ if

$$\inf_{x \in \overline{X}} f(x) = \inf_{x \in X} f(X).$$

As per the definition of infimum in (5.3), this means that

$$\text{cl conv} \bigcup_{x \in \overline{X}} f(x) = \text{cl conv} \bigcup_{x \in X} f(x).$$

Alternatively, we say that the set $\overline{X}$ is an *infimizer* of the problem. Combining both of these requirements, we arrive at an appropriate notion of a solution to a set optimization problem.

DEFINITION 5.2.10. Given $f : X \to \mathscr{G}(C)$, an infimizer $\overline{X} \subseteq X$ is called a *solution* to the problem

$$\min f(x) \text{ subject to } x \in X$$

if $\overline{X} \subseteq \text{Min } f[X]$. Similarly, we call an infimizer $\overline{X} \subseteq X$ a *full solution* to the problem if $\overline{X} = \text{Min } f[X]$.

In the typical optimization framework of the extended real numbers, the notion of an infimizer and minimizer coincide because the search for infimizers can be reduced to singleton sets. In the set-optimization setting, this is not the case, which warrants the above definition. The infimum of a problem is, in general, the closure of the union of function values, which is not necessarily a function value itself. Further details and a more in depth review of this issue can be found in [**HL14**].

We next review some important convex-analytic type properties for set-valued functions.

DEFINITION 5.2.11. A set valued function $f : X \to (\mathscr{G}(C), \supseteq)$ is said to be *convex* if for every pair x_1, $x_2 \in X$ and every $t \in (0,1)$

$$f(tx_1 + (1-t)x_2) \supseteq tf(x_1) + (1-t)f(x_2).$$

It is straight-forward to show that convexity of f is equivalent to convexity of the graph of f, where

$$\text{graph } f := \{(x,z) \in X \times Z | z \in f(x)\} \subseteq X \times Z.$$

We end this section with the following results, found in [**HL14**], which use convexity to simplify the computation of infimums and Minkowski sums.

PROPOSITION 5.2.12. *If $f : X \to \mathscr{P}(Z)$ is convex and*

$$f(x) = \text{cl}\,(f(x) + C),$$

then $f(x) \in \mathscr{G}(C)$

PROOF. We want to show that for each $x \in X$, $f(x) = \text{cl}\,\text{conv}\,(f(x) + C)$, given that f is convex and $f(x) = \text{cl}\,(f(x) + C)$. It suffices to show that $f(x) + C$ is convex, which will follow if $f(x)$ is convex because the Minkowksi sum of two convex sets is convex [**Roc97**]. But $f(x)$ is convex for every x, since for arbitrary $z_1, z_2 \in f(x)$, $t \in (0,1)$, we have

$$tz_1 + (1-t)z_2 \in tf(x) + (1-t)f(x) \subseteq f(tx + (1-t)x) = f(x)$$

where the last containment comes from the convexity of f. □

PROPOSITION 5.2.13. *If $f : X \to \mathscr{G}(C)$ and $g : X \to \mathscr{G}(D)$ are convex, then*

$$\inf_{(\mathscr{G}(C), \subseteq)} \{f(x) | x \in X, 0 \in g(x)\} = \text{cl} \bigcup \{f(x) | x \in X, 0 \in g(x)\},$$

so that the convex hull can be removed from the definition of infimum.

PROOF. We want to show that

$$\bigcup \{f(x) | x \in X, 0 \in g(x)\}$$

is convex. We begin by showing that $\{x \in X | 0 \in g(x)\}$ is convex. If 0 is contained in both $g(x_1)$ and $g(x_2)$, then for any $t \in (0,1)$,

$$0 \in tg(x_1) + (1-t)g(x_2) \subseteq g(tx_1 + (1-t)x_2),$$

so that $\{x \in X | 0 \in g(x)\}$ is convex.

Next, assume that $z_1, z_2 \in \bigcup\{f(x) | x \in X, 0 \in g(x)\}$. Then there are x_1, x_2 such that $z_1 \in f(x_1)$ and $z_2 \in f(x_2)$, with $0 \in g(x_1) \cap g(x_2)$. Thus for any $t \in (0,1)$

$$tz_1 + (1-t)z_2 \in tf(x_1) + (1-t)f(x_2) \subseteq f(tx_1 + (1-t)x_2).$$

Our initial claim gives that $0 \in g(tx_1 + (1-t)x_2)$, so that $\bigcup\{f(x) | x \in X, 0 \in g(x)$ is convex and we have our result. □

5.3. Problem Formulation

In this section, we explicitly formulate the multi-period utility maximization problem.

We consider a function $U(x) : \mathbb{R}^d \to \mathbb{R}^d$ which models the utility of an agent's assets x at the terminal time T. We make the following assumptions on U.

(1) U is a vector valued component-wise function

$$U(x) = (u_1(x_1), u_2(x_2), ..., u_d(x_d)), \;\; x = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$$

where each $u_i : \mathbb{R} \to \mathbb{R}$. Note each u_i is real-valued, as opposed to extended real valued. Thus U is defined even in the case of negative wealth.

(2) Each u_i is strictly concave, strictly increasing, and differentiable.

(3) Marginal utility tends to zero when wealth tends to infinity, so that

$$\lim_{x_i \to \infty} u_i'(x_i) = 0.$$

(4) u_i satisfies the *Inada condition*, so that the marginal utility tends to infinity when x_i tends to the infimum of the domain of u_i. In other words,

$$\lim_{x_i \to -\infty} u_i'(x) = \infty.$$

These assumptions are standard in the context of utility maximization problems [DS06].

Let the ordering cone $C = \mathbb{R}^d_+$. We define the objective function $F : L^p(\mathscr{F}_T, \mathbb{R}^d) \to \mathscr{G}(C)$ to be the expected utility of a random portfolio at terminal time.

$$F(x) = \mathbb{E}[-U(x)] + \mathbb{R}^d_+.$$

The expectation is taken with respect to the probability space $(\Omega, \mathscr{F}_T, \mathbb{P})$.

Note that in the definition of F, we have recast the utility maximization problem into a minimization framework. This is to establish more consistency with the set-valued optimization tools developed in [HHL$^+$ng], [Ham09], and [HL14], which cast their results in the traditional minimization framework of convex analysis. Of course, one could consider the maximization form of the problem without any loss of generality.

The portfolio optimization problem then takes the form

$$\text{minimize } F(x)$$

subject to the constraint that the portfolio x is the terminal result of a self-financing portfolio with initial endowment x_0. In other words, we have the problem

$$(\mathscr{P}) \qquad \begin{aligned} &\text{minimize } F(x) \\ &\text{subject to } x \in A_T(x_0) \end{aligned}$$

5.4. Duality in Set Optimization

In this section we recall the necessary results from set-valued duality [HL14] which we will use to prove our main result. ($\mathscr{P}$).

Set-valued Lagrange duality follows a similar theme to the real-valued case. Given convex cones $C \subseteq Z$ and $D \subseteq Y$, and convex functions $f : X \to \mathscr{G}(C) \subseteq \mathscr{P}(Z)$ and $g : X \to \mathscr{G}(D) \subseteq \mathscr{P}(Y)$, we are interested in the primal problem

$$(\mathscr{P}_e) \qquad \text{minimize } f(x) \text{ subject to } 0 \in g(x).$$

i.e. we search for a set $\bar{p} \subseteq Z$ where

$$\overline{p} := \inf_{(\mathscr{G}(C),\supseteq)} \{f(x)|x \in X, 0 \in g(x)\} = \text{cl conv} \bigcup \{f(x)|x \in X, 0 \in g(x)\}.$$

The first step is to define a set-valued Lagrangian function which recovers the objective, in the sense that the function $f(x)$ is the supremum of the Lagrangian over the set of dual variables.

For $y^* \in Y^*$ and $z^* \in Z^*$, where Y^* and Z^* denote the topological dual spaces of Y and Z, respectively, define the set-valued function $S_{(Y^*,Z^*)} : Y \to \mathscr{P}(Z)$ by

$$S_{(y^*,z^*)}(y) = \{z \in Z |\ y^*(y) \leqslant z^*(z)\}.$$

We use these functions to formulate a Lagrangian function.

DEFINITION 5.4.1. We define the Lagrangian $l : X \times Y^* \times C^+\backslash\{0\} \to \mathscr{G}(C)$ of the problem $(\mathscr{P}_e)$ by

$$l(x,y^*,z^*) = f(x) \oplus \bigcup_{y \in g(x)} S_{(y^*,z^*)}(y) = f(x) \oplus \inf\{S_{(y^*,z^*)}(y)|y \in g(x)\}.$$

We can recover the primal objective from the Lagrangian.

THEOREM 5.4.2. *[**HL14**][Prop 2.1] If $f(x) \neq Z$ for each $x \in X$, then*

$$\sup_{(y^*,z^*)\in Y^*\times C^+\backslash\{0\}} l(x,y^*,z^*) = \bigcap_{(y^*,z^*)\in Y^*\times C^+\backslash\{0\}} l(x,y^*,z^*) = \begin{cases} f(x): & 0 \in g(x) \\ \varnothing: & \textit{otherwise} \end{cases}$$

Under the condition in Theorem 5.4.2, we can formulate the problem $(\mathscr{P}_e)$ as

$$\inf_{x\in X} \sup_{(y^*,z^*)\in Y^*\times C^+\backslash\{0\}} l(x,y^*,z^*).$$

We define the dual problem

$$(\mathscr{D}_e) \qquad \sup_{(y^*,z^*)\in Y^*\times C^+\backslash\{0\}} \inf_{x\in X} l(x,y^*,z^*).$$

We denote by $h : Y^* \times C^+\backslash\{0\} \to G(C)$ the dual objective

$$h(y^*,z^*) := \inf_{x\in X} l(x,y^*,z^*) \tag{5.4}$$

and define $\overline{d}$ to be the set

$$\overline{d} := \sup_{(y^*,z^*)\in Y^*\times C^+\setminus\{0\}} h(y^*,z^*).$$

We have the following weak duality results for the problems ($\mathscr{P}_e$) and ($\mathscr{D}_e$).

PROPOSITION 5.4.3. *[**HL14**, Prop 6.2] Weak duality always holds for the problems ($\mathscr{P}_e$) and ($\mathscr{D}_e$). That is,*

$$\overline{d} = \sup\left\{h(y^*,z^*)|y^*\in Y^*, z^*\in C^+\setminus\{0\}\right\} \supseteq \inf\{f(x)|x\in X, 0\in g(x)\} = \overline{p}.$$

Strong duality, on the other hand, requires a constraint qualification. The problem ($\mathscr{P}_e$) is said to satisfy the *Slater condition* if

$$\exists \overline{x}\in \text{ dom } f: \ \ g(\overline{x})\cap \text{int}(-D)\neq\varnothing.$$

Slater's condition is sufficient for strong duality between ($\mathscr{P}_e$) and ($\mathscr{D}_e$).

THEOREM 5.4.4. *[**HL14**, Theorem 6.1] Assume $p\neq Z$. If $f: X\to\mathscr{G}(C)$ and $g: X\to\mathscr{G}(D)$ are convex and the Slater condition for problem ($\mathscr{P}_e$) is satisfied, then strong duality holds for ($\mathscr{P}_e$). That is,*

$$\overline{p} = \inf\{f(x)|0\in g(x)\} = \sup\left\{h(y^*,z^*)|y^*\in Y^*, z^*\in C^+\setminus\{0\}\right\} = \overline{d}.$$

Lastly, we introduce the notion of a set-valued Fenchel conjugate.

DEFINITION 5.4.5. The (negative) Fenchel conjugate of a function $f: X\to\mathscr{P}(Z)$ is the function $-f^*: X^*\times(C^+\setminus\{0\})\to\mathscr{G}(C)$ defined by

$$-f^*(x^*,z^*) = \text{cl}\bigcup_{x\in X}\left[f(x)+S_{(x^*,z^*)}(-x)\right].$$

Motivation for this definition and further details about the nature of the set-valued Fenchel conjugate can be found in [**Ham09**].

5.5. Duality in Portfolio Optimization

In this section we apply the tools introduced in the previous section for dualizing set-valued optimization problems to the portfolio optimization problem ($\mathscr{P}$).

We begin by showing that ($\mathscr{P}$) is well-defined.

LEMMA 5.5.1. *The functions* $F : L^P(\mathscr{F}_T, \mathbb{R}^d) \to \mathscr{G}(\mathbb{R}^d, \mathbb{R}^d_+)$, $F(x) = \mathbb{E}[-U(x)] + \mathbb{R}^d_+$ *and* $g : L^P(\mathscr{F}_T, \mathbb{R}^d) \to \mathscr{G}(L^p(\mathscr{F}_T, \mathbb{R}^d), L^p(\mathscr{F}_T, \mathbb{R}^d_+))$, $g(x) = x - A_T(x_0)$ *are well-defined and convex.*

PROOF. We begin with the function F. F clearly maps to $\mathscr{G}(\mathbb{R}^d, \mathbb{R}^d_+)$ because

$$F(x) + \mathbb{R}^d_+ = \mathbb{E}[-U(x)] + \mathbb{R}^d_+$$

is a polyhedral convex cone, and hence is a closed convex cone. We claim that F is also a convex map. More precisely, let $x_1, x_2 \in \mathbb{R}^d$, and $t \in (0,1)$. Then

$$\begin{aligned} & tf(x_1) + (1-t)f(x_2) \\ & = t\left(\mathbb{E}[-U(x_1)] + \mathbb{R}^+_d\right) + (1-t)\left(\mathbb{E}[-U(x_2)] + \mathbb{R}^+_d\right) \\ & = \mathbb{E}\left[t\left(-U(x_1)\right) + (1-t)\left(-U(x_2)\right)\right] + \mathbb{R}^+_d. \end{aligned}$$

By the assumptions on our objective function U, for each $1 \leqslant i \leqslant d$, $-u_i$ is convex, so that

$$t\left(-u_i(x_1(\omega))\right) + (1-t)\left(-u_i(x_2(\omega))\right) \geqslant -u_i\left(tx_1(\omega) + (1-t)x_2(\omega)\right)$$

for each $\omega \in \Omega$. It follows that

$$\mathbb{E}\left[t\left(-U(x_1)\right) + (1-t)\left(-U(x_2)\right)\right] + \mathbb{R}^+_d \subseteq \mathbb{E}\left[-U(tx_1 + (1-t)x_2)\right] + \mathbb{R}^+_d.$$

We conclude that F is convex by Definition 5.2.11.

Next we consider the function $g(x)$. We need to show that in $L^p(\mathscr{F}_T, \mathbb{R}^d)$,

$$\mathrm{cl}\,\mathrm{conv}\left(x - A_T(x_0) + L^p(\mathscr{F}_T, \mathbb{R}^d_+)\right) = x - A_T(x_0).$$

Observe that $-A_T(x_0)$ and $L^p(\mathscr{F}_T, \mathbb{R}^d_+)$ are convex, so their sum is as well [**Roc97**, Ch. 3] and the convex hull on the left side can be dropped. In addition, since K_T is a solvency cone, the cone $L^p(\mathscr{F}_T, \mathbb{R}^d_+)$ is contained in $L^p(\mathscr{F}_T, K_T)$, thus $x - A_T(x_0) + L^p(\mathscr{F}_T, \mathbb{R}^d_+) = x - A_T(x_0)$. Hence, it remains to show that $x - A_T(x_0)$ is closed in $L^p(\mathscr{F}_T, \mathbb{R}^d)$. By the assumptions (5.2.4), Ω is finite, and each element in $\Omega = \omega_1, \ldots, \omega_N$ has positive probability. The space of d-dimensional random variables can then be associated with Euclidean space of dimension $d \times N$ and inner product $\mathbb{E}\langle x, y\rangle$. Note that if $G = cone(\xi_1, ..., \xi_m)$ is a random convex cone generated by m $\mathscr{F}_t$-measurable random variables, then $L^p(G, \mathscr{F}_t)$ is the polyhedral

convex cone generated by $\xi_i I_{\Gamma_j}$ where $\{\Gamma_j\}_{j\in J}$ are the atoms of $\mathscr{F}_t$. We have established that each of the $L^p(\mathscr{F}_t, K_t)$ is finitely generated, so by the Farkas-Minkowski-Weyl Theorem, each is polyhedral. Since the finite sum of polyhedral cones is a polyhedral cone, we conclude that $x - A_T(x_0)$ is a polyhedral cone, and hence is closed. □

Note that, in the notation of the previous section, we have established that $X = L^p(\mathscr{F}_T, \mathbb{R}^d)$, $Y = L^p(\mathscr{F}_T, \mathbb{R}^d)$, $Z = \mathbb{R}^d$, $C = \mathbb{R}^d_+$, and $D = L^p(\mathscr{F}_T, \mathbb{R}^d_+)$.

We are now ready to state our main result, which is a formulation of the dual problem to the portfolio optimization problem ($\mathscr{P}$). We then examine the relationship between the primal and dual problems. Namely, we establish that strong duality holds.

THEOREM 5.5.2. *The dual problem to ($\mathscr{P}$) is the problem*

$$(\mathscr{D}) \qquad \sup_{(y^*,z^*)\in -A_T(x_0)^+ \times \mathbb{R}^d_+ \setminus \{0\}} h(y^*, z^*)$$

where $h : L^q(\mathscr{F}_T, \mathbb{R}^d) \times \mathbb{R}^d_+ \setminus \{0\} \to \mathscr{G}(\mathbb{R}^d_+)$ *is defined as*

$$h(y^*,z^*) = \begin{cases} \{z \in \mathbb{R}^d \,|\, \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} z^*(\mathbb{E}[-U(x)]) + y^*(x) \leqslant z^*(z)\} & \text{if } y^* \in (-A_T(x_0))^+ \\ \mathbb{R}^d & \text{otherwise.} \end{cases} \tag{5.5}$$

When $y^* \in -A_T(x_0)^+$ *and no components of* z^* *are zero, we can write the function* $h(y^*, z^*)$ *as*

$$\left\{ z \in \mathbb{R}^d | \mathbb{E}\left[\sum_{i=1}^d z_i^* \mathbb{1} u_i^* \left(\frac{y_i^*}{z_i^* \mathbb{1}} \right) \right] \leqslant z^* z \right\} \tag{5.6}$$

where u_i^* *denotes the concave conjugate of* u_i *[**Roc97**].*

To prove this result, we require the following lemma

LEMMA 5.5.3. *The lagrangian function for the problem ($\mathscr{P}$) is* $l : L^p(\mathscr{F}_T, \mathbb{R}^d) \times L^q(\mathscr{F}_T, \mathbb{R}^d) \times (\mathbb{R}^d_+ \setminus \{0\}) \to \mathscr{G}(\mathbb{R}^d, \mathbb{R}^d_+)$ *defined by*

$$l(x,y^*,z^*) = \begin{cases} \{z \in \mathbb{R}^d \,|\, z^*(\mathbb{E}[-U(x)]) + y^*(x) \leqslant z^*(z)\} & \text{if } y^* \in -A_T(x_0)^+ \\ \mathbb{R}^d & \text{otherwise.} \end{cases} \tag{5.7}$$

PROOF. Note that the positive dual cones of $L^p(\mathscr{F}_T,\mathbb{R}^d_+)$ and $\mathbb{R}^d_+$ are $L^q(\mathscr{F}_T,\mathbb{R}^d_+)$ and $\mathbb{R}^d_+$ respectively. Hence the Lagrangian function l has domain on $L^p(\mathscr{F}_T,\mathbb{R}^d)\times L^q(\mathscr{F}_T,\mathbb{R}^d)\times(\mathbb{R}^d_+\backslash\{0\})$, as per Definition 5.4.1.

Also from Definition 5.4.1, we see that

$$l(x,y^*,z^*) = F(x)\oplus \bigcup_{y\in x-A_T(x_0)} S_{y^*,z^*}(y) \tag{5.8}$$

The union on the right side can be written explicitly

$$\begin{aligned}\bigcup_{y\in x-A_T(x_0)}\{z\in\mathbb{R}^d\,|\,y^*(y)\leqslant z^*(z)\} &= \{z\in\mathbb{R}^d\,|\inf_{y\in x-A_T(x_0)} y^*(y)\leqslant z^*(z)\}\\ &= \{z\in\mathbb{R}^d\,|\inf_{y\in -A_T(x_0)} y^*(y)+y^*(x)\leqslant z^*(z)\}\,.\end{aligned}$$

Note that $A_T(x_0)=x_0-K_0-L^p(\mathscr{F}_1,K_1)-\cdots-L^p(\mathscr{F}_T,K_T)$ is a cone in $L^p(\mathscr{F}_T,K_T)$. The infimum on the right side is the support function on a cone [**Roc97**], and can be written

$$\inf_{y\in -A_T(x_0)} y^*(y) = -\delta_{-A_T(x_0)^+}(y^*)\,,$$

where $\delta_{-A_T(x_0)^+}(y^*)$ is the indicator function on $-A_T(x_0)^+$, equal 0 if y^* belongs to $-A_T^+$ and ∞ otherwise. Hence, identity (5.8) becomes

$$\begin{aligned}l(x,y^*,z^*) &= \mathbb{E}[-U(x)]+\{z\in\mathbb{R}^d\,|\,-\delta_{-A_T(x_0)^+}(y^*)+y^*(x)\leqslant z^*(z)\}\oplus\mathbb{R}^d_+\\ &= \{z\in\mathbb{R}^d\,|\,-\delta_{-A_T(x_0)^+}(y^*)+z^*(\mathbb{E}[-U(x)])+y^*(x)\leqslant z^*(z)\}\oplus\mathbb{R}^d_+\end{aligned}$$

Since $z^*\in\mathbb{R}^d_+\backslash\{0\}$,

$$\{z\in\mathbb{R}^d\,|\,a\leqslant z^*(z)\}\oplus\mathbb{R}^d_+ = \{z\in\mathbb{R}^d\,|\,a\leqslant z^*(z)\}$$

for any constant a in $\mathbb{R}\cup\{-\infty\}$. It follows that

$$l(x,y^*,z^*) = \{z\in\mathbb{R}^d\,|\,-\delta_{(-A_T(x_0))^+}(y^*)+z^*(\mathbb{E}[-U(x)])+y^*(x)\leqslant z^*(z)\}$$

which deduces identity (5.7). □

Recall that the coordinate functions $u_i(x)$ of the utility function $U(x)$ are real-valued for each $x \in \mathbb{R}^d$. Combining this with the fact that Ω is finite (and hence the expectation in the problem formulation is finite) yields the following proposition.

PROPOSITION 5.5.4. *The objective function of* ($\mathscr{P}$) *can be recovered from the lagrangian* (5.7)*. That is,*

$$\sup_{(y^*,z^*)\in L^q(\mathscr{F}_T,\mathbb{R}^d)\times\mathbb{R}^d_+\setminus\{0\}} l(x,y^*,z^*) = \begin{cases} \mathbb{E}[-U(x)]+\mathbb{R}^d_+ : & 0\in x-A_T(x_0) \\ \varnothing : & \text{otherwise} \end{cases}$$

PROOF. This follows immediately from the above comments and an application of Theorem 5.4.2. □

Using Lemma 5.5.3, we complete the proof of Theorem 5.5.2.

PROOF. From the definition of the dual objective (5.4)

$$h(y^*,z^*) = \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} l(x,y^*,z^*).$$

If $y^* \notin -A_T(x_0)^+$, this is $\mathbb{R}^d$. In the case that $y^* \in -A_T(x_0)^+$

$$\begin{aligned} &\inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} l(x,y^*,z^*) \\ &= \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \{z\in\mathbb{R}^d \mid -\delta_{(-A_T(x_0))^+}(y^*)+z^*(\mathbb{E}[-U(x)])+y^*(x)\leqslant z^*(z)\} \\ &= \mathrm{cl} \bigcup_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \{z | y^*(x)+z^*(\mathbb{E}[-U(x)])\leqslant z^*(z)\} \\ &= \mathrm{cl}\,\{z| \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} z^*(\mathbb{E}[-U(x)])+y^*(x)\leqslant z^*(z)\}. \end{aligned}$$

The infimum in the expression above is the Fenchel conjugate of a sum of convex functions. Since the Fenchel conjugate of a proper convex function is proper and lower semicontinuous [**RWW10**, Theorem 11.1], this infimum is attained, so we can drop the closure from the expression. Thus,

$$= \{z| \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} z^*(\mathbb{E}[-U(x)])+y^*(x)\leqslant z^*(z)\}$$

and the result is proven.

The final part of the theorem, in which we reformulate the dual objective in terms of concave conjugates, follows because

$$\begin{aligned}
&\inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} z^*(\mathbb{E}[-U(x)]) + y^*(x) \\
&= \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{E}[\langle z^*\mathbb{1}, -U(x)\rangle] + y^*(x) \\
&= \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{E}[\langle z^*\mathbb{1}, -U(x)\rangle + \langle y^*, x\rangle] \\
&= \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{E}[\sum_{i=1}^d -z_i^* u_i(x_i) + y_i^* x_i] \\
&= \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \sum_{\omega\in\Omega} \mathbb{P}[\omega]\left(\sum_{i=1}^d -z_i^* u_i(x_i(\omega)) + y_i^*(\omega)x_i(\omega)\right)
\end{aligned}$$

Exploiting separability over the sum gives

$$\begin{aligned}
&= \sum_{\omega\in\Omega} \mathbb{P}[\omega]\left(\sum_{i=1}^d \inf_{x_i(\omega)\in\mathbb{R}} -z_i^* u_i(x_i(\omega)) + y_i^*(\omega)x_i(\omega)\right) \\
&= \mathbb{E}\left[\sum_{i=1}^d \inf_{x_i\in L^p(\mathscr{F},\mathbb{R})} y_i^*(x_i) - z_i^*\mathbb{1}u_i(x_i)\right]
\end{aligned}$$

from which the result follows immediately. More details can be found in the next section, where we perform the details of this calculation more slowly with an example problem for context.

□

In the language of set-valued Fenchel conjugates, we have the following easy corollary.

COROLLARY 5.5.5. *The objective function of the dual problem is*

$$h(y^*,z^*) = \begin{cases} -F^*(-y^*,z^*) & \text{if } y^* \in -A_T(x_0)^+ \\ \mathbb{R}^d & \text{otherwise.} \end{cases} \tag{5.9}$$

where F^* is the Fenchel conjugate of F.

PROOF. We compute the Fenchel conjugate of F. For every $y^* \in L^q(\mathscr{F}_T, \mathbb{R}^d)$, $z^* \in \mathbb{R}^d_+ \backslash \{0\}$, it follows from Definition 5.4.5 that

$$\begin{aligned} -F^*(-y^*, z^*) &= \operatorname{cl} \bigcup_{x \in X} \big(F(x) + S_{(-y^*, z^*)}(-x)\big) \\ &= \operatorname{cl} \bigcup_{x \in L^p(\mathscr{F}_T, \mathbb{R}^d)} \big(\mathbb{E}[-U(x)] + \mathbb{R}^d_+ + \{z | -y^*(-x) \leqslant z^*(z)\}\big) \\ &= \operatorname{cl} \bigcup_{x \in L^p(\mathscr{F}_T, \mathbb{R}^d)} \big\{z + \mathbb{E}[-U(x)] + \mathbb{R}^d_+ | -y^*(-x) \leqslant z^*(z)\big\} \end{aligned} \tag{5.10}$$

(5.11)

Since $z^* \in \mathbb{R}^d_+ \backslash \{0\}$, $z^*(r) \geqslant 0$ for each $r \in \mathbb{R}^d_+$, (5.10) becomes

$$\begin{aligned} &= \operatorname{cl} \bigcup_{x \in L^p(\mathscr{F}_T, \mathbb{R}^d)} \{z | y^*(x) \leqslant z^*(z - \mathbb{E}[-U(x)]\} \\ &= \operatorname{cl} \left\{ z | \inf_{x \in L^p(\mathscr{F}_T, \mathbb{R}^d)} y^*(x) + z^*(\mathbb{E}[-U(x)]) \leqslant z^*(z) \right\} \\ &= \left\{ z | \inf_{x \in L^p(\mathscr{F}_T, \mathbb{R}^d)} y^*(x) + z^*(\mathbb{E}[-U(x)]) \leqslant z^*(z) \right\}. \end{aligned}$$

This agrees with (5.5). □

THEOREM 5.5.6. *Strong duality holds between the problems ($\mathscr{P}$) and ($\mathscr{D}$). That is,*

$$\begin{aligned} \overline{p} = \inf F(x) = & \qquad\qquad \sup(y^*, z^*) = \overline{d} \\ \textit{subject to } x \in A_T(x_0) & \qquad \textit{subject to } y^* \in -A_T(x_0)^+ \\ & \qquad\qquad z^* \in \mathbb{R}^d_+ \backslash \{0\}. \end{aligned}$$

PROOF. By Theorem 5.4.4 and Lemma 5.5.1, it suffices to show that $\overline{p} = \inf_{x \in A_T(x_0)} F(x) \neq \mathbb{R}^d$ and that Slater's condition is satisfied.

For the first part, we use weak duality. By Proposition 5.4.3, $\overline{p} \subseteq \overline{d}$, so it suffices to show that $\overline{d} \neq \mathbb{R}^d$. Lemma 5.2.9 give that $-A_T(x_0)^+ \cap \operatorname{int}(L^P(\mathscr{F}_T, \mathbb{R}^d_+)$ is nonempty, so there exists $\tilde{y}^* \in -A_T(x_0)^+$ with

$\tilde{y}^*(\omega)_i < 0$ for each $\omega \in \Omega$, $1 \leqslant i \leqslant d$. Then

$$\overline{p} \subseteq \sup_{(y^*,z^*)\in -A_T(x_0)^+ \times z^* \in \mathbb{R}^d_+ \setminus \{0\}} h(y^*,z^*) \subseteq \{z | \inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{1}_d(\mathbb{E}[-U(x)]) + y_0^*(x) \leqslant \mathbb{1}_d(z)\}.$$

The last containment follows by taking $y^* = \tilde{y}^*$ and $z^* = \mathbb{1}_d$, the d-dimensional vector consisting of all ones. So it suffices to show that

$$\inf_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{1}_d(\mathbb{E}[-U(x)]) - \tilde{y}^*(x) > -\infty,$$

which is equivalent to

$$\sup_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \tilde{y}^*(x) - \mathbb{1}_d(\mathbb{E}[-U(x)]) < \infty.$$

Note that the left hand side of this expression is simply the Fenchel-Conjugate of the function $\mathbb{1}_d(\mathbb{E}[-U(x)])$. We have

$$\begin{aligned}
&\sup_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \tilde{y}^*(x) - \mathbb{1}_d(\mathbb{E}[-U(x)]) \\
&= \sup_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \mathbb{E}[\langle \tilde{y}^*(\omega), x(\omega)\rangle] - \mathbb{E}[\langle \mathbb{1}, -U(x(\omega))\rangle] \\
&= \sup_{x\in L^p(\mathscr{F}_T,\mathbb{R}^d)} \sum_{\omega=\omega_1}^{\omega_n} \mathbb{P}[\omega] \left[\sum_{i=1}^{d} \tilde{y}_i^*(\omega) x_i(\omega) - (-u_i(x_i(\omega))) \right] \\
&= \sum_{i=1}^{d} \sum_{\omega=\omega_1}^{\omega_n} \mathbb{P}[\omega] \left[\sup_{x_i(\omega)} \tilde{y}_i^*(\omega) - (-u_i(x_i(\omega))) \right]. \qquad (5.12)
\end{aligned}$$

The first equality follows from the definition of the inner product in $L^p(\mathscr{F}_T,\mathbb{R}^d)$. The second and third come from the finiteness of Ω and the separability of the expression, respectively.

Since each u_i' is continuous with range $(-\infty,0)$, and each $\tilde{y}_i^*(\omega) < 0$, the intermediate value theorem gives that, for each $\omega \in \Omega$, $1 \leqslant i \leqslant d$, there exists $\tilde{x}_i(\omega)$ such that $u_i'(\tilde{x}_i(\omega)) = \tilde{y}_i^*(\omega)$. By [**Roc97**, Theorem 23.5],

$$\sup_{x(\omega)_i} \tilde{y}_i^*(\omega) - (-u_i(\tilde{x}_i(\omega)))$$

achieves its supremal value at $\tilde{x}_i(\omega)$. It follows that (5.12) is finite, from which we conclude that $\overline{p} \neq \mathbb{R}^d$.

Next we show that Slater's condition is satisfied. We want to find an $\bar{x} \in \operatorname{dom} F$ such that $x - A_T(x_0) \cap \operatorname{int}(-L^p(\mathscr{F}_T,\mathbb{R}^d_+)) \neq \varnothing$. Recall from the problem formulation that $\operatorname{dom} F = L^p(\mathscr{F}_T,\mathbb{R}^d)$, so the first part

of Slater's condition is not a restriction. Note that since

$$A_T(x_0) = x_0\mathbb{1} - K_0\mathbb{1} - L^p(\mathscr{F}_1, K_1) - \ldots - L^p(\mathscr{F}_T, K_T)$$

where each K_t is a solvency cone, $x_0\mathbb{1} \in A_T(x_0)$. Then, choose $\bar{x}$ such that $\bar{x}_i(\omega) < (x_0)_i$ for each component $1 \leqslant i \leqslant d$ and for all $\omega \in \Omega$. We have that $\bar{x} - x_0\mathbb{1} \in x - A_T(x_0)$ and also $\bar{x} - x_0\mathbb{1} \in \text{int}(-L^p(\mathscr{F}_T, \mathbb{R}^d_+))$, so Slater's condition is satisfied. □

5.6. An Example

In this final section, we explore an example which we hope will help to illustrate the theoretical results from previous sections.

We consider a market with 2 assets and 3 time steps, so that the time step t ranges from 0 to 3. The probability space $\Omega = \times_{i=1}^{3}\{-1,0,1\}$, with the probability measure $\mathbb{P}$ defined uniformly on this set. In other words, the possible outcomes ω are defined as a tuple $\omega = (\omega_1, \omega_2, \omega_3)$, $\omega_1, \omega_2, \omega_3 \in \{-1,0,1\}$. From the decision maker's perspective, we have that at each time step t the random variable taking values ω_t becomes known. Thus the filtration $((F_t)_{t=0}^3)$ is defined by $\mathscr{F}_t = \sigma(\omega_i|_{i\leqslant t})$, the sigma algebra generated by these random variables. We also take $\mathscr{F} = \mathscr{F}_3 = \sigma(\omega_1, \omega_2, \omega_3)$, the sigma-algebra of full information.

The bid-ask process $(\Pi_t)_{t=0}^T$ is defined as follows:

$$\Pi_t(\omega) = \begin{bmatrix} 1 & 1 \\ 8 \cdot 2^{\sum_{i\leqslant t}\omega_i} & 1 \end{bmatrix}.$$

This is obviously $(\mathscr{F}_t)_{t=0}^3$ adapted, and one can also easily check that the properties of bid-ask matrix are satisfied for each realization. The solvency cones generated by this process are

$$K_t(\omega) = \text{cone}\{(1,-1), (-1, 8\cdot 2^{\sum_{i\leqslant t}\omega_i})\} = \{(x,y)|x+y \geqslant 0, 8\cdot 2^{\sum_{i\leqslant t}\omega_i}x + y \geqslant 0\}.$$

Figure (5.1) illustrates these cones for various times and realizations of ω.

We define our vector-valued objective function to be

$$U(x) = (-e^{-x_1}, -e^{-x_2})^T$$

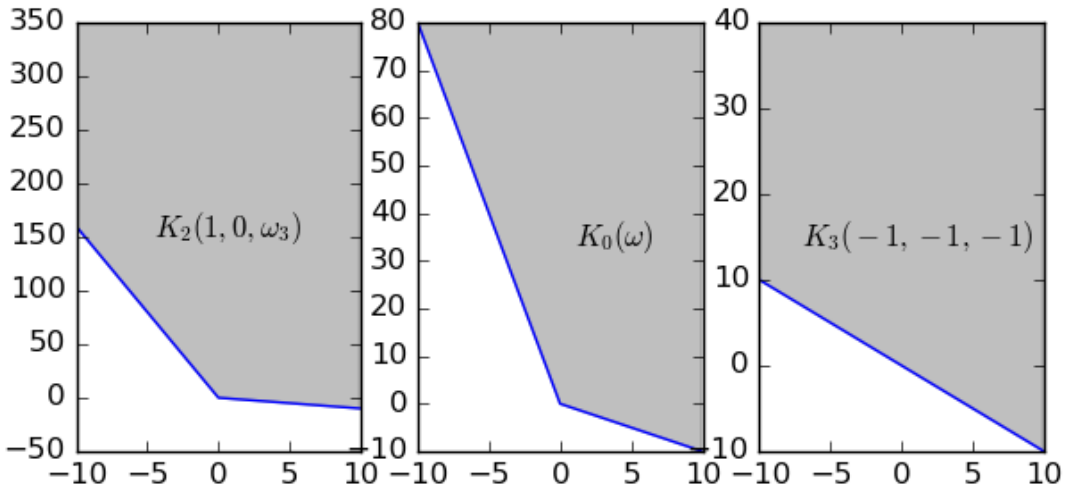


FIGURE 5.1. The solvency cones $K_t(\omega)$ for various t and ω.

where $x = (x_1, x_2)^T$ is the quantity of physical assets we have at terminal time. Our set-valued objective function to be minimized is then

$$F(x) = \mathbb{E}\left[\left(e^{-x_1}, e^{-x_2}\right)^T\right] + \mathbb{R}^2_+.$$

We assume that our initial endowment is $(0,0)^T$. The set of self-financing portfolios is

$$A_3 = -K_0\mathbb{1} - L^p(\sigma(\omega_1); K_1) - L^p(\sigma(\omega_1, \omega_2), K_2) - L^p(\sigma(\omega_1, \omega_2, \omega_3), K_3)$$

where $K_t(\omega)$ are given as above.

We can then formulate the primal portfolio optimization problem as

$$(\mathscr{P}_{ex}) \qquad \begin{aligned} &\text{minimize } \mathbb{E}\left[\left(e^{-x_1}, e^{-x_2}\right)^T\right] + \mathbb{R}^2_+ \\ &\text{subject to } x \in A_3 \end{aligned}$$

According to theorem (5.5.2), the dual problem is then

$$(\mathscr{D}_{ex}) \qquad \begin{aligned} &\sup\{z \mid \inf_{x \in L^p(\mathbb{R}^2, \mathscr{F}_3)} z^*(\mathbb{E}\left[(-e^{x_1}, -e^{x_2})^T\right]) + y^*(x) \leqslant z^*(z)\} \\ &\text{subject to } y^* \in -A_3^+ \\ &z^* \in \mathbb{R}^2_+ \setminus \{0\} \end{aligned}$$

First, we investigate the nature of the set $-A_3^+$. From [**Roc97**][Cor 16.4.2],

$$-A_3^+ = (K_0\mathbb{1} + L^p(\mathscr{F}_1, K_1) + L^p(\mathscr{F}_2, K_2) + L^p(\mathscr{F}_3, K_3))^+$$

$$= \bigcap_{i=0}^{3} (L^p(\mathscr{F}_i, K_i))^+.$$

In other words, $y^* \in L^q(\mathscr{F}, \mathbb{R}^2)$ is in $-A_3^+$ when $y(\omega) \in K_t(\omega)^+$ for each $t = 0, ..., 3$.

We can explicitly compute the cones $K_t(\omega)^+$. We have that

$$K_t(\omega) = \text{cone}\left(\text{conv}\left\{\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 8 \cdot 2^{\sum_{i \leqslant t} \omega_i} \end{pmatrix}\right\}\right)^+.$$

Hence the dual cone is

$$K_t(\omega)^+ = \text{cone}\left(\text{conv}\left\{\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -8 \cdot 2^{\sum_{i \leqslant t} \omega_i} \\ -1 \end{pmatrix}\right\}\right).$$

Next we take the intersection of these cones to form $A_3^+(\omega)$. For a fixed ω, let $s(\omega) = \min_{j=0,1,2,3} \sum_{i=1}^{j} \omega_i$. Then

$$-A_3(\omega)^+ = \text{cone}\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -8 \cdot 2^{s(\omega)} \\ -1 \end{pmatrix}\right).$$

Figure (5.2) illustrates $K_t^+(\omega)$ for various times and realizations of ω.

Now we work with to simplify the dual problem ($\mathscr{D}_{ex}$). The objective function is

$$h(y^*, z^*) = \left\{ z \middle| \inf_{x \in L^p(\mathscr{F}, R^2)} z^*(\mathbb{E}\left[(e^{-x_1}, e^{-x_2})^T\right]) + y^*(x) \leqslant z^*(z) \right\}$$

$$= \left\{ z \middle| \inf_{x \in L^p(\mathscr{F}, R^2)} \mathbb{E}\left[z_1^* \mathbb{1} e^{-x_1} + z_2^* \mathbb{1} e^{-x_2}\right] + y^*(x) \leqslant z^*(z) \right\}$$

Recall that the nature of linear functionals $y^* \in L^q(\mathscr{F}, \mathbb{R}^d)$ is $y^*(x) = \mathbb{E}\langle y^*, x \rangle$. Hence the objective becomes

$$\left\{ z \middle| \inf_{x \in L^p(\mathscr{F}, R^2)} \mathbb{E}\left[z_1^* \mathbb{1} e^{-x_1} + z_2^* \mathbb{1} e^{-x_2} + y_1^* x_1 + y_2^* x_2\right] \leqslant z^*(z) \right\}$$

Using the fact that our probability space is finite, we expand the expectation

$$\left\{ z \middle| \inf_{x \in L^p(\mathscr{F}, \mathbb{R}^2)} \frac{1}{27} \sum_{\omega \in \Omega} z_1^* e^{-x_1(\omega)} + z_2^* e^{-x_2(\omega)} + y_1^*(\omega) x_1(\omega) + y_2^*(\omega) x_2(\omega) \leqslant z^*(z) \right\}$$

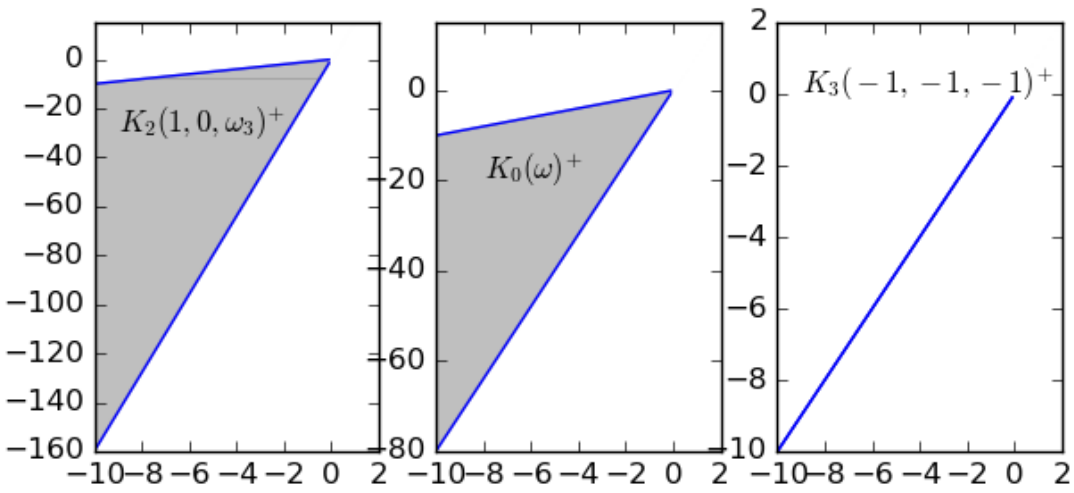


FIGURE 5.2. The positive polar cones $K_t(\omega)^+$ for various t and ω.

This infimum is separable over the $x_i(\omega)$ variables. Hence,

$$= \left\{ z | \frac{1}{27} \sum_{\omega \in \Omega} \inf_{x_1(\omega) \in \mathbb{R}} \{ z_1^* e^{-x_1(\omega)} + y_1^*(\omega) x_1(\omega) \} + \inf_{x_2(\omega) \in \mathbb{R}} z_2^* e^{-x_2(\omega)} + y_2^*(\omega) x_2(\omega) \leqslant z^*(z) \right\}$$

We can compute each of these infimums explicitly. Note that

$$\inf_{x_1(\omega) \in \mathbb{R}} z_1^* e^{-x_1(\omega)} + y_1^*(\omega) x_1(\omega) = -f^*(-y_1^*(\omega))$$

where $f(x) = z_1^* e^{-x}$ and f^* denotes the convex conjugate of f. Recall that for $g : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}$ (See [**Roc97**])

$$(ag(\cdot))^*(x^*) = ag^*(x^*/a)$$

$$(g(a\cdot))^*(x^*) = g^*(x^*/a)$$

$$g^*(x^*) = \begin{cases} x^* \ln(x^*) - x^* & \text{if } x^* > 0 \\ 0 & \text{if } x^* = 0 \\ \infty \text{ otherwise} \end{cases} \quad \text{when } g(x) = e^x.$$

Hence the objective function becomes

$$\left\{ z | \frac{1}{27} \sum_{\omega \in \Omega} y_1^*(\omega) - y_1^*(\omega) \ln \left(\frac{y_1(\omega)^*}{z_1^*} \right) + y_2^*(\omega) - y_2^*(\omega) \ln \left(\frac{y_2^*(\omega)}{z_2^*} \right) \leqslant z^*(z) \right\}$$

when $y_1(\omega), y_2(\omega), z_1, z_2 \neq 0$.

Therefore, when $y_1(\omega), y_2(\omega), z_1, z_2 \neq 0$, we have the following formulation of the dual problem: find the supremum of

$$\left\{ (z_1, z_2)^T \left| \frac{1}{27} \sum_{\omega \in \Omega} y_1^*(\omega) - y_1^*(\omega) \ln \left(\frac{y_1^*(\omega)}{z_1^*} \right) + y_2^*(\omega) - y_2^*(\omega) \ln \left(\frac{y_2^*(\omega)}{z_2^*} \right) \leqslant z_1^* z_1 + z_2^* z_2 \right. \right\}$$

$$\text{subject to } \begin{pmatrix} z_1^* \\ z_2^* \end{pmatrix} \in \mathbb{R}^2 \setminus \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$$

$$\begin{pmatrix} y_1^*(\omega) \\ y_2^*(\omega) \end{pmatrix} \in A_3(\omega) \ \ \forall \omega \in \Omega$$

When either z_1 or z_2 equals 0, we only consider x_2 or x_1, respectively, in our objective because the other terms vanish. Likewise, the case that $y_1(\omega) = 0$ eliminates the expressions with $y_1(\omega)$ in the objective, because of the conjugation result on the previous page. The case that $y_2(\omega) = 0$ is symmetric.

Bibliography

[ABP13] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, *Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory*, The Journal of Machine Learning Research **14** (2013), no. 1, 2689–2728.

[AT16] T. B. Arnold and R. J. Tibshirani, *Efficient implementations of the generalized lasso dual path algorithm*, Journal of Computational and Graphical Statistics **25** (2016), no. 1, 1–27.

[AW94] H. Attouch and R. J.-B. Wets, *Epigraphical processes: laws of large numbers for random lsc functions*, Dép. des Sciences Mathématiques, 1994.

[BB05] M. Bagnoli and T. Bergstrom, *Log-concave probability and its applications*, Economic theory **26** (2005), no. 2, 445–469.

[BD18] R. Bassett and J. Deride, *Maximum a posteriori estimators as a limit of bayes estimators*, Mathematical Programming (2018).

[Bee93] G. Beer, *Topologies on closed and closed convex sets*, Mathematics and Its Applications, Springer, 1993.

[Ber09] D. P. Bertsekas, *Convex optimization theory*, Athena Scientific Belmont, 2009.

[Ber14] ———, *Constrained optimization and lagrange multiplier methods*, Academic press, 2014.

[BL16] R. Bassett and K. Le, *Multistage portfolio optimization: A duality result in conic market models*, arXiv preprint arXiv:1601.00712 (2016).

[BM97] L. Birgé and P. Massart, *From model selection to adaptive estimation*, Festschrift for lucien le cam, Springer, 1997, pp. 55–87.

[BPC+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends® in Machine learning **3** (2011), no. 1, 1–122.

[BR06] L. Birgé and Y. Rozenholc, *How many bins should be put in a regular histogram*, ESAIM: Probability and Statistics **10** (2006), 24–45.

[BT09] A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Transactions on Image Processing **18** (2009), no. 11, 2419–2434.

[BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[Car18] S. Caron, *QPSolvers: Wrappers for quadratic programming in python*, https://github.com/stephane-caron/qpsolvers, 2016-2018.

[CDSS14] S.-O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun, *Efficient density estimation via piecewise polynomial approximation*, Proceedings of the 46th Annual ACM Symposium on Theory of Computing, ACM, 2014, pp. 604–613.

[Cor09] T. H. Cormen, *Introduction to algorithms*, MIT press, 2009.

[Cox64] H. Cox, *On the estimation of state variables and parameters for noisy dynamic systems*, Automatic Control, IEEE Transactions on **9** (1964), no. 1, 5–12.

[Dav77] M. Davis, *Linear estimation and stochastic control*, Chapman and Hall mathematics series, Chapman and Hall, 1977.

[DG85] L. Devroye and L. Györfi, *Nonparametric density estimation: The l1 view*, New York: John Wiley & Sons, 1985.

[DJD88] S. Dharmadhikari and K. Joag-Dev, *Unimodality, convexity, and applications*, Elsevier, 1988.

[DJKP96] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, *Density estimation by wavelet thresholding*, The Annals of Statistics (1996), 508–539.

[DM09] L. Denby and C. Mallows, *Variations on the histogram*, Journal of Computational and Graphical Statistics **18** (2009), no. 1, 21–31.

[DN90] M. H. Davis and A. R. Norman, *Portfolio selection with transaction costs*, Mathematics of Operations Research **15** (1990), no. 4, 676–713.

[DR] L. Dümbgen and K. Rufibach, *logcondens: Computations related to univariate log-concave density estimation*.

[DR+09] L. Dümbgen, K. Rufibach, et al., *Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency*, Bernoulli **15** (2009), no. 1, 40–68.

[DS06] F. Delbaen and W. Schachermayer, *The mathematics of arbitrage*, Springer Finance, Springer Berlin Heidelberg, 2006.

[DW88] J. Dupacová and R. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, The annals of statistics (1988), 1517–1549.

[FD81] D. Freedman and P. Diaconis, *On the histogram as a density estimator: L 2 theory*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **57** (1981), no. 4, 453–476.

[FHT] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1.

[Fig04] M. A. Figueiredo, *Lecture notes on bayesian estimation and classification*, Instituto de Telecomunicacoes-Instituto Superior Tecnico (2004).

[FST14] W. Fithian, D. Sun, and J. Taylor, *Optimal inference after model selection*, arXiv preprint arXiv:1410.2597 (2014).

[Gee00] S. A. Geer, *Empirical processes in m-estimation*, vol. 6, Cambridge university press, 2000.

[Gew05] J. Geweke, *Contemporary bayesian econometrics and statistics*, vol. 537, John Wiley & Sons, 2005.

[Gey94] C. J. Geyer, *On the asymptotics of constrained m-estimation*, The Annals of Statistics (1994), 1993–2010.

[GS02] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, International statistical review **70** (2002), no. 3, 419–435.

[Ham09] A. H. Hamel, *A duality theory for set-valued functions i: Fenchel conjugation theory*, Set-Valued and Variational Analysis (2009), 153–182.

[HHL$^+$ng] A. Hamel, F. Heyde, A. Lø"hne, B. Rudloff, and C. Schrage, *Set optimization-a rather short introduction*, Set Optimization and Applications in Finance, Springer PROMS series (Forthcoming).

[HL14] A. H. Hamel and A. Löhne, *Lagrange duality in set optimization*, J. Optim. Theory Appl. **161** (2014), no. 2, 368–397.

[HLD$^+$13] W. Hoegele, R. Loeschel, B. Dobler, O. Koelbl, and P. Zygmanski, *Bayesian estimation applied to stochastic localization with constraints due to interfaces and boundaries*, Mathematical Problems in Engineering **2013** (2013).

[HLR14] A. H. Hamel, A. Löhne, and B. Rudloff, *Benson type algorithms for linear vector optimization and applications*, Journal of Global Optimization **59** (2014), no. 4, 811–836.

[HW88] P. Hall and M. P. Wand, *Minimizing l1 distance in nonparametric density estimation*, Journal of Multivariate Analysis **26** (1988), no. 1, 59–88.

[Jar15] K. Jarosz, *Function spaces in analysis*, vol. 645, American Mathematical Society, 2015.

[Kab99] Y. Kabanov, *Hedging and liquidation under transaction costs in currency markets*, Finance and Stochastics **3** (1999), no. 2, 237–248.

[Kal60] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Transactions of the ASME–Journal of Basic Engineering **82** (1960), no. Series D, 35–45.

[KF00] K. Knight and W. Fu, *Asymptotics for lasso-type estimators*, Annals of statistics (2000), 1356–1378.

[KK96] J.-Y. Koo and W.-C. Kim, *Wavelet density estimation by approximation of log-densities*, Statistics & probability letters **26** (1996), no. 3, 271–278.

[KK00] J.-Y. Koo and C. Kooperberg, *Logspline density estimation for binned data*, Statistics & probability letters **46** (2000), no. 2, 133–147.

[KKBG09] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, ℓ_1 *trend filtering*, SIAM review **51** (2009), no. 2, 339–360.

[KM06] R. Koenker and I. Mizera, *Density estimation by total variation regularization*, Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum (2006), 613–634.

[Kni99] K. Knight, *Epi-convergence in distribution and stochastic equi-semicontinuity*, Unpublished manuscript **37** (1999).

[Kni01] _____, *Limiting distributions of linear programming estimators*, Extremes **4** (2001), no. 2, 87–103.

[KR93] A. J. King and R. T. Rockafellar, *Asymptotic theory for solutions in statistical estimation and stochastic programming*, Mathematics of Operations Research **18** (1993), no. 1, 148–162.

[KS01] Y. M. Kabanov and C. Stricker, *The harrison–pliska arbitrage pricing theorem under transaction costs*, Journal of Mathematical Economics **35** (2001), no. 2, 185–196.

[KW91] A. J. King and R. J. Wets, *Epi-consistency of convex stochastic programs*, Stochastics and Stochastic Reports **34** (1991), no. 1-2, 83–92.

[LC12] L. Le Cam, *Asymptotic methods in statistical decision theory*, Springer Science & Business Media, 2012.

[LCY12] L. Le Cam and G. L. Yang, *Asymptotics in statistics: some basic concepts*, Springer Science & Business Media, 2012.

[LD07] G. LaFree and L. Dugan, *Introducing the global terrorism database*, Terrorism and Political Violence **19** (2007), no. 2, 181–204.

[LeC73] L. LeCam, *Convergence of estimates under dimensionality restrictions*, The Annals of Statistics (1973), 38–53.

[Lee12] P. M. Lee, *Bayesian statistics: an introduction*, John Wiley & Sons, 2012.

[LMSW16] H. Li, A. Munk, H. Sieling, and G. Walther, *The essential histogram*, arXiv preprint arXiv:1612.07216 (2016).

[LRU14] A. Löhne, B. Rudloff, and F. Ulus, *Primal and dual approximation algorithms for convex vector optimization problems*, Journal of Global Optimization **60** (2014), no. 4, 713–736.

[MC76] M. J. Magill and G. M. Constantinides, *Portfolio selection with transactions costs*, Journal of Economic Theory **13** (1976), no. 2, 245–263.

[MD17] A. Meir and M. Drton, *Tractable post-selection maximum likelihood inference for the lasso*, arXiv preprint arXiv:1705.09417 (2017).

[Mer71] R. C. Merton, *Optimum consumption and portfolio rules in a continuous-time model*, Journal of economic theory **3** (1971), no. 4, 373–413.

[MvdG+97] E. Mammen, S. van de Geer, et al., *Locally adaptive regression splines*, The Annals of Statistics **25** (1997), no. 1, 387–413.

[NN94] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*, vol. 13, Siam, 1994.

[Ope17] OpenStreetMap contributors, *Planet dump retrieved from https://planet.osm.org* , https://www.openstreetmap.org, 2017.

[Pfl91] G. C. Pflug, *Asymptotic dominance and confidence for solutions of stochastic programs*, International Institute for Applied Systems Analysis, 1991.

[Pfl95] ———, *Asymptotic stochastic programs*, Mathematics of Operations Research **20** (1995), no. 4, 769–789.

[PSST16] O. H. M. Padilla, J. G. Scott, J. Sharpnack, and R. J. Tibshirani, *The dfs fused lasso: Linear-time denoising over general graphs*, arXiv preprint arXiv:1608.03384 (2016).

[PWM07] J. K. Pal, M. Woodroofe, and M. Meyer, *Estimating a polya frequency function*, Lecture Notes-Monograph Series (2007), 239–249.

[Rob07] C. Robert, *The bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media, 2007.

[Roc74] R. T. Rockafellar, *Conjugate duality and optimization*, vol. 14, SIAM, 1974.

[Roc97] R. Rockafellar, *Convex analysis*, Princeton landmarks in mathematics and physics, Princeton University Press, 1997.

[Roc99] R. Rockafellar, *Duality and optimality in multistage stochastic programming*, Annals of Operations Research **85** (1999), 1–19.

[Roc15] R. T. Rockafellar, *Convex analysis*, Princeton university press, 2015.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: nonlinear phenomena **60** (1992), no. 1-4, 259–268.

[Ruf] K. Rufibach, *Log-concave density estimation and bump hunting for iid observations*, Ph.D. thesis.

[Ruf07] ———, *Computing maximum likelihood estimators of a log-concave density function*, Journal of Statistical Computation and Simulation **77** (2007), no. 7, 561–574.

[RW09] R. Rockafellar and R. Wets, *Variational analysis*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2009.

[RW13a] J. O. Royset and R. J. Wets, *Nonparametric density estimation via exponential epi-eplines: Fusion of soft and hard information*, Tech. report, 2013.

[RW13b] ———, *Nonparametric density estimation via exponential epi-eplines: Fusion of soft and hard information*, Tech. report, 2013.

[RWW10] R. Rockafellar, M. Wets, and R. Wets, *Variational analysis*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2010.

[SBG+17] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, *Osqp: An operator splitting solver for quadratic programs*, arXiv preprint arXiv:1711.08013 (2017).

[Sch04] W. Schachermayer, *The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time*, Math. Finance **14** (2004), no. 1, 19–48.

[Sco79] D. W. Scott, *On optimal and data-based histograms*, Biometrika **66** (1979), no. 3, 605–610.

[Sha91] A. Shapiro, *Asymptotic analysis of stochastic programs*, Annals of Operations Research **30** (1991), no. 1, 169–186.

[Sil82] B. W. Silverman, *On the estimation of a probability density function by the maximum penalized likelihood method*, The Annals of Statistics (1982), 795–810.

[SNF+13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*, IEEE Signal Processing Magazine **30** (2013), no. 3, 83–98.

[SS94] S. E. Shreve and H. M. Soner, *Optimal investment and consumption with transaction costs*, The Annals of Applied Probability (1994), 609–692.

[ST10] S. Sardy and P. Tseng, *Density estimation by total variation penalized likelihood driven by the sparsity 1 information criterion*, Scandinavian Journal of Statistics **37** (2010), no. 2, 321–337.

[Stu26] H. A. Sturges, *The choice of a class interval*, Journal of the american statistical association **21** (1926), no. 153, 65–66.

[SW86] G. Salinetti and R. J.-B. Wets, *On the convergence in distribution of measurable multifunctions (random sets) normal integrands, stochastic processes and stochastic infima*, Mathematics of Operations Research **11** (1986), no. 3, 385–419.

[Tal06] M. Talagrand, *The generic chaining: upper and lower bounds of stochastic processes*, Springer Science & Business Media, 2006.

[Tod08] E. Todorov, *General duality between optimal control and estimation*, Decision and Control, 2008. CDC 2008. 47th IEEE Conference on, IEEE, 2008, pp. 4286–4292.

[TSR+05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 1, 91–108.

[TT+12] R. J. Tibshirani, J. Taylor, et al., *Degrees of freedom in lasso problems*, The Annals of Statistics **40** (2012), no. 2, 1198–1232.

[VDVW96] A. W. Van Der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, Springer, 1996.

[Wah90] G. Wahba, *Spline models for observational data*, vol. 59, Siam, 1990.

[Wal09] G. Walther, *Inference and modeling with log-concave distributions*, Statistical Science (2009), 319–327.

[Wan11] S. Wang, *The utility maximization problem in markets with transaction costs: A set-valued approach*, 2011, Princeton University Senior Thesis, with contributions form Andreas Ham and Amit Singer.

[WN07] R. M. Willett and R. D. Nowak, *Multiscale poisson intensity and density estimation*, IEEE Transactions on Information Theory **53** (2007), no. 9, 3171–3187.

[Wri97] S. J. Wright, *Primal-dual interior-point methods*, Siam, 1997.

[WSST16] Y.-X. Wang, J. Sharpnack, A. Smola, and R. J. Tibshirani, *Trend filtering on graphs*, Journal of Machine Learning Research **17** (2016), no. 105, 1–41.

[YB99] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics (1999), 1564–1599.