THREE ESSAYS ON NONPARAMETRIC ESTIMATION

BY

RYAN CUMINGS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Roger Koenker, Chair
Professor Anil Bera
Assistant Professor Ji Hyung Lee
Assistant Professor Minchul Shi

Abstract

The first essay describes a shape constrained density estimator, which, in terms of the assumptions on the functional form of the population density, can be viewed as a middle ground between fully nonparametric and fully parametric estimators. For example, typical constraints require the estimator to be log $-$concave or, more generally, $\rho-$concave; (Koenker and Mizera, 2010). In cases in which the true population density satisfies the shape constraint, these density estimators often compare favorably to their fully nonparametric counterparts; see for example, (Cule et al., 2010; Koenker and Mizera, 2010). The particular shape constrained density estimator proposed here is defined as the minimum of the entropy regularized Wasserstein metric provided by Cuturi (2013), which can be found with a nearly linear time complexity in the number of points in the mesh (Altschuler et al., 2017). It is also a common thread that links all three essays.

After providing results on consistency, limiting distribution, and rate of convergence for the estimator, the paper moves onto testing if a population density satisfies a shape constraint. This is done by deriving the limiting distribution of the regularized Wasserstein metric at the estimator. In the interest of tractability these results are initially described in terms of $\rho-$concave shape constraints. The final result provides the additional requirements that arbitrary shape constraints must satisfy for these results to hold. The generalization is then used to explore whether the California Department of Transportation's decision to award construction contracts with the use of a first price auction is cost minimizing. The shape constraint in this case is given by Myerson's (1981) regularity condition, which is a requirement for the first price auction to be cost minimizing.

The next essay provides a novel nonparametric estimator of the mode of a random variable conditional on covariates, which is also known as a modal regression. The estimator is defined by first finding several paths through the data, and then aggregating these paths together with the use of a regularized Wasserstein barycenter. The initial paths each minimize a combination of distance between subsequent points as well as the curvature along the path. The paper provides a result on consistency, and shows that the estimator has a time complexity that is $O(n^2)$, where n is the number of points in the sample. An approximation is also provided that has a time complexity of $O(n^{1+2\beta})$, where $\beta \in (0, 1/2)$. Simulations are also provided, and then the estimator is used in an application to find the mode of undergraduate GPA, conditional on high school GPA and college entrance tests on cumulative undergraduate GPA.

Note that the estimators provided in these first two essays require minimizing the regularized Wasserstein metric. In their textbook treatment, Peyré and Cuturi (2019) state that second order methods are not an "applicable" approach for optimization in

this setting. This is because of the large scale of many applications of interest, as well as because the Hessian is dense, poorly scaled, and not expressible in closed form. In the third paper we provide functions to evaluate products with this Hessian and its inverse that have a nearly linear time complexity. We also provide recommendations to allow these methods to be applied in standard second order optimization implementations, and discuss how in certain favorable cases these approaches have a provably nearly linear time complexity. Afterward, numerical experiments are carried out outside of these favorable cases. When the derivatives of the constraints are sufficiently sparse, the computational efficiency of the approach in this setting is also favorable.

Acknowledgements

I owe a debt of gratitude to the many individuals who have been there for me throughout the years as I have pursued my research and my PhD at the University of Illinois. In particular, I would like to thank my Committee Chair, Roger Koenker for his patience and insightful comments through the many iterations of these chapters. His curiosity and pursuit of novel ideas have had a large impact on my own interests in econometrics. I would also like to thank the other members of my committee, Minchul Shin, Anil Bera, and Ji Hyung Lee, for our many engaging and insightful conversations. These conversations shaped the way I wrote and viewed the concepts in these essays. I would also like to thank those that have motivated and inspired me to continue my academic pursuits, including Brent Hickman, Charles Frohman, and Ron Harstad.

I am also grateful to my friends and family that have come along with me in this journey. I would like to thank my brother and "sis," Drew and Cassie, who have been continual sources of encouragement, and my parents, Michael and Roxanne, for their unconditional love and the value they have placed on education. I would also like to thank Shashikala and Vinod, for warmly welcoming me into their family. Finally, I would like to thank my loving wife, Suvarna, for always being there, literally, as she wrote alongside me in all the cafes in Urbana, and, more importantly, figuratively. Thank you for your support throughout this process and for your patience, and even encouragement, while discussing contrived analogies to bowls, sand, and walks across campus.

Table of Contents

Chapter 1

Shape-Constrained Density Estimation Via Optimal Transport

Abstract

Constraining the maximum likelihood estimator to satisfy a sufficiently strong constraint, log −concavity being a common example, has the effect of restoring consistency without requiring additional parameters. Since many results in economics require densities to satisfy a shape constraint, these estimators are also attractive for the structural estimation of economic models. In all the examples provided by Bagnoli and Bergstrom (2005) and Ewerhart (2013), log −concavity is sufficient to ensure that the density satisfies the required conditions. However, in many cases log −concavity is far from necessary, and it has the unfortunate side effect of ruling out sub-exponential tail behavior.

In this paper, we use optimal transport to formulate a shape constrained density estimator. We initially describe the estimator using a $\rho-$concavity constraint. In this setting we provide results on consistency, asymptotic distribution, convexity of the optimization problem defining the estimator, and formulate a test for the null hypothesis that the population density satisfies a shape constraint. Afterward, we provide sufficient conditions for these results to hold using an arbitrary shape constraint. This generalization is used to explore whether the California Department of Transportation's decision to award construction contracts with the use of a first price auction is cost minimizing. We estimate the marginal costs of construction firms subject to Myerson's (1981) regularity condition, which is a requirement for the first price reverse auction to be cost minimizing. The proposed test fails to reject that the regularity condition is satisfied.

1.1 Introduction

Nonparametric density estimation has the advantage over its parametric counterparts of not requiring the underlying population density to belong to a specific family. In the case of distribution functions, Kiefer and Wolfowitz (1956) showed that the empirical distribution function is a maximum likelihood estimator; however, attempting to use this distribution function to directly define a nonparametric density estimate results in a series of point masses located at each of the datapoints. Grenander (1956) provided the first example of a shape constrained density estimator as a way to extricate the maximum likelihood estimator from this "Dirac catastrophe." Specifically, he showed that maximizing the likelihood function subject to a monotonicity constraint on the estimator results in density estimates without point masses.

A great deal of progress was made in subsequent decades by adding penalty terms to the maximum likelihood objective function to restore the parsimony of the density estimator; for example, see (Silverman, 1986). Parzen (1962) also showed that kernel density estimators resulted in consistent density estimators and derived the rates of convergence. Unlike Grenander's (1956) approach, the performance of maximum penalized likelihood estimators and kernel density estimators is highly dependent on the specification of penalty terms and bandwidths, respectively, which can be difficult to choose.

Partly for this reason, recently there has been a renewed interest in ensuring parsimony of the maximum likelihood density estimator through conditioning on the information provided by the shape of the underlying density. In particular, significant progress has been made on the maximum likelihood density estimator subject to the constraint that the logarithm of the density is a concave function, which defines a log $-$concave density (Dümbgen and Rufibach, 2009; Cule, Samworth, and Stewart, 2010; Kim and Samworth, 2016).

One early pioneer on the advantages of log $-$concavity for both statistical testing as well as estimation was Karlin (1968). Suppose the distribution function $F : \mathbb{R} \to [0, 1]$ has a density function denoted by $f : \mathbb{R} \to \mathbb{R}_+$. Some examples of log $-$concavity's many implications include that the density $f(x - \theta)$ has a monotonic likelihood ratio if and only if $f(\cdot)$ is log $-$concave, products and convolutions between log $-$concave densities are log $-$concave, and that the hazard function of the log $-$concave density $f(x)$, defined by $f(x)/(1 - F(x))$, is increasing. Bagnoli and Bergstrom (2005) also provide a survey of economic models in which log $-$concavity of a density is a sufficient condition for the existence or uniqueness of an equilibrium. Chen and Samworth (2013) as well as Dümbgen, Samworth, and Schuhmacher (2011) provide tests for a population density satisfying log $-$concavity, and Carroll, Delaigle, and Hall (2011) provide a test for a population density satisfying a more general set of shape

constraints.

A wide variety of the random variables in the economics literature, such as annual income or changes in stock prices, are thought to exhibit sub-exponential tail behavior, so a $\log-$concavity constraint would not result in a consistent estimator in these cases. Koenker and Mizera (2010) generalized the $\log-$concave maximum likelihood estimator by maximizing Rényi entropy of order $\rho \in \mathbb{R}$ subject to the $\rho-$concavity constraint,

$$f\left(\alpha x_0 + (1-\alpha)x_1\right) \geq \left(\alpha f(x_0)^\rho + (1-\alpha)f(x_1)^\rho\right)^{1/\rho},$$

for all $\alpha \in [0,1]$. This estimator converges to the maximum likelihood estimator subject to a $\log-$concavity constraint in the limit as $\rho \to 0$.[1] Decreasing ρ corresponds to a relaxation of this shape constraint, so if $f(x)$ satisfies the constraint for some ρ, then it also satisfies the constraint for all $\rho' < \rho$. Also, this constraint is equivalent to concavity when ρ is equal to one, and the cases of $\log-$concavity and quasi-concavity can be derived in the limit as $\rho \to 0$ and $\rho \to -\infty$ respectively. Koenker and Mizera (2010) place particular emphasis on the case in which $\rho = -1/2$, partly because most standard densities are $-1/2-$concave. For example, all Student t_v densities with $v \geq 1$ satisfy this constraint.

$\rho-$concavity constraints provide a considerable relaxation over $\log-$concavity constraints, while restricting the set of feasible densities sufficiently to ensure parsimony of the density estimator. These constraints are also sufficient conditions for many results in economics, including the uniqueness or existence of equilibria in a variety of models; see for examples, (Ewerhart, 2013; Bagnoli and Bergstrom, 2005). However, in many cases the necessary and sufficient conditions for these results are considerably weaker, so inference and estimation based on these stronger conditions can provide misleading results. For example, inferring whether a population density satisfies these weaker conditions based on tests for their more restrictive counterparts is generally not possible. Things are less straightforward in the case of estimation because shape constraints are generally the source of the density estimator's parsimony. However, since using a shape constraint that is not satisfied by the population density would not result in a consistent estimator, it is prudent to err toward the weakest constraint that theory predicts a population density would satisfy, or when using the estimate in a structural model, the weakest constraint that a model requires a population density to satisfy.

For a concrete example in the economics literature, given a density of private valuations of risk neutral agents, Myerson (1981) defined the virtual valuations function as $x - (1 - F(x))/f(x)$, and showed that the first price auction is revenue maximizing if this function is

[1] Maximum likelihood is equivalent to maximizing Shannon entropy, and Rényi entropy of order ρ converges to Shannon entropy as $\rho \to 0$.

strictly increasing. Sufficient conditions for Myerson's (1981) regularity condition, ordered from strongest to weakest, are log $-$concavity, a monotonic hazard rate, and $\rho-$concavity for $\rho > -1/2$ (Ewerhart, 2013). While the first two sufficient conditions are commonly cited in mechanism design, they both imply exponential tail behavior. Since it seems plausible that willingness to pay is influenced by ability to pay, allowing valuations to have sub-exponential tail behavior may be a reasonable modeling choice, given that wealth and income are typically modeled with sub-exponential tails. Luckily, Myerson's (1981) regularity condition does not exclude these densities. For example, while the log $-$normal density has sub-exponential tails, it also satisfies this condition when $\sigma^2 < 2$, which holds in the structural model provided by Laffont, Ossard, and Vuong (1995).

This paper provides a framework for estimating and performing inference with shape constrained densities using regularized optimal transport (Cuturi, 2013). This objective function has the advantage of having an unconstrained global optimum that is a consistent density estimator, which ameliorates the requirement that the shape constraint is the only source of parsimony. At first we motivate the method using a $\rho-$concavity constraint, but one of the advantages of the method is that the estimator is consistent when this constraint is replaced by a wide variety of alternative shape constraints. We also provide a consistent test for whether or not a population density satisfies a shape constraint based on comparing the objective function at the unconstrained optimum to the constrained optimum.

After introducing density estimation with this more general class of shape constraints, we use the proposed estimator to explore whether or not the California Department of Transportation's decision to use a reverse first price auction to award construction contracts is cost minimizing. To do this, we use the method provided by Guerre, Perrigne, and Vuong (2000) to calculate the firms' marginal costs using data on their bids. We find that a kernel density estimator of these costs does not satisfy Myerson's (1981) regularity condition everywhere; however, the proposed density estimate, subject to the constraint that Myerson's (1981) regularity condition is satisfied, appears to follow the data closely. Our test also fails to reject that the population density satisfies Myerson's (1981) regularity condition.

In addition to the flexibility offered by the proposed framework, there are three other advantages of the proposed method. First, the notion of fidelity to the data that we optimize is independent of the constraint, including the choice of ρ in the case of $\rho-$concavity constraints. Note that the objective function used in Koenker and Mizera's (2010) approach, Rényi entropy of order ρ, is dependent on this constraint parameter. Also, adding a $\rho-$concavity constraint to the maximum likelihood estimator would not provide a convincing way to achieve this goal since this would not provide a convex optimization problem for values of $\rho < 0$ and the estimator does not exist when $\rho < -1$ (Doss and Wellner, 2016).

Second, the shape constraints of the estimator only binds in regions in which it would not otherwise be satisfied. This is advantageous when using shape constraints that are not sufficiently restrictive to ensure parsimony by themselves. Moreover, the existence of the unconstrained minimizer ensures that the density estimator exists, regardless of the strength of the shape constraint. As discussed in the preceding paragraph, this is not the case for the maximum likelihood estimator. Although it is not our primary focus, we also provide an option for relying on the shape constraint for parsimony.

Third, the proposed algorithm solves an optimization problem over a set of variables that grows sub-linearly in the sample size, so the time complexity of the proposed algorithm compares favorably to other shape constrained density estimators.

The next section outlines the aspects of optimal transport that are required to formulate our estimator. Galichon (2016) provides a more comprehensive overview of the optimal transport literature, including its many applications in economics. The third section defines the estimator, provides the rate of convergence, and the asymptotic distribution of the estimator. This section also provides a test for the null hypothesis that the population density satisfies the shape constraint. The fourth section proves that the optimization problem defining the estimator is convex and briefly outlines a simple approach for finding the estimator. A more computationally efficient approach is provided in the third chapter, as it is outside of the scope of this chapter. The sixth section generalizes the framework presented here to allow for density estimation and inference subject to a much larger class of shape constraints, with a focus on shape constraints that arise in the economics literature. Specifically, this section provides sufficient conditions for each of the results in the paper to hold under an arbitrary set of shape constraints. This generalization is used in the seventh section to provide evidence that the firms bidding on the California Department of Transportation's construction contracts have marginal cost distributions that satisfy Myerson's (1981) regularity condition.

A few notational conventions will be useful in the subsequent sections. For $x \in \mathbb{R}^m$, we will denote the vector with an i^{th} element defined by $\exp(x_i)$ as $\exp(x)$, and a similar convention will be used for $\log(x)$ and x^ρ. Also, a diagonal matrix with a diagonal equal to the vector x will be denoted by D_x, an $m \times 1$ vector of ones by $\mathbf{1}_m$, the identity matrix by I, element-wise division of the two vectors x and y by $x \oslash y$, element-wise multiplication by $x \otimes y$, the Moore-Penrose pseudoinverse of the matrix A as A^+, the convolution between $f : \mathbb{R}^d \to \mathbb{R}^1$ and $g : \mathbb{R}^d \to \mathbb{R}^1$ by $g(x) * f(x)$, the derivative of $f(x)$ with respect to x by $\nabla_x f(x)$, and a Dirac delta function centered at z by $\delta_z(x)$. Also, $sgn(x)$ will be used to denote a function that is 1 when $x \geq 0$ and -1 when $x < 0$. Since the proposed method requires discretizing densities, say $\mu : \mathcal{A} \to \mathbb{R}^1$ for $\mathcal{A} \subset \mathbb{R}^d$, we will denote the points in the mesh as $\{\mathbf{a}_i\}_{i=1}^m$, where $\mathbf{a}_i \in \mathbb{R}^d$. We will also continue to include parenthesis after functions,

as in $\mu(x)$ or $\mu(\cdot)$, and exclude parenthesis when denoting $\mu \in \mathbb{R}^m$ with elements $\mu_i = \mu(\mathbf{a}_i)$.

1.2 Optimal Transport

Gaspard Monge formulated the theory of optimal transport in the 18^{th} century in order to derive the optimal method of moving a pile of sand to a nearby hole of the same volume. Specifically, suppose that both the pile of sand and the hole are defined on $\mathcal{A} \subset \mathbb{R}^d$, and we use the measures $\mathcal{M}_0 : \mathcal{A} \to \mathbb{R}_+$ and $\mathcal{M}_1 : \mathcal{A} \to \mathbb{R}_+$ to define the volume of the pile and the hole respectively. Monge sought to find a transportation plan, $T : \mathcal{A} \to \mathcal{A}$, that minimizes transportation costs while ensuring that the hole is completely filled.

Kantorovitch (1958) generalized this problem by describing the transportation plan by the absolutely continuous measure $\psi : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$. For example, given $a_1 \in \mathcal{A}$ and $a_2 \in \mathcal{A}$, we can view the Radon-Nikodym derivative of $\psi(\cdot)$, $d\psi(a_1, a_2)$, as the amount of mass moved from a_1 to a_2 under the transportation plan, or *coupling*, $\psi(\cdot)$. Feasibility of $\psi(\cdot)$ simply requires $\psi(a, \mathcal{A}) = \mathcal{M}_1(a)$ and $\psi(\mathcal{A}, a) = \mathcal{M}_0(a)$ for all $a \in \mathcal{A}$.

When $\mathcal{M}_i(\cdot)$ is absolutely continuous, there exists $\mu_i(a)$ such that $\mathcal{M}_i(A) = \int_A \mu_i(a)da$ for each $A \subset \mathcal{A}$ by the Radon-Nikodym theorem. When $\mathcal{M}_i(\cdot)$ also satisfies $\mathcal{M}_i(\mathcal{A}) = 1$, then $\mu_i(a)$ is also a probability density function. Optimal transport can be described without assuming $\mathcal{M}_0(\cdot)$ and $\mathcal{M}_1(\cdot)$ satisfy these conditions; however, since our goal is density estimation, we will generally restrict our attention to these cases for the rest of the paper. In addition we will define (or constrain) all density functions to be continuous, with the exception of $\delta_z(\cdot)$. We will use this notation to define $\Psi(\mu_0(\cdot), \mu_1(\cdot))$ as the set of feasible couplings.

The most common cost function in optimal transport is simply squared Euclidean distance. In this case the cost of moving one unit of earth from $a_1 \in \mathcal{A}$ to $a_2 \in \mathcal{A}$ is proportional to $\|a_1 - a_2\|^2$. The resulting minimization problem is then given by

$$W_0(\mu_0(\cdot), \mu_1(\cdot)) := \min_{\psi \in \Psi(\mu_0(\cdot), \mu_1(\cdot))} \int_{\mathcal{A}\times\mathcal{A}} \|a_1 - a_2\|^2 \, d\psi(a_1, a_2), \tag{1.1}$$

which we will refer to as the squared Wasserstein distance (Mallows, 1972). $W_0(\mu_0(\cdot), \mu_1(\cdot))$ has many desirable properties, one being that $\sqrt{W_0(\mu_0(\cdot), \mu_1(\cdot))}$ satisfies all of the usual properties of a distance metric. $W_0(\mu_0(\cdot), \mu_1(\cdot))$ also metrizes weak convergence and convergence in the first two moments. In other words, given a sequence of densities, $\{\mu_0(\cdot), \mu_1(\cdot), ..., \mu_n(\cdot)\}$, we have $\lim_{i\to\infty} W_0(\mu_0(\cdot), \mu_i(\cdot)) = 0$ if and only if $\mathcal{M}_i$ converges weakly to $\mathcal{M}_0$ and the first two moments of μ_i converge to the first two moments of μ_0.

The Wasserstein distance between two distributions can be viewed as a measure of distance over the domain of the densities rather than in the direction of their range. For example, the Fréchet mean of two Dirac delta functions, centered at a and b, in the spaces of densities equipped with an L^2 norm is given by $\arg\min_{\nu(x)} \|\delta_a(x) - \nu(x)\|^2 + \|\delta_b(x) - \nu(x)\|^2 = \delta_a(x)/2 + \delta_b(x)/2$, while a similar notion of average in the spaces of densities equipped with the Wasserstein distance is $\arg\min_{\nu(x)} W_0(\delta_a(x), \nu(x)) + W_0(\delta_b(x), \nu(x)) = \delta_{a/2+b/2}(x)$. To make this intuition more explicit, when $\mathcal{A} \subset \mathbb{R}^1$, one can show $W_0(\mu_0(\cdot), \mu_1(\cdot))$ can also be expressed as $\int_0^1 (Q_0(\tau) - Q_1(\tau))^2 d\tau$, where $Q_0(\tau)$ and $Q_1(\tau)$ are the quantile functions corresponding to $\mu_0(\cdot)$ and $\mu_1(\cdot)$ respectively. Thus, $(Q_0(\tau) - Q_1(\tau))^2$ represents a squared distance between two points in $\mathcal{A}$ (Villani, 2003).

In practice augmenting the Wasserstein distance with a regularization term ameliorates some numerical difficulties, which will be described below in more detail. The regularized squared Wasserstein distance is a generalization of $W_0(\mu_0(\cdot), \mu_1(\cdot))$, and is defined by

$$W_\gamma(\mu_0(\cdot), \mu_1(\cdot)) := \min_{\psi \in \Psi(\mu_0(\cdot), \mu_i(\cdot))} \int_{\mathcal{A}\times\mathcal{A}} \|a_1 - a_2\|^2 \, d\psi(a_1, a_2) - \gamma H(\psi(\cdot)), \tag{1.2}$$

where $\gamma \geq 0$ and $H(\psi(\cdot)) := -\int_{\mathcal{A}\times\mathcal{A}} \log \psi(a_1, a_2) d\psi(a_1, a_2)$ is the Shannon entropy of $\psi(\cdot)$ (Cuturi, 2013; Cuturi and Doucet, 2014). In the shape constrained density estimation setting, the addition of this entropy term is advantageous for several reasons. First, the objective function is strictly convex when $\gamma > 0$, so the optimal coupling will always be unique. Second, in practice $\mathcal{A}$ must be discretized before finding the unregularized Wasserstein distance, and the computational cost of solving for the optimal coupling scales at least cubically in the number of points in the mesh. Third, after discretizing, the minimizer of $W_\gamma(\mu_0, \mu_1)$ with respect to μ_0 is often a more accurate representation of the minimizer of $W_0(\mu_0(\cdot), \mu_1(\cdot))$, when γ is set to a reasonably small value.[2] Four, using $W_\gamma(\mu_0(\cdot), \mu_1(\cdot))$ allows us to avoid assumptions in the next section regarding the existence of the second moments of μ_0 and μ_1. Lastly, we can find the minimizer of (2) with a very computationally efficient algorithm after discretizing, which we will describe next.

To introduce the discretized counterparts of $d\psi(\cdot), \mu_1(\cdot)$, and $\mu_0(\cdot)$, recall our uniform mesh over $\mathcal{A}$ contains the vertices $\{\mathbf{a}_i\}_{i=1}^m$, and let μ_0, μ_1 define $\mu_0(\mathbf{a}_i), \mu_1(\mathbf{a}_i)$ respectively. Also, let $M_{m\times m}$ so that $M_{ij} := \|\mathbf{a}_i - \mathbf{a}_j\|^2$, and $\psi_{m\times m}$ so that $\psi_{i,j} := d\psi(\mathbf{a}_i, \mathbf{a}_j)$. After discretizing, (2) can be written as

$$W_\gamma(\mu_0, \mu_1) := \min_{\psi} \sum_{i,j} \psi_{ij} M_{ij} + \gamma \psi_{ij} \log(\psi_{ij}) \text{ subject to:} \tag{1.3}$$

[2]Minimizing $W_0(\mu_0, \mu_1)$ with respect to μ_0 generally results in a minimizing density with many large discrete changes. For more detail, see Figures 3.1, 3.2, and the accompanying explanation in (Cuturi and Peyré, 2016).

$$\sum_{j}^{m} \psi_{ij} = \mu_{0i} \; \forall i \in \{1, 2, ..., m\} \tag{1.4}$$

$$\sum_{i}^{m} \psi_{ij} = \mu_{1j} \; \forall i \in \{1, 2, ..., m\}. \tag{1.5}$$

The corresponding Lagrangian is given by

$$\mathcal{L} = \left(\sum_{i,j} \gamma \psi_{ij} \log(\psi_{ij}) + \psi_{ij} M_{ij} \right) - \lambda_0^T \left(\sum_{j}^{m} \psi_{\cdot j} - \mu_0 \right) - \lambda_1^T \left(\sum_{i}^{m} \psi_{i\cdot} - \mu_1 \right), \tag{1.6}$$

and the first order conditions imply

$$\psi_{ij} = \exp(\lambda_{0i}/\gamma - 1/2) \exp(-M_{ij}/\gamma) \exp(\lambda_{1j}/\gamma - 1/2).$$

In other words, there exists $v, w \in \mathbb{R}_+^m$ such that the optimal coupling has elements $\psi_{ij} = K_{ij} w_i v_j$, where $K_{ij} := \exp\left(-\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2 / \gamma\right)$, a symmetric $m \times m$ matrix. This can also be written as,

$$\psi = D_w K D_v, \tag{1.7}$$

so adding the entropy term to the objective function reduces the dimensionality of the optimization problem from m^2 to $2m$. Sinkhorn (1967) shows that ψ is unique. The iterative proportional fitting procedure (IPFP) is an efficient method of computing these vectors; see Krupp (1979). This method iteratively redefines w so that $D_w K v = \mu_0$, and subsequently redefines v so that $D_v K w = \mu_1$, as summarized in Algorithm 1. Note that after combining these equalities we have $D_w K(\mu_1 \oslash (Kw)) = \mu_0$, which will be used in subsequent sections.

Algorithm 1 The iterative proportional fitting procedure.

function `IPFP`(K, μ_0, μ_1)
$w \leftarrow \mathbf{1}_m$
until convergence:
  $v \leftarrow \mu_1 \oslash (Kw)$
  $w \leftarrow \mu_0 \oslash (Kv)$
return w, v

In the rest of the paper we will make substantial use of the dual of (3)-(5). Cuturi and Doucet (2014) show that the dual is given by the unconstrained optimization problem,

$$W_\gamma(\mu_0, \mu_1) = \max_{(x,y) \in \mathbb{R}^{2m}} x^T \mu_0 + y^T \mu_1 - \gamma \sum_{i,j} \exp((x_i + y_j - M_{ij})/\gamma). \tag{1.8}$$

Note that $K = \exp(-M/\gamma)$, so the first order conditions of (8) can be written as,

$$\mu_{0i} / \left(\sum_j \exp(y_j/\gamma) K_{ij} \right) = \exp(x_i/\gamma) \tag{1.9}$$

$$\mu_{1j} / \left(\sum_i \exp(x_i/\gamma) K_{ij} \right) = \exp(y_j/\gamma). \tag{1.10}$$

Note that after replacing $\exp(x/\gamma)$ with w and $\exp(y/\gamma)$ with v, these formulas are equivalent to the updates of w and v given in Algorithm 1. Also, given x and y that satisfy (9)-(10), consider the vectors $\tilde{x} := x - c$ and $\tilde{y} := y + c$, where $c \in \mathbb{R}$. Since μ_0 and μ_1 have the same sum, the objective function of (8), evaluated at $\tilde{x}$ and $\tilde{y}$ must equal the objective function evaluated at x and y. Since $\exp(\tilde{y}_j/\gamma) = \exp(y_j/\gamma)\exp(c/\gamma)$ and $\exp(\tilde{x}_i/\gamma) = \exp(x_i/\gamma)\exp(-c/\gamma)$, $\tilde{x}$ and $\tilde{y}$ must also satisfy (9)-(10). In other words, while v and w are unique up to a multiplicative constant on w and one over this constant on v, y and x are unique up to the additive constant c.

A few comments regarding the effect of γ on the optimal coupling will also be useful in subsequent sections. Higher values of γ correspond to placing a higher penalty on the negative entropy of the coupling, so the optimal coupling becomes more dispersed as this parameter is increased. Also, in the limit $\gamma \to 0$, $W_\gamma(\mu_0, \mu_1)$ converges to $W_0(\mu_0, \mu_1)$ at the rate $O(\exp(-1/\gamma))$ and ψ converges to the optimal unregularized coupling at this same rate; see Benamou et al. (2015) and Cuturi (2013). In the next section will use $W_\gamma(\cdot)$ as an objective function to define the proposed estimator and show how γ impacts this estimator in more detail.

1.3 Shape-Constrained Density Estimation

The input of the density estimator proposed in this paper is a kernel density estimator, μ, based on N i.i.d. datapoints, $\{z_i\}_{i=1}^N$, drawn from a uniformly continuous population density, $\mu^\star : \mathbb{R}^d \to \mathbb{R}^1_+$, with a bandwidth of $\sigma \geq 0$. Our results also hold when γ and σ are redefined to be functions of the form $c_1(\{z_i\}_i)\gamma$ and $c_2(\{z_i\}_i)\sigma$, where $c_j(\{z_i\}_i) \xrightarrow{p} c_j$ for $j \in \{1, 2\}$ at any polynomial rate. In the interest of the brevity of notation, we will only write the parameters in this way when their randomness would have a non-trivial impact on the result. γ, σ and m will also be dependent on N, but we will suppress this input throughout the paper and discuss our recommendations for defining them after Theorem 2.

With this notation in mind we can define the shape-constrained density estimator as,

$$\min_f W_\gamma(f, \mu) \text{ subject to:}$$

$$sgn(\rho) f^{\rho} \in \mathcal{K}, \tag{1.11}$$

where $\mathcal{K}$ is the cone of concave functions. Although $\mathcal{K}$ is a convex set, the set of $\rho-$concave densities is generally not convex. To ameliorate this problem, we will use a similar formulation as Koenker and Mizera (2010) and solve for $g := f^{\rho}$. Note that no generality is lost in doing so, as there is a one to one correspondence between g and f.

For the clarity of the derivations in the next section, we will also define g to be a vector of length $m-1$ and refer to the element of the vector f, of length m, that corresponds to this omitted value as f_k. We will set f_k so that the density sums to m. In other words, the elements of the density f that do not correspond to this k^{th} element, denoted by f_{-k}, will be set equal to $g^{1/\rho}$ and f_k will be set equal to $m - \sum_i g_i^{1/\rho}$.

Unlike optimizing over g rather than f, this can be seen as a slight modification of our initial optimization problem, since we will also exclude the shape constraints that depend on the k^{th} element of f. However, as discussed in the next section in more detail, one can choose k to correspond to an element on the boundary of the mesh so that the estimator satisfies the shape constraint everywhere on the interior of its domain.

In a slight abuse of notation, we will also denote the objective function as $W_{\gamma}(g^{1/\rho}, \mu)$. In summary, we will define $W_{\gamma}(g^{1/\rho}, \mu)$ by,

$$\max_{(x,y)\in\mathbb{R}^{2m}} x_{-k}^T g^{1/\rho} + x_k \left(m - \sum_i g_i^{1/\rho} \right) + y^T \mu - \gamma \sum_{i,j} \exp\left((x_i + y_j - M_{ij})/\gamma \right), \tag{1.12}$$

and the final form of our optimization problem is,

$$\min_g W_{\gamma}(g^{1/\rho}, \mu) \text{ subject to:} \tag{1.13}$$

$$g_i = \alpha_i + \beta_i \mathrm{a}_i, \;\; sgn(\rho)\left(\alpha_i - \alpha_j + (\beta_i - \beta_j)^T \mathrm{a}_i\right) \leq 0 \;\; \forall\, i,j \in \{2, ..., m-1\}, \tag{1.14}$$

where $\alpha_i \in \mathbb{R}^1$, $\beta_i \in \mathbb{R}^d$, and d is the dimension of the support of μ and f. These inequality constraints are used by Afriat (1972) to estimate production functions with concavity constraints. They tend to provide a gain in numerical accuracy relative to local concavity constraints.

The following Lemma provides the limiting distribution of the estimator after removing the shape constraints. This is achieved by showing that the resulting density can be viewed as a kernel density estimator with a bandwidth of $\sqrt{\sigma^2 + \gamma/2}$ away from the edges of the mesh

over $\mathcal{A}$. To avoid these boundary value effects, we recommend enlarging the domain of μ and $\hat{f}$ to include regions within approximately $3\sqrt{\sigma^2 + \gamma/2}$ of the datapoints. Alternatively, one could replace the matrix K with the direct application of a Gaussian filter. This approach has the added benefit of reducing the computational complexity of Algorithm 1 to $O(m \log m)$, so this is our recommended approach when $d > 2$; see Solomon et al. (2015) for more details on this method.

Lemma 1: *Suppose μ is a kernel density estimate, generated with a Gaussian kernel and a bandwidth σ and $\mu^\star(\cdot)$ is uniformly continuous. Also, suppose σ, γ and m are chosen so that $\sqrt{\sigma^2 + \gamma/2} \xrightarrow{p} 0$, $N\sqrt{\sigma^2 + \gamma/2} \xrightarrow{p} \infty$, and $\min_{i \neq j} \|\boldsymbol{a}_i - \boldsymbol{a}_i\| / \sqrt{\gamma} \to 0$ as $N \to \infty$. Then there exists $c \in \mathbb{R}^1$ so that the limiting density of $f_{unc} := arg\ min_f\ W_\gamma(f, \mu)$ is given by,*[3]

$$\sqrt{N(\sigma^2 + \gamma/2)^{d/2}} \left(f_{unc,i} - \mu_i^\star + c(\sigma^2 + \gamma/2)^d \right) \xrightarrow{d} N\left(0, \mu_i^\star / \left(2\sqrt{\pi}\right)^d\right)$$

Proof: To find f_{unc}, consider the optimization problem,

$$\min_{\psi} \sum_{i,j} \psi_{ij} M_{ij} + \gamma \psi_{ij} \log(\psi_{ij}) \text{ subject to:} \tag{1.15}$$

$$\sum_{i}^{m} \psi_{ij} = \mu_j \ \forall j \in \{1, 2, ..., n\} \tag{1.16}$$

The corresponding Lagrangian is

$$\mathcal{L} = \left(\textstyle\sum_{i,j} \gamma \psi_{ij} \log(\psi_{ij}) + \psi_{ij} M_{ij} \right) + \lambda_0^T \left(\textstyle\sum_i \psi_{i\cdot} - \mu \right),$$

and the first order conditions imply that there exists $v \in \mathbb{R}^m$ such that

$$\psi = K D_v. \tag{1.17}$$

Note that convexity of negative entropy implies that the optimal coupling will correspond to a minimizer. After combining this equality with the constraints, we have

$$\sum_{i}^{m} \psi_{ij} = v_j \sum_{i}^{m} K_{ij} = \mu_j, \tag{1.18}$$

which implies

$$v_j = \mu_j / \left(\textstyle\sum_i^m K_{ij} \right).$$

[3]When $\mu^\star(\cdot)$ is differentiable at $\boldsymbol{a}_i$, $c = \triangle \mu_i^\star(\boldsymbol{a}_i)/2$, where $\triangle \mu_i^\star(\boldsymbol{a}_i)$ denotes the Laplacian, $\sum_{j=1}^d \nabla_{x_j, x_j} \mu^\star(x)$ at $\boldsymbol{a}_i$. However, c can be defined without assuming $\mu^\star(\cdot)$ is differentiable; Karunamuni and Mehra (1991) provide more details on this approach.

Let $\kappa := K\mathbf{1}$. Now we can find f_{unc} by finding the other marginal of ψ,

$$f_{unc} := K\left(\mu \oslash \kappa\right).$$

Let $\phi_\eta : \mathbb{R}^d \to \mathbb{R}^1$ be a Gaussian density with variance ηI_d and mean $\mathbf{0}_d$. Suppose $\nu : \mathbb{R}^d \to \mathbb{R}^1$ is a continuous function. Then,

$$K\nu(\mathbf{a}) = \textstyle\sum_{i=1}^m \exp(-\left\|\mathbf{a}_i - \mathbf{a}_j\right\|/\gamma)\nu(\mathbf{a}_i)$$

$$\approx \tfrac{m\sqrt{\pi}}{\gamma} \textstyle\sum_{i=1}^m \phi_{\gamma/2}\left(\mathbf{a}_i - \mathbf{a}_j\right)\nu(\mathbf{a}_i)$$

by the definition of K, so K can be viewed as a linear operator that discretizes the convolution $\nu(\cdot) * \phi(\cdot)\sqrt{\pi}/\gamma$. Since the distance between adjacent points decreases at a rate that is faster than $\sqrt{\gamma/2}$, and the convex hull of the data is a strict subset of the convex hull of $\{\mathbf{a}_i\}_i$, the error of this approximation converges to zero on the convex hull of the data. Thus, as N diverges we have $\kappa_i \to \sqrt{\pi}/(m\gamma)$ for all i in the convex hull of the data.

Since μ is a kernel density estimator, it can be expressed as $\mu = \left(\sum_i \delta_{z_i}(\cdot) * \phi_{\sigma^2}/N\right), (\mathbf{a})$ which implies that f_{unc} converges to $\left(\sum_i \delta_{z_i}(\cdot) * \phi_{\sigma^2} * \phi_{\gamma/2}/N\right)(\mathbf{a})$. The convolution of two Gaussian densities is $\phi_{\sigma^2} * \phi_{\gamma/2} = \phi_{\sigma^2+\gamma/2}$, so $f_{unc} = \left(\sum_i \delta(z_i)\right) * \phi_{\sigma^2+\gamma/2}(y)/N$, which defines a kernel density estimator with a bandwidth of $\sqrt{\sigma^2 + \gamma/2}$, and Parzen (1962) provides the limiting density of the kernel density estimator when $\sqrt{\sigma^2 + \gamma/2} \to 0$ and $N\sqrt{\sigma^2 + \gamma/2} \to \infty$.

□

The following theorem provides the limiting density of the estimator with the primary additional assumption that $(\mu^\star(x))^\rho$ is strictly concave. Afterward, we will move onto results that relax this assumption.

Theorem 2: *Suppose the assumptions from Lemma 1 hold. If $\mu^\star$ is in the interior of the feasible set, $N\sqrt{c_2(\{z_i\}_i)\sigma^2 + c_1(\{z_i\}_i)\gamma/2}/\log(N) \to \infty$, and for $j \in \{1, 2\}$, $c_j(\{z_i\}_i)$ converges in probability to a constant, then there exists $c \in \mathbb{R}^1$ so that,*

$$\sqrt{N(\sigma^2 + \gamma/2)^{d/2}}\left(\hat{f}_i - \mu_i^\star + c(\sigma^2 + \gamma/2)^d\right) \xrightarrow{d} N\left(0, \mu_i^\star/\left(2\sqrt{\pi}\right)^d\right).$$

Proof: Since $W_\gamma(\cdot)$ is differentiable, we can apply the envelope theorem to the dual problem (8) to show that $\nabla_f W_\gamma(f, \mu) = x$. This implies that the $\tilde{f} \in \mathbb{R}^m$ is a critical point of $f \mapsto W_\gamma(f, \mu)$ if and only if it has a corresponding dual variable in (8) of $x = \mathbf{0}$, so f_{unc} is

the unique minimizer of $W_\gamma(\cdot)$ by the proof of Lemma 1 and the definition $w := \exp(x/\gamma)$.[4] This implies $\hat{f} = f_{unc}$ when f_{unc} is feasible. Einmahl and Mason (2005) show that $f_{unc} \xrightarrow{a.s.} \mu^\star$ under the additional assumptions of the theorem when using data dependent bandwidths, as in the statement of the theorem. The result follows from the fact that $\mu^\star$ is in the interior of the feasible set.

□

Remark 1: Algorithm 1 can be slow to converge when γ is chosen to be too small, but our assumption that $\gamma \to 0$ as $N \to \infty$ is not problematic when the assumptions of the theorem hold. To see this note that f_{unc} can be written as $K(\mu \oslash (K\mathbf{1}))$. Evaluating Algorithm 1 at input densities μ and f_{unc} would result in the algorithm first initializing w as $\mathbf{1}_m$ and then define v to be $\mu \oslash (K\mathbf{1})$. The $w-$update would redefine w to be $f_{unc} \oslash K(\mu \oslash (K\mathbf{1}))$, but since this is also equal to $\mathbf{1}$, the algorithm has already converged. For input densities f and μ, it is also generally the case that Algorithm 1 tends to converge at a faster rate for densities that are closer to f_{unc}.

□

Remark 2: Our more general convergence result (Theorem 4) provides the same rate of convergence as Theorem 2 without requiring that $\mu^\star$ is an interior point of the feasible set, so we can compare this rate with the corresponding rate of convergence of shape constrained maximum likelihood estimators. Seregin and Wellner (2010) show that the pointwise minimax absolute error loss of the $\log-$concave constrained maximum likelihood estimators at $x \in \mathcal{A}$ is $N^{-2/(d+4)}$ when $\mu^\star(x)$ is twice differentiable, x is in the interior of the domain of $\mu^\star(\cdot)$, and the Hessian has full rank. Theorem 2 implies that our estimator also obtains these bounds when $\sqrt{\sigma^2 + \gamma/2}$ is chosen so that it converges to zero at the optimal rate, in the sense of minimizing mean squared error, of $O_p(N^{-1/(d+4)})$.

□

In practice, $\sqrt{\sigma^2 + \gamma/2}$ has a more noticeable impact on the resulting density estimator than the individual values of σ and γ, so we recommend fixing γ/σ^2 to be a constant and focusing on the choice of $\sqrt{\sigma^2 + \gamma/2}$. Since increasing γ/σ^2 tends to result in an increase in the rate of convergence and numerical stability of Algorithm 1, our recommendation is $\gamma/\sigma^2 = 8$, which was used in all of the applications and examples given below.

[4]Uniqueness of f_{unc} can also be shown using strict convexity of $W_\gamma(f, \mu)$ in f, which will be shown in Theorem 6.

From a numerical standpoint, it is possible to set $\sqrt{\sigma^2+\gamma/2}$ so that the smoothing provided by γ and σ is negligible and parsimony is almost entirely ensured by the shape constraints. Since the stability of Algorithm 1 is the only limiting factor on how small we can make $\sqrt{\sigma^2+\gamma/2}$, we can use this fact to estimate a lower bound on the value $\sqrt{\sigma^2+\gamma/2}$. Specifically, we can estimate the lowest possible value of $\sqrt{\sigma^2+\gamma/2}$ that still results in convergence of Algorithm 1 to within a given tolerance in a fixed number of iterations, say 100, which can be achieved using a root finding algorithm. This requires an approximation of the density estimate, and which is described in the third chapter in more detail. When running the main optimization algorithm described in third chapter, we recommend increasing γ after this root is found by approximately one fourth and increasing the number of iterations used in Algorithm 1 by a factor of approximately ten to ensure the accuracy of the Hessian.

In our tests this approach resulted in density estimates that were similar to estimates using the method described by Koenker and Mizera (2010), in the sense that the shape constraint binds almost everywhere. However, note that it is possible that the condition $N\sqrt{\sigma^2+\gamma/2}/\log(N) \to \infty$ would not hold in this case.

The mean squared error of f_{unc} is minimized when $\sqrt{\sigma^2+\gamma/2}$ is $O(N^{-1/(d+4)})$. Any of the standard techniques for setting the bandwidth of kernel density estimators are reasonable methods of setting $\sqrt{\sigma^2+\gamma/2}$, including cross validation or a rule-of-thumb bandwidth estimator; for examples of rule-of-thumb bandwidth estimators see (Silverman, 1986; Scott, 1992). Since the shape constraint also helps to ensure the parsimony of the density estimate, these rules should generally be regarded as upper bounds on $\sqrt{\sigma^2+\gamma/2}$. In practice, using Scott's (1992) rule-of-thumb multiplied by 2/3 works well. After dividing each dimension of the dataset, $\{z_{i,j}\}_i$ for $j \in \{1,2,...,d\}$, by $\min(\text{IQR}(\{z_{i,j}\}_i)/1.349, \hat{\sigma}(\{z_{i,j}\}_i))$, combining this with $\gamma/\sigma^2 = 8$, results in $\sigma \approx N^{-1/(d+4)}/3$ and $\gamma \approx N^{-2/(d+4)}4/5$.

The choice of m is likely less critical than that of $\sqrt{\sigma^2+\gamma/2}$. The effect of m on the estimator can be viewed as analogous to the effect of the size of a Gaussian filter on its output. In practice, these filters are often constructed by discretizing a Gaussian density over a mesh with a resolution equal to the standard deviation. Using this as a lower bound in our setting yields, $m\sqrt{\gamma/2}/d \geq 1$. The gain in accuracy from increasing m beyond the point $m\sqrt{\gamma/2}/d = 1$ generally diminishes fairly quickly when m is set to larger values, so we recommend choosing m so that $m = \left\lceil N^{\beta} d/\sqrt{\gamma/2} \right\rceil$, where $\lceil\cdot\rceil$ is the ceiling function, for any $\beta > 0$. In the interest of specificity, we recommend using $m = \left\lceil N^{1/5} d/\sqrt{\gamma/2} \right\rceil$.

If the set of constraints that bind asymptotically is known, it is straightforward to find the asymptotic distribution of $\hat{f}$ when $\mu^\star$ is not $\rho-$concave; however, this is rarely the case in practice. For this reason, finding confidence intervals for estimators subject to shape con-

straints is an active area of research in nonparametric econometrics. The difficulties in these cases are due to the estimators, when viewed as functions of the data, not being sufficiently smooth (either differentiable or Hadamard differentiable) at points in which an inequality constraint goes from slack to active, and these notions of smoothness are requirements of many of the commonly used methods for deriving limiting distributions (Andrews, 2000).

The following theorem establishes the limiting distribution of the estimator conditional on the set of constraints that bind asymptotically. Note that the set of active constraints converges when $\sqrt{\sigma^2 + \gamma/2}$ converges to zero at its optimal rate. Thus, the limiting distribution part of the proof might be useful with a rather large sample size. However, since the set of constraints that binds in a finite sample may not be equal to the set of constraints that bind asymptotically, in many cases the primary value of this result is that it provides the limiting value of the estimator without assuming that $(\mu^\star)^\rho$ is strictly concave.

There are several possible methods to avoid conditioning on the set of active constraints, which is an area of ongoing research. One interesting possibility would be to adopt a similar method as Horowitz and Lee (2017) to the present setting. Their technique involves explicitly defining the set of active constraints as those constraints that would bind otherwise or that would "nearly bind." Asymptotically correct coverage follows from consistency of this conservative estimate of the active set.

Theorem 3: *Suppose the assumptions from Lemma 1 hold. In addition, suppose* $\sigma N \to \infty$. *Then, conditional on the set of constraints that are active asymptotically, say* Ω, *there exists* G, *as defined in (23-25), and* $c \in \mathbb{R}^m$ *such that*

$$\sqrt{N(\sigma^2 + \gamma/2)^{d/2}} \left(\hat{f} - \tilde{\mu} + D_c(\sigma^2 + \gamma/2)^d \right) \xrightarrow{d} N\left(0, G^T D_{\mu^\star/(2\sqrt{\pi})^d} G\right),$$

where $\tilde{\mu}$ *is the Wasserstein projection of* $\mu^\star$ *onto the set of feasible densities. Specifically, if* $\tilde{g}$ *is the minimizer of*

$$\min_g W_\gamma(g^{1/\rho}, \mu^\star) \text{ subject to:}$$

$$sgn(\rho) g \in \mathcal{K},$$

then $\tilde{\mu}_{-k} := \tilde{g}^{1/\rho}$ *and* $\tilde{\mu}_k = m - \sum_i \tilde{g}_i^{1/\rho}$ *in the case of a* $\rho-$*concavity constraint.*

Proof: We only consider the case in which $d = 1$ and $k = m$ for the sake of clarity, but generalizing the proof to higher dimensions is straightforward. Since $d = 1$, we will assume all sets containing the numerical labels of vertices in the mesh are ordered sequentially. For

example, Ω_i will be viewed as the i^{th} largest vertex label in which the constraint binds. Let $\dot{\Omega} := \{i \mid i \notin \Omega \wedge i \pm 1 \in \Omega\}$, the boundary of the complement of Ω.

Since $\sqrt{\sigma^2 + \gamma/2} \to 0$, we have $\sigma \to 0$. Since this is the case, and since $\sigma N \to \infty$ and $\mu^\star$ is continuous, μ converges uniformly to $\mu^\star$ (Parzen, 1962). We can view $\hat{f}$ as a continuous function of μ, so the continuous mapping theorem implies $\hat{f}$ converges to $\tilde{\mu}$.

Let the matrix $A_{|\Omega|\times(|\Omega|+|\dot{\Omega}|)}$ be defined so that $Ag_{\Omega\cup\dot{\Omega}}$ is the second difference of $g_{\Omega\cup\dot{\Omega}}$, so the binding constraints can be denoted by $Ag_{\Omega\cup\dot{\Omega}} \leq \mathbf{0}$.[5] Since we assumed that Ω is known, $\hat{g}$, the vector of Lagrange multipliers, λ, and x are defined by the following system of equations,

$$(x_{\Omega\cup\dot{\Omega}} - x_k) \otimes \hat{g}_{\Omega\cup\dot{\Omega}}^{1/\rho-1}/\rho = -A^T\lambda \tag{1.19}$$

$$(x_{-\Omega\cap-\dot{\Omega}} - x_k) \otimes \hat{g}_{-\Omega\cap-\dot{\Omega}}^{1/\rho-1}/\rho = \mathbf{0} \tag{1.20}$$

$$A\hat{g}_\Omega = \mathbf{0} \tag{1.21}$$

$$\left[\ \left(\hat{g}^{1/\rho}\right)^T \quad m - \textstyle\sum_i \hat{g}_i^{1/\rho}\ \right]^T = \exp(x/\gamma) \otimes (K\,(\mu \oslash (K\exp(x/\gamma)))), \tag{1.22}$$

where the final equality results from combining the first order conditions of the optimization problem defining $W_\gamma(g^{1/\rho}, \mu)$, which are given by (9) and (10).

Theorem 5 will establish that $W_\gamma(g^{1/\rho}, \mu)$ is strictly convex in g. Since this is the case, the solution of this system is unique. The implicit function theorem, applied to (19)-(22), implies that we can view $\hat{g}$, and thus also $\hat{f}$, as a differentiable function of μ. After simplifying this system, we will find this limiting density using the delta method.

We will begin by deriving $\nabla_\mu x$ using (22). Recall $w := \exp(x/\gamma)$, and let $h_1(x, \mu)$ be defined as,

$$h_1(x, \mu) := \hat{f} - D_w K\,(\mu \oslash (Kw))\,.$$

(22) can be written as $h_1(x, \mu) = \mathbf{0}$. Recall the following equalities from the previous section: $\psi = D_w K D_v$, $f = D_w K v$, and $\mu = D_v K w$. These imply

$$\nabla_\mu h_1(x, \mu) = -D_w K D_{\mathbf{1}\oslash(Kw)}$$

$$= -D_w K D_{v\oslash\mu} = -\psi D_{\mathbf{1}\oslash\mu},$$

[5] In the interest of the simplicity of the exposition, we are defining A using local concavity constraints rather than Afriat's (1972) formulation of concavity constraints. The nonzero elements of each row of A are given by $(A_{i,j-1}, A_{i,j}, A_{i,j+1}) = sgn(\rho)(1, -2, 1)$. Note that $j > 1$, and, since the sets Ω and $\Omega \cup \dot{\Omega}$ are ordered by the vertex labels, this matrix is in row echelon form by construction. Thus, A has full rank, so AA^T is nonsingular.

and

$$\nabla_w h_1(x,\mu) = D_w K D_{\mu\oslash(Kw)^2} K - D_{K(\mu\oslash(Kw))}$$

$$= D_w K D_v D_{\mathbf{1}\oslash\mu} D_v K - D_{\hat{f}\oslash w}$$

$$= \left(\psi D_{\mathbf{1}\oslash\mu}\psi^T - D_{\hat{f}}\right) D_{\mathbf{1}\oslash w}.$$

The implicit function theorem implies,

$$\nabla_\mu w = D_w \left(\psi D_{\mathbf{1}\oslash\mu}\psi^T - D_{\hat{f}}\right)^+ \psi D_{\mathbf{1}\oslash\mu}.$$

Since $x = \gamma \log(w)$, we have,

$$\nabla_\mu x = \gamma \left(\psi D_{\mathbf{1}\oslash\mu}\psi^T - D_{\hat{f}}\right)^+ \psi D_{\mathbf{1}\oslash\mu}.$$

(21) implies that each element in $\hat{g}_\Omega$ can be expressed as a mean of its neighbors, so we can express all of the elements of $\hat{g}_\Omega$ as a weighted mean of $\hat{g}_{\dot{\Omega}}$. In other words, there exists C such that $\hat{g}_{\Omega\cup\dot{\Omega}} = C\hat{g}_{\dot{\Omega}}$. (19) implies that, given x and $g_{\dot{\Omega}}$, λ is given by,

$$\lambda = -\left(AA^T\right)^{-1} A D_{x_{\Omega\cup\dot{\Omega}} - x_k} (C\hat{g}_{\dot{\Omega}})^{1/\rho - 1}/\rho.$$

We will use the additional $|\Theta|$ equations in (19) to define $\hat{f}_{\dot{\Omega}}$. After replacing each instance of $\hat{g}_{\dot{\Omega}}$ with $\hat{f}^\rho_{\dot{\Omega}}$ and using this definition of λ, we can write (19) as $h_2(x, \hat{f}_{\dot{\Omega}}) = \mathbf{0}$, where $h_2(x, \hat{f}_{\dot{\Omega}})$ is defined as

$$(x_{\dot{\Omega}} - x_k) \otimes \hat{f}^{1-\rho}_{\dot{\Omega}} + \tilde{A}\left((x_{\Omega\cup\dot{\Omega}} - x_k) \otimes (C\hat{f}^\rho_{\dot{\Omega}})^{1/\rho - 1}\right),$$

and $\tilde{A} := A^T_{\cdot,\dot{\Omega}} \left(AA^T\right)^{-1} A$. This implies

$$\nabla_{\hat{f}_{\dot{\Omega}}} h_2(\cdot) = (1-\rho)\left(D_{(x_{\dot{\Omega}} - x_k)} \hat{f}^{-\rho}_{\dot{\Omega}} + \tilde{A} D_{x_{\Omega\cup\dot{\Omega}} - x_k} D_{(C\hat{f}^\rho_{\dot{\Omega}})^{1/\rho - 2}} C D_{\hat{f}^{\rho-1}_{\dot{\Omega}}}\right),$$

and

$$\nabla_{x_{\Omega\cup\dot{\Omega}} - x_k} h_2(\cdot) = B + \tilde{A} D_{(C\hat{f}^\rho_{\dot{\Omega}})^{1/\rho - 1}},$$

where $B_{|\dot{\Omega}|\times|\Omega\cup\dot{\Omega}|}$ is defined so that $B_{i,j}$ is equal to $\hat{f}^{1-\rho}_{\dot{\Omega}}$ when i, j satisfies $\dot{\Omega}_i = \{\Omega \cup \dot{\Omega}\}_j$ and zero otherwise. The implicit function theorem implies

$$\nabla_{x_{\Omega\cup\dot{\Omega}} - x_k} \hat{f}_{\dot{\Omega}} = -\left(D_{(x_{\dot{\Omega}} - x_k)} \hat{f}^{-\rho}_{\dot{\Omega}} + \tilde{A} D_{(x_{\Omega\cup\dot{\Omega}} - x_k)\otimes(C\hat{f}^\rho_{\dot{\Omega}})^{1/\rho - 2}} C D_{\hat{f}^{\rho-1}_{\dot{\Omega}}}\right)^{-1}$$

$$\cdot \left(B + \tilde{A} D_{(C\hat{f}^\rho_{\dot{\Omega}})^{1/\rho - 1}}\right)/(1-\rho),$$

so

$$G_{\dot{\Omega},\cdot} := \nabla_\mu \hat{f}_{\dot{\Omega}} = \nabla_{x_{\Omega\cup\dot{\Omega}}-x_k} \hat{f}_{\dot{\Omega}} P \nabla_\mu x =$$

$$-\left(D_{(x_{\dot{\Omega}}-x_k)}\hat{f}_{\dot{\Omega}}^{-\rho} + \tilde{A}D_{(x_{\Omega\cup\dot{\Omega}}-x_k)\otimes(C\hat{f}_{\dot{\Omega}}^{\rho})^{1/\rho-2}}CD_{\hat{f}_{\dot{\Omega}}^{\rho-1}}\right)^{-1}$$

$$\cdot\left(B + \tilde{A}D_{(C\hat{f}_{\dot{\Omega}}^{\rho})^{1/\rho-1}}\right)P\left(\psi D_{\mathbf{1}\oslash\mu}\psi^T - D_{\hat{f}}\right)^{+}\psi D_{\mathbf{1}\oslash\mu}\gamma/(1-\rho), \tag{1.23}$$

where $P_{\left|\Omega\cup\dot{\Omega}\right|\times m} := \nabla_x(x_{\Omega\cup\dot{\Omega}} - x_k)$ is defined so that $P_{i,j}$ is equal to 1 when i, j satisfies $\{\Omega\cup\dot{\Omega}\}_i = j$, -1 when $j = m$, and zero otherwise.

Note that (20) implies $x_i = x_k$ for all $i \in \{j \mid j \notin \Omega \cap j \notin \dot{\Omega}\}$. The proof of Theorem 2 shows that this condition defines $f_{unc,i}$, so we have $f_{unc,i} = \hat{f}_i$ for all such i. This implies,

$$G_{-\Omega\cap-\dot{\Omega},\cdot} := \nabla_\mu \hat{f}_{-\Omega\cap-\dot{\Omega}} = K_{-\Omega\cap-\dot{\Omega},\cdot} \tag{1.24}$$

Lastly, writing the equations defining $\hat{f}_\Omega$ in terms of $\hat{f}_{\dot{\Omega}}$ yields $\hat{f}_\Omega = (C\hat{f}_{\dot{\Omega}}^{\rho})^{1/\rho}$, so we have,

$$G_{\Omega,\cdot} := \nabla_\mu \hat{f}_\Omega = D_{(C\hat{f}_{\dot{\Omega}}^{\rho})^{1/\rho-1}}CD_{\hat{f}_{\dot{\Omega}}^{\rho-1}}G_{\dot{\Omega},\cdot} \tag{1.25}$$

□

There are also a few options for testing if a population density satisfies a shape constraint. Hypothetically, one could consistently test the null hypothesis that a population density is $\rho-$concave using any consistent shape constrained density estimator. This can be done by estimating the population distribution subject to the shape constraint and then using one of the classic tests for whether or not the empirical distribution of the data is equal to this estimate; see for example (Smirnov, 1948; Anderson and Darling, 1952). In these cases, choosing a test with a statistic that is closely related to the fidelity criterion used for estimation allows for a more straightforward interpretation of the result. For example, if the test statistic is equal to the fidelity criterion that the estimator optimizes, we would reject the null if and only if we would also reject the null for every density that satisfies the shape constraint.

The following theorem provides the distribution of $W_\gamma(f_{unc}, \mu)$ and a consistent test for the null hypothesis that $\mu^\star(x)$ satisfies the shape constraint based on this distribution. The method also has the straightforward interpretation from the preceding paragraph, so, if we denote the set of $\rho-$concave densities by $\mathcal{K}_\rho$, the null is rejected $\min_{f\in\mathcal{K}_\rho} W_\gamma(f,\mu) - W_\gamma(f_{unc},\mu)$ is statistically significant. Since $W_\gamma(f,\mu)$ is differentiable in f, this can be achieved without conditioning on the set of active constraints.

Theorem 4: *Suppose the assumptions from Lemma 1 hold and that $\sigma N \to \infty$. Let T be defined as $N(\sigma^2+\gamma/2)^{d/2}\left(W_\gamma(\mu^\star,\mu) - W_\gamma(f_{unc},\mu)\right)$, ψ as the optimal coupling between μ to*

f_{unc}, $\psi^\star$ as the optimal coupling between $\mu^\star$ and itself, $B^\star$ as $\gamma(D_{\mu^\star} - \psi^\star D_{\mathbf{1}\oslash\mu^\star}\psi^{\star T})^+/2$, and B as $\gamma(D_{f_{unc}} - \psi D_{\mathbf{1}\oslash\mu}\psi^T)^+/2$. Then, $T \xrightarrow{d} Z^T B^\star Z$, where the iid elements of $Z \in \mathbb{R}^m$ are distributed $Z_i\sqrt{N(\sigma^2+\gamma/2)^{d/2}} \sim N\left(\triangle\mu_i^\star/2, \mu_i^\star/\left(2\sqrt{\pi}\right)^d\right)$.

Also, the hypothesis that $\mu^\star$ satisfies the shape constraint can be consistently tested at a significance level of α by rejecting the null when $N(\sigma^2+\gamma/2)^{d/2}\left(W_\gamma(\hat{f},\mu) - W_\gamma(f_{unc},\mu)\right) \geq c_\alpha$, where c_α satisfies $P(X^T B X \geq c_\alpha) = \alpha$ and the iid elements of $X \in \mathbb{R}^m$ are distributed $X_i\sqrt{N(\sigma^2+\gamma/2)^{d/2}} \sim N\left(0, \mu_i/\left(2\sqrt{\pi}\right)^d\right)$.

Proof: The gradient and Hessian of $W_\gamma(f,\mu)$ with respect to f are $\nabla_f W_\gamma(f,\mu) = x_{\mu,f}$ and $\nabla_{f,f} W_\gamma(f,\mu) = \nabla_f x_{\mu,f} = \gamma(D_f - \psi_{\mu,f} D_{\mathbf{1}\oslash\mu}\psi_{\mu,f}^T)^+$, which are derived in the proof of the next theorem. Thus, the second order Taylor series expansion about $\mu^\star = f_{unc}$ is given by,

$$W_\gamma(\mu^\star,\mu) = W_\gamma(f_{unc},\mu) + x_{\mu,f_{unc}}^T(\mu^\star - f_{unc})$$
$$+(f_{unc}-\mu^\star)^T B(f_{unc}-\mu^\star) + O(\|f_{unc}(\mu)-\mu^\star\|^3).$$

Since $f_{unc} := \arg\min_f W_\gamma(f,\mu)$ and $\nabla_f W_\gamma(f,\mu) = x$, we have $x = \mathbf{0}$. Given the general discretized densities $\mu_0, \mu_1 \in \mathbb{R}^m$, the next theorem will also shows that the matrix $D_{\mu_0} - \psi D_{\mathbf{1}\oslash\mu_1}\psi^T$ has one eigenvalue that is zero and $m-1$ eigenvalues that are strictly positive. Note that the pseudo inverse is a continuous function when its domain is restricted to the set of matrices with the same rank (Stewart, 1969). Since μ and f_{unc} converge in probability to $\mu^\star$, the Slutsky theorem implies $B \xrightarrow{p} B^\star$. Lemma 1 implies that $N(\sigma^2+\gamma/2)^{d/2}\|f_{unc}-\mu^\star\|^3$ converges to zero in probability at a rate of $O_p\left(N^{-1/2}(\sigma^2+\gamma/2)^{-d/4}\right)$. Combining these results with the limiting distribution of f_{unc} given in Lemma 1 implies that the Taylor series expansion given above can be written as,

$$T := N(\sigma^2+\gamma/2)^{d/2}\left(W_\gamma(\mu,\mu^\star) - W_\gamma(f_{unc},\mu)\right) \xrightarrow{d} Z^T B^\star Z.$$

Since μ is a consistent estimator for $\mu^\star$, we also have $X \xrightarrow{d} Z$, so we also have $T \xrightarrow{d} X^T B^\star X$.

To show the test is consistent, suppose $\mu^\star(x)$ is not $\rho-$concave. Recall that $W_\gamma(\hat{f},\mu)$ converges to the metric $W_0(\hat{f}(x),\mu(x))$ asymptotically (Benamou et al, 2015) and that Wasserstein distance metrizes weak convergence of distributions (and convergence in the first two moments). Continuity of the functions $\hat{f}(x), \mu(x)$ and $\mu^\star(x)$ implies that the distributions corresponding to these densities converge weakly if and only if the densities themselves converge to one another in probability. Since $\hat{f}$ is in the feasible set and $\mu^\star$ is not, $\hat{f}(x)$ cannot converge to $\mu^\star(x)$ in probability, so $W_0(\mu^\star,\hat{f})$ does not converge to zero. Also, asymptotically we have $W_0(\hat{f},\mu) - W_0(\mu^\star,\mu) > W_0(\mu^\star,\hat{f})$ by the triangle inequality, so $N(\sigma^2+\gamma/2)^{d/2}\left(W_\gamma(\hat{f},\mu) - W_\gamma(f_{unc},\mu)\right)$ diverges under the alternative hypothesis.

□

1.4 The Optimization Method

In this section we will briefly outline a trust region sequential quadratic programming algorithm to find the global minimum of (13) and (14), while the third chapter describes this method in more detail. To do this, we will require the gradient and the Hessian of $W_\gamma(\mu, g^{1/\rho})$. The following Theorem provides these values and shows that the optimization problem is convex for cases in which $\rho \neq 0$. The derivations for the $\log-$concave case, are given in the final chapter, along with definitions and properties of the derivatives in a more general context than $\rho-$concavity. For notational convenience, the Hessian given below corresponds to the case in which the index k is set equal to m; although this is not a requirement of the theorem.

Theorem 5: *The gradient of the* $W_\gamma(\mu, g^{1/\rho})$ *is*

$$r := \nabla_g W_\gamma(\mu, g^{1/\rho}) = D_{g^{1/\rho-1}/\rho}(x_{-k} - x_k), \tag{1.26}$$

and the Hessian is

$$H := \nabla_g^2 W_\gamma(\mu, g^{1/\rho}) = ABA^T + C, \tag{1.27}$$

where $A := \begin{bmatrix} D_{g^{1/\rho-1}/\rho} & -g^{1/\rho-1}/\rho \end{bmatrix}$, $B := \gamma(D_{g^{1/\rho}} - \psi D_{\mathbf{1}\oslash\mu}\psi^T)^+$, *and* $C := \frac{1-\rho}{\rho^2}\ D_{g^{1/\rho-2}}$ $D_{x_{-k}-x_k}$. *In addition,* $W_\gamma(f, \mu)$ *is strictly convex in* f *when* k *is chosen to be arg min*$_i$ x_i *and* $\gamma > 0$, *and the optimization problem given in (13)-(14) is convex.*

Proof: Since $W_\gamma(\mu, g^{1/\rho})$ is differentiable, the envelope theorem implies that the gradient of the objective function in (12) is equal to the gradient of the function given in (13), so $\nabla_g W_\gamma(\mu, g^{1/\rho}) = D_{g^{1/\rho-1}/\rho}(x_{-k} - x_k)$.

The derivative of (26) yields the sum of two matrices. Specifically,

$$H = \nabla_g^2 W_\gamma(\mu, f(g))$$

$$= D_{g^{1/\rho-1}/\rho}\nabla_g(x_{-k}(g) - x_k(g)) + \left(\nabla_g g^{1/\rho-1}/\rho\right) D_{x_{-k}-x_k}. \tag{1.28}$$

We will begin by deriving the first term, which will require several intermediate derivatives.

First, since, $f = (g^{1/\rho}, m - \mathbf{1}\cdot g^{1/\rho})$, we have

$$\nabla_g f(g) = \begin{bmatrix} D_{g^{1/\rho-1}/\rho} \\ -g^{1/\rho-1}/\rho \end{bmatrix}.$$

Second, we will view w as a function of f in order to find $\nabla_f w(f)$. (9) and (10) can be combined to yield the equality $f - D_w K\left(\mu \oslash (Kw)\right) = 0$, and implicit differentiation of this equality implies,

$$\nabla_f w(f) = (D_{f \oslash w} - \psi D_{\mathbf{1} \oslash \mu} \psi^T D_{\mathbf{1} \oslash w})^+.$$

Third, the definition $w := \exp\left(x/\gamma\right)$ implies $x = \gamma \log(w)$, so

$$\nabla_w x(w) = \gamma D_{\mathbf{1} \oslash w}.$$

Lastly, let $\tilde{x}(x) := x_{-k} - x_k$, so $\nabla_x \tilde{x}(x) = \begin{bmatrix} I & -\mathbf{1} \end{bmatrix}_{m-1 \times 1}$. After combining all four derivatives, we have

$$D_{g^{1/\rho-1}/\rho} \nabla_g \tilde{x}(x(w(f(g)))) = D_{g^{1/\rho-1}/\rho} \nabla_x \tilde{x}(x) \nabla_w x(w) \nabla_f w(f) \nabla_g f(g)$$

$$= \gamma D_{g^{1/\rho-1}/\rho} \begin{bmatrix} I & -\mathbf{1} \end{bmatrix} D_{\mathbf{1} \oslash w} (D_{f \oslash w} - \psi D_{\mathbf{1} \oslash \mu} \psi^T D_{\mathbf{1} \oslash w})^+ \begin{bmatrix} D_{g^{1/\rho-1}/\rho} \\ -g^{1/\rho-1}/\rho \end{bmatrix}$$

$$= \gamma \begin{bmatrix} D_{g^{1/\rho-1}/\rho} & -g^{1/\rho-1}/\rho \end{bmatrix} (D_f - \psi D_{\mathbf{1} \oslash \mu} \psi^T)^+ \begin{bmatrix} D_{g^{1/\rho-1}/\rho} \\ -\left(g^{1/\rho-1}\right)^T/\rho \end{bmatrix}$$

$$= ABA^T.$$

Since $\nabla_g g^{1/\rho-1}/\rho = (1-\rho)/\rho^2 D_{g^{1/\rho-2}}$, the second matrix in (28) is given by C.

Convexity requires that this Hessian is positive semidefinite. If k is chosen as the index corresponding to the minimum of x, then $x_{-k} - x_k \geq 0$. Since this is the case, C is a diagonal matrix with nonnegative diagonal elements, so C is positive semidefinite.

Next we will establish that ABA^T is positive definite in several steps. First, note that ABA^T is symmetric if B is symmetric. Also, since D_f and $\psi D_{\mathbf{1} \oslash \mu} \psi^T$ are symmetric B^+ is also symmetric. Since the Moore-Penrose pseudo inverse preserves symmetry, B is also symmetric. We will proceed by establishing a few intermediary results on the components of ABA^T.

Since $D_{\mathbf{1} \oslash \mu}$ is positive semidefinite, $\psi D_{\mathbf{1} \oslash \mu} \psi^T$ is as well. Since ψ is a coupling of the densities μ and f, we have

$$D_{\mathbf{1} \oslash f} \psi D_{\mathbf{1} \oslash \mu} \psi^T \mathbf{1}$$

$$= D_{\mathbf{1} \oslash f} \psi \mathbf{1} = \mathbf{1}.$$

In other words, $\mathbf{1}$ is an eigenvector of $D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T$ with a corresponding eigenvalue of 1. The Perron-Frobenius theorem states that an $m \times m$ matrix with all positive elements and columns that sum to one has a unique eigenvalue that is equal to one and $m-1$ eigenvalues that are strictly less than one. Note that each eigenvalue, say λ, and corresponding eigenvector, say p, of $D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T D_{\mathbf{1}\oslash f}$ satisfies,

$$\left(D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T - \lambda I\right) p = 0,$$

$$\implies \left(I - D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T - \tilde{\lambda} I\right) p = 0$$

so, p is an eigenvector of $I - D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T D_{\mathbf{1}\oslash f}$, with an eigenvalue corresponding to $\tilde{\lambda} := 1 - \lambda$. This implies that $I - D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T$ is a positive semidefinite matrix with rank $m-1$. Since multiplication by a positive definite matrix and applying the pseudoinverse preserve both the rank and the signs of the eigenvalues, $B = \gamma\left(D_f\left(I - D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T\right)\right)^+$ is also a positive semidefinite matrix with rank $m-1$.

Observation 7.1.8 in Horn and Johnson (1990) implies that the nullspace of ABA^T is the same as the nullspace of BA^T. Since the eigenvector of B that corresponds to the eigenvalue of zero is $\mathbf{1}$, ABA^T is positive definite if there does not exist $v \in \mathbb{R}^{m-1}$ such that $A^T v = \mathbf{1}_m$. This system of equations is equivalent to $g_i^{1/\rho-1} v_i = \rho$ for all $i \in \{1, ..., m-1\}$ and $\sum_i g_i^{1/\rho-1} v_i = -\rho$, which does not have a solution, so ABA^T is positive definite.

This, along with the fact that the constraints in (14) are equivalent to constraining $sgn(\rho)g$ to be in the set of concave functions, which is a convex cone, implies that the optimization problem is convex.

□

Remark 3: Choosing k to be $arg\ min_i\ x_i$ is a sufficient, but not necessary, condition to guarantee convexity. Choosing k to correspond to the element on the boundary of the mesh over $\mathcal{A}$, with the lowest corresponding value of w_i, ensures that the density estimate satisfies the shape constraints everywhere on the interior of its domain and rarely results in the objective function being nonconvex along the convergence path. Note that the third chapter outlines a method to formulate the estimator using $g \in \mathbb{R}^m$ rather than $g \in \mathbb{R}^{m-1}$, along with an alternative proof of this result.

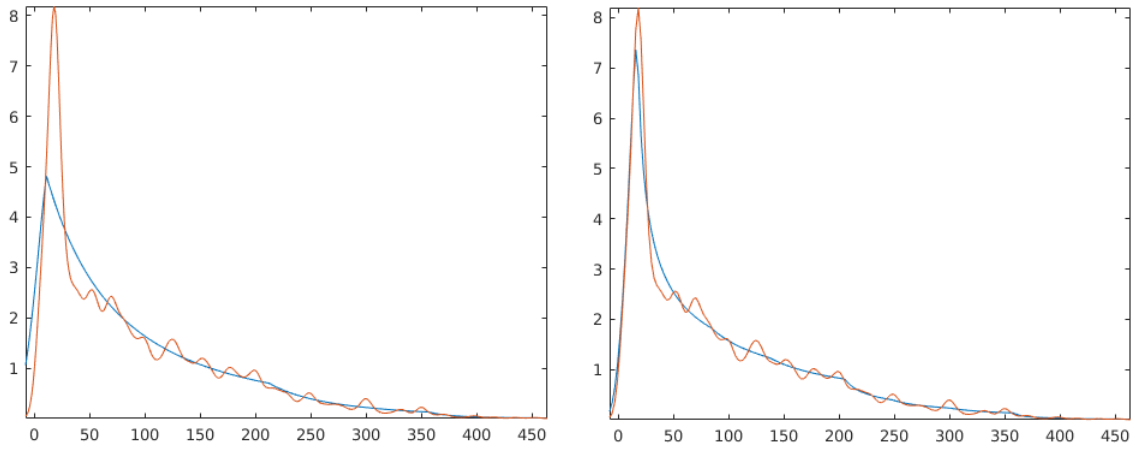
□

Having already derived the gradient and Hessian of $W_\gamma(\mu, g^{1/\rho})$, it is straightforward to create a trust region algorithm. The algorithm takes an initial density estimate, $f^{(0)}$, as input and in iteration i the algorithm solves

$$\Delta \leftarrow \arg\min_d \left\{ d^T H d/2 + d^T r \;\middle|\; \left(d + g_{-k}^{(i-1)}\right) sgn(\rho) \in \mathcal{K},\ \|d\| \leq c \right\}.$$

If the value of the objective function evaluated at $g_{-k}^{(i-1)} + \Delta$ results in an improvement over its value at $g_{-k}^{(i-1)}$, then $g_{-k}^{(i)}$ is defined to be $g_{-k}^{(i-1)} + \Delta$. If the improvement was significant, then the radius of the trust region, c, is increased, and otherwise it is decreased and the value of $g_{-k}^{(i)}$ is defined to be $g_{-k}^{(i-1)}$. Note that third chapter provides more detail on this, as well as methods for finding products with H and its pseudoinverse, both with a time complexity of $O(m \log(m))$.

Figures 1 and 2 illustrate two examples of the output of the optimization procedure fully described chapter 3. Figure 1 provides density estimates of the rotational velocity of stars that are constrained to be $-2-$concave and $-1/2-$concave, respectively. Figure 2 provides a plot of a two dimensional density; to illustrate the tail behavior of the density more clearly, the logarithm of the density is shown. This estimator uses a dataset containing the height and left middle finger length of 3,000 British criminals that was analyzed by Macdonell (1902) and Student (1908).



(a) (b)

Figure 1: The red density in each plot is a kernel density estimator of the rotational velocity of stars from from Hoffleit and Warren (1991). The blue density is the density estimate. ρ was set equal to -0.5 in (a) and -2 in (b). γ was set using using the first method described above, so the constraints are binding almost everywhere.

The data used to generate both figures are also used by Koenker and Mizera (2010) to compare $\log-$concave density estimates with $-1/2-$concave density estimates. In the case of the dataset for the rotational velocity of stars, they show that the former provides a monotonic density in the region in which the speed of rotation is strictly positive, while the

latter density has a peak near the mode of the kernel density estimator shown in Figure 1. This peak is also present in both of the shape-constrained densities shown in Figure 1.

For the dataset used in Figure 2, Koenker and Mizera (2010) show that the logarithm of the maximum likelihood density estimator subject to a log $-$concavity constraint is below -24 near the observation at the very top of Figure 2, so observations this far from the rest of the data would be fairly unlikely to occur if the density was in fact log $-$concave. The logarithm of the $-1/2-$concave density given below is approximately -7.2 near this observation.

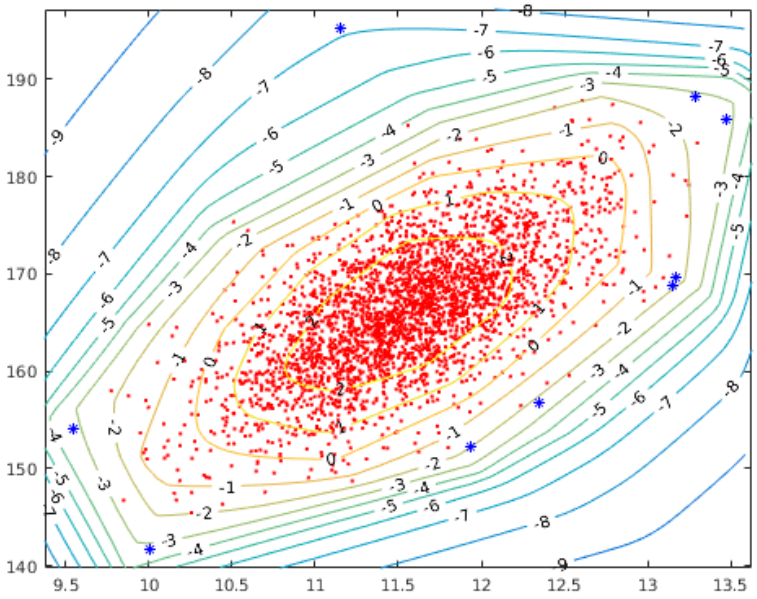


Figure 2: The data points (shown in red) consists of the height and finger length of 3,000 criminals from Macdonell (1902). The points on the convex hull of the data are illustrated with blue asterisks. The contour plot depicts the logarithm of the $-1/2-$concave density estimate to illustrate the tail behavior of the density.

The next section provides results on a generalization of the optimization problem given in (13)-(14). One of these results provides sufficient conditions for convexity of this more general optimization problem.

1.5 More General Shape-Constraints

Although log $-$concavity, and $\rho-$concavity more generally, are the most commonly studied shape constraints, the results provided in the previous sections are also applicable to a large

class of new shape constraints. Specifically, this section considers solutions to the general optimization problem given by,

$$\min_{g} W_\gamma(\mu, \alpha(g)) \tag{1.29}$$

$$\text{subject to: } \cap_i \beta_i(g) \leq 0, \tag{1.30}$$

where $\alpha_i : \mathbb{R}^{m-1} \to \mathbb{R}_+$ and $\beta_i : \mathbb{R}^{m-1} \to \mathbb{R}$. Analogous to the estimator described in the preceding sections, we will denote the minimizer of (29)-(30) as $\tilde{g}$ and the generalized density estimator as $\tilde{f} := f(g)$, where $f(g) := [\alpha(g)^T, m - \mathbf{1}^T\alpha(g)]^T$. The following theorem provides sufficient conditions for the results provided in Theorems 2, 4, and 5 to hold in this more general setting. The statements of these theorems were purposefully ambiguous in regards to the constraint, so we simply provide additional requirements on the functions $\alpha(\cdot)$ and $\beta(\cdot)$ for these generalizations. We also provide sufficient conditions for Theorem 6, but this requires restating the result in its entirety using the new notation. Lastly, point (4) of the theorem provides a new result for strict convexity of $W_\gamma(\mu, \alpha(g))$ in the neighborhood of its global minimum, without requiring the assumption that $\alpha(\cdot)$ is convex or concave.

Theorem 6: *Let $i(x)$ be defined as arg $min_i \|x - \mathbf{a}_i\|$, Θ as the set of proper and uniformly continuous density functions, Ω_N as $\{g \mid \cap_i \beta_i(g) \leq 0 \wedge \alpha(g) \in \mathbb{R}^m \wedge \sum_i \alpha_i(g) \leq m\}$, the set Λ so that $f(x) \in \Lambda$ if and only if there exists a sequence $\{g^{(N)}\}_N^\infty$ such that $g^{(N)} \in \Omega_N$ for all N sufficiently large and $f(x) = \lim_{N\to\infty}[\alpha(g^{(N)})^T, m - \mathbf{1}^T\alpha(g^{(N)})]_{i(x)}$.*

If $\mu^\star(x) \in \Theta$, both of the sets $\Theta \cap \Lambda$ and Ω_N are nonempty, the function $\alpha(\cdot)$ is invertible, then the following additional conditions are sufficient for the applicability of Theorems 2, 4, 5, and 6 to $\tilde{f}$.

(1) If the codomain of the function $g \mapsto [\alpha(g)^T, m - \mathbf{1}^T\alpha(g)]$ contains f_{unc} then Theorems 2 and 5 hold for $\tilde{f}$.

(2) If each of the functions $\beta_i(g)$, as well as $\alpha(g)$, are differentiable in the neighborhood of $\tilde{g}$ asymptotically, $\tilde{g}$ converges to a point on the interior of the domains of these functions, and $\nabla_g \beta_i(g) \mid_{g=\tilde{g}} \neq 0$ for each i and $\nabla_g \alpha(g) \mid_{g=\tilde{g}} \neq 0$, then Theorem 4 holds for $\tilde{f}$.

(3) Suppose $\{g \mid \beta_i(g) \leq 0\}$ is convex for all i, and $\alpha_j(g)$ is convex (respectively, concave) for all j. If $k := arg\ min_i\ x_i$ ($k := arg\ max_i\ x_i$), then (29)-(30) is a convex optimization problem. Also, the gradient and Hessian of $W_\gamma(\mu, \alpha(g))$ exist almost everywhere, and at these points they are given by

$$\nabla_g W_\gamma(\mu, \alpha(g)) = (x_{-k} - x_k)\nabla_g \alpha(g),$$

$$\nabla_g^2 W_\gamma(\mu, \alpha(g)) = ABA^T + C,$$

where $A := \begin{bmatrix} \nabla_g \alpha(g) & -\nabla_g \alpha(g)\mathbf{1} \end{bmatrix}$, $B := \gamma(D_{\alpha(g)} - \psi D_{\mathbf{1} \oslash \mu} \psi^T)^+$, *and* $C := \sum_j (x_j - x_k) \nabla_g^2 \alpha_j(g)$.

(4) Let $d : \mathbb{R}^{m-1} \times \mathbb{R}^{m-1} \to \mathbb{R}^1_+$ *denote an arbitrary distance measure. If* $\alpha(g) \in C^2$ *and* $\nabla_g \alpha(g)$ *has full rank, then there exists* $\delta > 0$ *such that* $W_\gamma(\mu, \alpha(g))$ *is convex in* g *when* $d([\alpha(g)^T, m - \mathbf{1}^T \alpha(g)], f_{unc,-k}) \leq \delta$.

Proof: (1): The proof of theorems 2 and 5 do not use the $\rho-$concavity constraint, so they still hold as long as there exists g_{unc} such that $f_{unc} = f(g_{unc})$. Since f_{unc} is the global minimum of $f \mapsto W_\gamma(f, \mu)$, $\tilde{f}$ will be equal to f_{unc} whenever g_{unc} is feasible.

(2): The proof given for Theorem 4 uses the first order delta method, which requires the conditions given in the theorem. These conditions are also sufficient for the application of the continuous mapping theorem, which was used to show convergence in probability of the estimator.

(3): Since each $\alpha_j(g)$ is convex, Aleksandrov's theorem implies that $\nabla_g \alpha_j(g)$ and $\nabla_g^2 \alpha_j(g)$ exist almost everywhere. We will begin by deriving the gradient and Hessian at these points. Let $\omega(x, y, g)$ be defined as

$$x_{-k}^T \alpha(g_{-k}) + x_k \left(m - \mathbf{1}_{m-1}^T \alpha(g_{-k})\right) + y^T \mu - \gamma \sum_{i,j} \exp\left((x_i + y_j - M_{ij})/\gamma\right),$$

so that we can write $W_\gamma(\mu, \alpha(g))$ as

$$\max_{x,y} \omega(x, y, g).$$

Since $\omega(x, y; g)$ is differentiable in all of its arguments, the envelope theorem implies

$$\nabla_g W_\gamma(\mu, \alpha(g)) = \nabla_g \omega(x, y, \alpha(g)) = (x_{-k} - x_k) \nabla_g \alpha(g).$$

By the same logic used in the proof of Theorem 4, we can write the Hessian as,

$$\nabla_g^2 W_\gamma(\mu, \alpha(g)) = \left(\nabla_g f(g)\right) \left(\nabla_f x(w(f))\right) \left(\nabla_g f(g)\right)^T + \sum_{l \neq k} (x_l - x_k) \nabla_g^2 \alpha_l(g)$$

$$= ABA^T + C.$$

ABA^T is positive definite by the same argument used in Theorem 6, which is also expanded on further in the proof of statement (2). When each $\alpha_j(g)$ is convex and $\nabla_g^2 \alpha_j(g)$ exists, then $\nabla_g^2 \alpha_j(g)$ is a positive semidefinite matrix. In this case, $k := \arg\min_i x_i$, so each element of $(x_{-k} - x_k \mathbf{1})$ is nonnegative. This implies each term in the sum defining C is a positive semidefinite matrix, so C is positive semidefinite. Also, when each $\alpha_j(g)$ is concave and $\nabla_g^2 \alpha_j(g)$ exists, then $\nabla_g^2 \alpha_j(g)$ is negative semidefinite. Choosing $k := \arg\max_i x_i$ implies

$(x_j - x_k)\nabla_g^2 \alpha_j(g)$ is a positive semidefinite matrix, so C is also positive semidefinite in this case.

Since $\{g \mid \beta_i(g) \leq 0\}$ is convex for all i, the intersection of these sets is also convex. This, combined with the strict convexity of $W_\gamma(\mu, \alpha(g))$ in g, implies convexity of the optimization problem given in (29)-(30).

(4): The argument given in Remark 1 implies that each of the elements of x are the same when $f(g) = f_{unc}$. When this occurs, we have $x_{-k} - x_k = 0$, so $C = \mathbf{0}_{m-1 \times m-1}$ when $f(g) = f_{unc}$.

The argument in the second to last paragraph of Theorem 6 implies that ABA^T is positive definite when $\nabla_g \alpha(g)$ has full rank and there does not exist $v \in \mathbb{R}^{m-1}$ such that $A^T v = \mathbf{1}_m$. This system of equations requires $(\nabla_g \alpha(g))\, v = \mathbf{1}_{m-1}$ and $-\mathbf{1}_{m-1}^T \nabla_g \alpha(g) v = 1$. However, if v satisfies $(\nabla_g \alpha(g))\, v = \mathbf{1}_{m-1}$, then $-\mathbf{1}_{m-1}^T \nabla_g \alpha(g) v = 1 - m$, so there is not a solution to this system of equations. This, combined with the fact that $\nabla_g \alpha(g)$ has full rank, implies ABA^T is positive definite.

Since $\alpha(g) \in C^2$, the eigenvalues of the $\nabla_g^2 W_\gamma(\mu, \alpha(g))$ are continuous in g. Since we have shown that these eigenvalues are strictly positive when $f(g) = f_{unc}$, continuity implies that there exists $\delta > 0$ such that all eigenvalues are nonnegative when $d(f(g), f_{unc}) \leq \delta$.

□

Remark 4: Some of the assumptions given above were made to simplify the exposition rather than necessity. For example, we can replace assumptions regarding differentiability and rank for all g with similar assumptions in the neighborhood of $\tilde{g}$. The assumption regarding the existence of the inverse of $\alpha(\cdot)$ is worth mentioning in particular. Cases in which either $\alpha(\cdot)$ or $\alpha^{-1}(\cdot)$ cannot be expressed in a closed form appear to be fairly common, but closed form solutions are not a requirement of the theorem since they can be replaced with their numerical counterparts. The application provided in the next section is one example of this case.

□

Mechanism design appears to be a particularly fruitful source for applications of this generalized shape constrained density estimator. For example, consider a private values auction model in which bidders have valuations that are drawn from the density $f(x)$. Myerson (1981) defines the virtual valuations function as $J_f(x) := x - (1 - F(x))/f(x)$, and shows that if $J_f(x)$ is monotone increasing, then an auction that awards the item to the highest bidder is optimal in the sense of maximizing expected revenue. It is common in mechanism design to make the stronger assumption that the hazard function, defined by $f(x)/(1 - F(x))$, is

increasing or the even stronger assumption that $f(x)$ is $\log-$concave. McAfee and McMillan (1987) show that monotonicity of $x-(1-F(x))/f(x)$ is equivalent to convexity of the function $g(x)=1/(1-F(x))$, which is used in the the next section to formulate a density estimate subject to the constraint that $f(x)$ satisfies Myerson's (1981) regularity condition.[6]

In addition, Myerson and Satterthwaite (1983) show that bilateral bargaining between a buyer and a seller will only result in trade when the virtual valuation of the buyer and the virtual cost of the seller, defined by $x+F(x)/f(x)$, are both increasing functions. Note that we can define $H(x):=1-F(x)$ and $h(x):=H'(x)=-f(x)$ to show that this last condition is equivalent to monotonicity of $x-(1-H(x))/h(x)$. A reformulation of McAfee and McMillan's (1987) condition for monotonicity of $J_f(x)$ shows that this is equivalent to convexity of $g(x):=1/(1-H(x))$. This allows for the formulation of this shape constraint in an analogous manner as the method used to formulate the shape constraint in the next section.

It would also be interesting to explore constraining a density to have an increasing hazard function. Wellner and Laha (2017) show that this is equivalent to constraining $g(x)=-\log(1-F(x))$ to be convex. In all three of the examples given above, guaranteeing that C is positive definite requires the density estimate to satisfy a set of inequalities that do not appear to have an obvious interpretation. Regardless, statement (2) in Theorem 5 implies that $W_\gamma(\mu,\alpha(g))$ is still convex as long as $f(g)$ is sufficiently close to f_{unc}. Initializing the density near this unconstrained minimizer and then checking for convexity in each iteration often results in local convexity of $W_\gamma(\mu,\alpha(g))$ along the path of convergence.

Since the eigenvalues of the positive definite matrix ABA^T are increasing in γ, $\sqrt{\gamma/2+\sigma^2}$ can also be increased to ensure that $W_\gamma(\mu,\alpha(g))$ is convex over a larger set. This has the added benefit of increasing the dispersion of f_{unc}, which results in f_{unc} moving closer to the feasible set in the case of most shape constraints, including all the examples discussed so far. In some cases ensuring convexity by increasing $\sqrt{\gamma/2+\sigma^2}$ may result in the density being too disperse. If this occurs, it would be best to compare the resulting density estimate with an estimate subject to a stronger constraint, that allows for the formulation of a convex optimization problem, and check which density estimate fits the data more closely. For example, Ewerhart (2013) shows that a sufficient condition for a density to satisfy Myerson's (1981) regularity condition is $\rho-$concavity for $\rho>-1/2$, and a $\log-$concavity constraint can be used to ensure that the hazard function is monotonic.

Many other examples of constraints that are commonly imposed on densities in economics are given by Ewerhart (2013). Even though the examples cited in this paper all constrain

[6]This example also demonstrates that $\alpha(\cdot)$ and $\beta(\cdot)$ do not need to be unique, since we could constrain the discretized counterparts of $1/(1-F(x))$ to be convex, $\nabla_x 1/(1-F(x))$ to be monotonic, or $\nabla_{x,x}1/(1-F(x))$ to be positive.

$g(x)$ to be concave or convex, this is by no means a requirement of Theorem 7. For example, one could define a density estimate of the form given by (29)-(30) to estimate densities subject to any of the examples of shape constraints that are given by Ewerhart (2013).

1.6 Myerson's (1981) Regularity Condition

The California Department of Transportation (Caltrans) uses first price auctions to allocate construction contracts. In this section we use data on the bids submitted to Caltrans in 1999 and 2000 to explore whether or not this choice of auction format minimizes the costs to the state of California. As discussed in the previous section, if $f(x)$ is the density of private valuations for the bidders, with a distribution function denoted by $F(x)$, and if the bidders are risk neutral, Myerson (1981) shows that auctions that award the item to the person with the highest bid are optimal when the virtual valuations of the bidders is monotonically increasing.

To examine whether this condition is plausible, we need to estimate the valuations (or, in this case, marginal costs) of the construction firms. Guerre, Perrigne, and Vuong (2000) used the fact that the best response function of bidders in a first price sealed bid auction is an increasing function of the bidders's valuations to show that the valuation of bidder i can be estimated by

$$b_i + \frac{\hat{F}_b(b_i)}{(l-1)\hat{f}_b(b_i)}, \tag{1.31}$$

where l is the numbers of bidders participating in the auction, b_i is i's bid, $\hat{f}_b(\cdot)$ is a consistent estimate of the density of bids, and $\hat{F}_b(\cdot)$ is its corresponding distribution function. To control for the size of each project, we normalize each bid by Caltrans's engineers's estimates of the cost of each project before estimating $\hat{f}_b(\cdot)$ and $\hat{F}_b(\cdot)$ for each auction size.

Bajari, Houghton and Tadelis (2006) use the same dataset to estimate the costs of each firm. We follow a similar estimation strategy but make some modifications because our focus is on the costs to the state of California. Specifically, we did not subtract transportation costs from the cost estimates or treat bids from small firms differently than larger firms. Each bid consists of a unit cost bid on each item that the contract requires, and the total bid is equal to the dot product of the number of items required and the unit bid of each item. If small modifications are made to the contract after it is awarded, the final payment to the firm is equal to the dot product of the modified quantities and the unit costs in the original bid. Bajari, Houghton and Tadelis (2006) found evidence that firms are able to make accurate forecasts of these final quantities, so we follow their recommendation and

replace the first term in (26) with the final amount that is paid to the firm (normalized by the Caltran's engineers' estimate of the project cost). We also exclude all auctions in which these modifications resulted in a change in the payment received by the firm by more than 3%. After excluding these auctions we were left with 1,393 bids. Lastly, Hickman and Hubbard (2015) showed that the accuracy of the valuations estimates can be improved by applying a boundary correction to $\hat{f}_b(\cdot)$, which we also employed in our estimation procedure.

After we estimated the valuations for each firm, we estimated f_{unc} by setting $\sqrt{\gamma/2+\sigma^2}$ using Scott's (1992) rule of thumb; however, the resulting virtual valuations function was not monotonic. This could be an innocuous idiosyncrasy of the data or it could be evidence that Caltran's choice of auction format is suboptimal.

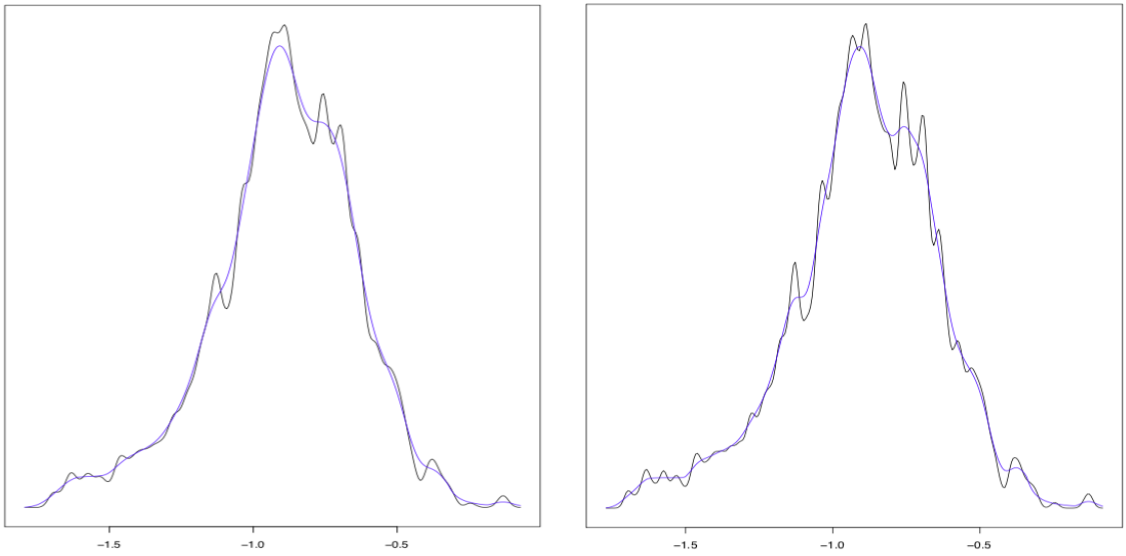
To investigate which possibility is more plausible, we find the proposed density estimate subject to Myerson's (1981) regularity condition. To define $\alpha(\cdot)$ we solved for $F(x)$ in the equation introduced in the previous section, $g(x) = 1/(1-F(x))$. This derivative is $f(x) = mg'(x)/g(x)^2$, and after discretizing, we defined $\alpha_j(g)$ as $(g_j - g_{j+1})/g_j^2$. The convexity of the objective function was maintained along the entire path of convergence, without requiring that we increase $\sqrt{\gamma/2+\sigma^2}$ above the recommendation given in the third section. The input density and the estimate are shown in Figure 3.

We also performed the test described in Theorem 5. We failed to reject the null hypothesis that the objective function, evaluated at $\hat{f}$, was equal to the objective function evaluated at its unconstrained counterpart, f_{unc}, with a $p-$value of 0.93. This is similar to the result of the Kolmogorov-Smirnov (1948) test and the Anderson-Darling (1952) test for the null hypothesis that the sample was drawn from the distribution function of $\hat{f}$; these tests also failed to reject the null with $p-$values of 0.98 and 0.94, respectively. In this case the constraints are inactive at all but 24 points in the right tail in a mesh of 300 points.[7]

Since the density already appears parsimonious, there is little motivation to decrease $\sqrt{\gamma/2+\sigma^2}$ further; however, in the interest of comparing these three tests further, we also estimated the density using Scott's (1992) rule of thumb multiplied by $1/2$ rather than $2/3$. In this case the $p-$value of our test decreased to 0.32, while the $p-$values of other two tests both increased to 0.99. This divergence in $p-$values underlines the difference between these two approaches. Specifically, as we decrease the smoothing, the distribution function converges to the empirical distribution function over the vast majority of the domain, so tests based on comparisons between a distribution and its empirical counterpart are less likely to reject. In contrast, our statistic measures the discrepancy between the global unconstrained minimum of $f \mapsto W_\gamma(\mu, f)$ and the set of feasible densities. The test is most reliable when

[7]Myerson's (1981) regularity condition can also be expressed as $f(x)^2 + f'(x)(1-F(x)) \geq 0$, so it is always satisfied when the density is increasing. In this case, f_{unc} decreases rapidly to the right of the mode, as shown in Figure 3, so it is not in the set of feasible densities.

f_{unc} is a reasonable estimate of $\mu^\star$, so we do not recommend setting $\sqrt{\gamma/2+\sigma^2}$ to a value that under-smooths f_{unc} in this way.



(a) (b)

Figure 3: The input density and the output density, subject to Myerson's (1981) regularity condition, are shown in black and blue, respectively. (a) shows the estimate when $\sqrt{\gamma/2+\sigma^2}$ is set using Scott's (1992) rule of thumb multiplied by 2/3, while in (b) a factor of 1/2 is used instead. The absolute value of the valuations can be viewed as the cost per dollar of the Caltrans's engineer's estimates.

1.7 Conclusion

This paper proposes a density estimator that is defined as the density that minimizes a regularized Wasserstein distance from the input kernel density estimator subject to $\rho-$concavity constraints. This framework provides the advantages of convexity and consistency, and it allows for a generalization that is capable of estimating densities subject to a large class of alternative shape constraints. In addition, it allows for a test of the impact of the shape constraints on the fidelity criterion.

The framework presented here can also be extended to allow γ and σ to take different values at each column of the matrix K, which would be appealing in two situations. When one would like $f^\star$ to be as close as possible to μ, γ and σ can be decreased below what would have otherwise been possible in regions where the shape-constrained density estimator is closer to μ, without interfering with convergence of the algorithm described in chapter 3. Secondly, this would allow for the development of methods that set $\sqrt{\gamma/2+\sigma^2}$ using an

adaptive approach that is similar to the one described by Sheather and Jones (1991) for kernel density estimators.

Another promising area for future research that has not already been mentioned would be to extend this framework to allow for the estimation of a regression and the density of residuals simultaneously. Dümbgen, Samworth, and Schuhmacher (2011) showed that this does not result in a convex optimization problem in the maximum likelihood setting, so verifying convexity of the objective function in this case is an active area of research. Note that extending the framework presented here to estimating the mode of a data generating process conditional on covariates, or a modal regression, is straightforward. For example, this could be done by imposing a $\rho-$concavity constraint on the conditional density of the dependent variable. Using a relatively low value of ρ, say -1 or -2, could be viewed as similar to a quasi-concavity constraint.[8] Convexity of the optimization problem in this case follows from Theorem 7.

[8]Note that quasi-concavity is equivalent to uni-modality if and only if $d = 1$.

Chapter 2

A Nonparametric Modal Regression

Abstract

We propose a nonparametric estimator of the mode of a dependent variable conditional on covariates, or a modal regression. Methods from graph theory are used to provide a consistency result and a computationally efficient algorithm. We then apply the regression to a dataset from the Office of Undergraduate Admissions at the University of Illinois at Urbana-Champaign (UIUC) that consists of UIUC GPA, high school GPA, and the ACT scores for all UIUC undergraduate students from 2000 to 2010. There are three primary advantages of the proposed estimator. First, given a dataset consisting of n observations, the proposed algorithm has a worst-case time complexity of $O(n^2)$. We also provide a generalization of this algorithm that provides an approximation of the regression at a strictly faster rate, which depends on the desired level of accuracy of the approximation. This approximation converges to the proposed estimator with probability one as n diverges, and, since it simply adds a constraint on the distance between subsequent points in the paths, it is itself a reasonable estimate of the mode. Second, the accuracy of the method compares favorably to the classic methods when the modal regression and the value of the joint density along the regression are both nonlinear. Third, when the conditional mode can be expressed as a function of the covariates, the estimator does not require the specification of smoothing parameters.

2.1 Introduction

There are three classic estimators of the univariate mode that have been proposed in the statistics literature. For example, Parzen (1962) showed that the modes of kernel density estimators provide consistent estimates of the modes. In a subsequent paper, Chernoff (1964) derived a modal estimator based on the midpoint of an interval of a fixed width containing the

most observations, and later Venter (1967) proposed estimating the mode with the midpoint of the smallest possible interval containing k data points.

The first two of these univariate estimators have been generalized to higher dimensional contexts, resulting in either parametric or nonparametric regressions that converge to the mode of the dependent variable conditional on the covariates. Like their univariate counterparts, these classic modal regressions require the choice of bandwidth parameters and are solutions to nonconvex optimization problems.

There are also three primary advantages to both of these approaches. First, the modal regression augments information provided by the regression to the mean and all of the quantiles, and there are cases in which the slope coefficient of a modal regression has an opposite sign as a regression to the mean, as well as all of the quantiles (Baldauf and Silva, 2012; Kemp and Silva, 2012).

Second, the conditional mode is a very simple notion of central tendency when the dependent variable is typically equal to a default value conditional on the covariates. For example, Cardoso and Portugal (2005) use a modal regression to describe the typical wage in a market with collective bargaining. More generally, whenever the decisions of agents can be modeled with a zero-one loss function, possibly due to either a high cost of deviating or a cost of reevaluating this default choice, a modal regression may more aptly describe this behavior than a regression to the mean.

Third, the inherently local nature of the mode makes it one of the more robust notions of central tendency. For example, modal estimators are robust to outliers moving further away from the mode, switching from one side of the mode to the other, or being removed from the dataset entirely. Moreover, with the exception of the estimators described by Lee (1989; 1993), all of the classic modal regressions have a breakdown point that converges to one asymptotically.

Lee (1989; 1993) introduced modal regressions into the economics literature with a parametric regression that can be viewed as a multivariate generalization of Chernoff's (1964) modal estimator. Subsequently, parametric estimators were proposed that relax Lee's (1989; 1993) assumption that the distribution of the error terms have a density that are either symmetric or homoscedastic; see for example, (Kemp and Silva, 2012; Ho, Damien, and Walker, 2017). Relative to other methods of estimating the conditional mode, one of the primary advantages of these approaches is statistical efficiency. In particular, the estimator proposed by Lee (1993) is $\sqrt{n}-$consistent.

Meanwhile, progress in the statistics literature has followed a different path. Scott (1992) formulated a multivariate nonparametric generalization of Parzen's (1962) modal estimator. These approaches are fully nonparametric, often even avoiding the assumption that the

conditional modes can be described by functions; see for example, (Chen et al., 2016; Yao, et al., 2012; Einbeck and Tutz, 2006). This generality comes at a cost. First, since estimation relies on various algorithms that can all be formulated as gradient ascent, with a time complexity of $O(n^2)$ in each iteration, the computational requirements of accurate estimates can be high in large datasets (Chen, 2018). Also, the specification of bandwidths is much less straightforward in this setting than in the kernel density estimation setting (Zhou and Huang, 2018; Chen, 2018).[1]

This paper proposes an alternative formulation of modal nonparametric regressions that are defined by combining several paths through a geometric graph with vertex locations given by the data. It is similar in spirit to the univariate modal estimator derived by Venter (1967), in the sense that these paths minimize the distance between consecutive datapoints. Aggregating several of these paths provides three primary advantages to the method. First, this approach allows us to leverage the large computational literature on shortest paths to provide a convergence result and an efficient estimation method. Given a sample of n observations, the proposed algorithm has a worst-case time complexity of $O(n^2)$, and, particularly in large samples, its computational efficiency compares favorably to the other nonparametric approaches. Also, a slight generalization of the algorithm provides an approximation of the regression with a worst-case time complexity of $O(n^{1+2\beta})$, where $\beta \in (0, 1/2]$. This approximation is equal to the proposed regression when $\beta = 1/2$ or, with probability approaching one, as n diverges.

Second, like methods based on kernel density estimators, we only assume that the conditional mode of the dependent variable is a manifold rather than a function. This allows the estimator to reveal traits of the data-generating process that would otherwise be difficult to ascertain, including regions of the covariates that contain multiple modes. An example of a modal manifold for which the dependent variable cannot be expressed as a function of the independent variable is shown in Figure 4.

Third, while the estimator allows for the specification of a smoothing parameter to increase the statistical efficiency of the individual paths, this is not a requirement. Our recommended approach is to forego the specification of this parameter when the modal manifold can be expressed as a function of the covariates. In these cases, the parsimony of the estimator stems from combining several, likely under-smoothed, paths. Moreover, the method also does not require any other choice parameters including bandwidth parameters or tuning parameters for optimization heuristics.

After introducing notation in Section 2, Section 3 introduces random geometric graphs. Next, Section 4 defines the paths that form the proposed estimator and provides a conver-

[1]This is described in more detail in Section 6.

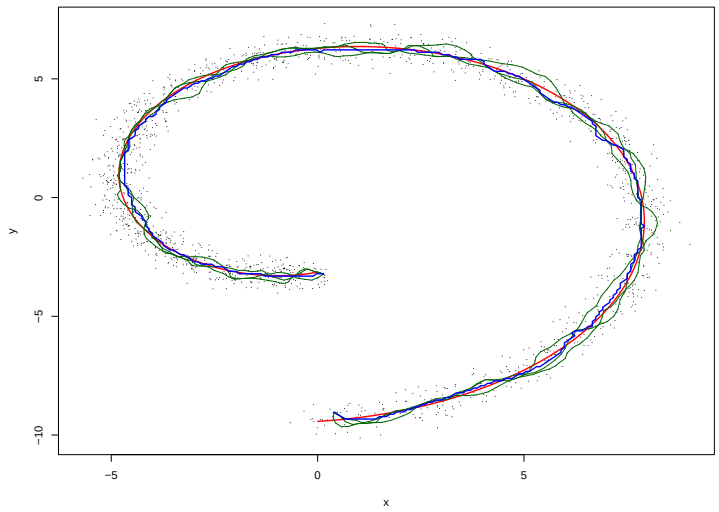


Figure 4: The blue curve shows an example of the proposed estimator of the conditional mode, the red curve is the true conditional mode, and the green curves are the three paths through the data that define the estimator.

gence result. The estimator is defined as a Frechét mean of these paths, which is discussed, along with other computational aspects of the estimator, in Section 5. Section 6 provides simulations to compare the performance of the estimator to methods based on Parzen's (1962) approach. The Section 7 applies the estimator to data from the Office of Undergraduate Admissions at the University of Illinois at Urbana-Champaign, and the final section concludes.

2.2 Notation

The input of the proposed estimator is a dataset composed of n i.i.d. observations, drawn from a distribution with a continuously differentiable joint density function $f : X \times Y \to \mathbb{R}_+$, where $X \subset \mathbb{R}^{d-1}$ is the set of independent variables and $Y \subset \mathbb{R}^1$ is the set of dependent variables. We will denote the density of y conditional x as $f(y \mid x)$, and the dataset by $\{(x_i, y_i)\}_{i=1}^n$. When there is little risk of confusion, we will refer to the set $X \times Y$ as Z and the elements $(x_i, y_i) \in X \times Y$ as z_i. We will also assume that the local modes of $f(y \mid x)$ form a manifold and parameterize this manifold as $((x^\star(t), y^\star(t)) =)\ z^\star(t) : [0, 1] \to Z$. In other words, $(\tilde{x}, \tilde{y}) \in \{z^\star(t) \mid t \in [0, 1]\}$ if and only if $\tilde{y}$ is a local maximum of $f(y \mid \tilde{x})$, which, given our assumption that $f(x, y)$ is differentiable, is equivalent to,

$$\left.\frac{\partial f(y|x)}{\partial y}\right|_{(x,y)=(\tilde{x},\tilde{y})} = 0 \text{ and } \left.\frac{\partial^2 f(y|x)}{\partial^2 y}\right|_{(x,y)=(\tilde{x},\tilde{y})} < 0.$$

As mentioned in the introduction, the proposed estimator is defined by paths through the datapoints. The fifth section provides a way to avoid specifying the endpoints of these paths, but to simplify the exposition, we will assume these endpoints are fixed in the following two sections. Specifically, suppose these endpoints are defined as the datapoint that is closest to the estimate of the univariate mode of $\{y_i \mid |x_i - x^\star(0)| \leq \epsilon_n\}$ (respectively, $\{y_i \mid |x_i - x^\star(1)| \leq \epsilon_n\}$), where $\epsilon_n \to 0$ and $n\epsilon_n \to \infty$. We will denote these estimates as $z_{\underline{i}}$ ($z_{\bar{i}}$).

Some additional definitions will also be useful. The floor of $x \in \mathbb{R}$ will be denoted by $\lfloor x \rfloor$, the Dirac delta function centered at z by $\delta_z(\cdot)$, the Euclidean distance between z_i and z_j by $\|z_i - z_j\|$, the ball in $\mathbb{R}^d$ centered at z with radius r by $B_r(z)$, and an n dimensional vector with each element equal to the constant c will be denoted by $\mathbf{c}_n$. In addition, $|\cdot|$ will be used to denote either the absolute value or the L_0-"norm" when there is little risk of confusion. The following section provides a brief introduction to geometric graphs.

2.3 Geometric Graphs

A graph consists of vertices and the links between these vertices, or edges. Also, a complete graph is a graph that contains an edge between every pair of vertices. One can also generalize this definition of graph to add attributes to either the edges or the vertices. For example, a weighted graph also assigns a weight, generally a scalar, to each edge in the graph.

Similarly, geometric graphs assign a location to each vertex in the graph. For example, one can construct a $K-$nearest neighbors graph by assigning a location to each vertex, and then specifying that vertex i is adjacent to vertex j if and only if either point has a location that is among the closest K points of the other's location. The limiting behavior of geometric graphs with random variables defining the location of vertices was first studied by Beardwood, Halton, and Hammersley (1959), and subsequently, a large literature has grown around this topic.

We will define the geometric graph G as a graph with vertices given by $V := \{1, 2, ..., n\}$. The location of the i^{th} vertex will be defined by z_i. Paths through the graph will be represented as partial permutation of V, and we will use Ω to denote the set of all partial permutations of V such that for every $\omega \in \Omega$, $\omega(1) = \underline{i}$ and $\omega(|\omega|) = \bar{i}$. Howard and Newman (1997, 2001) define the power weighted path length as

$$L(z, \tilde{\omega}) := \sum_{i=1}^{|\tilde{\omega}|-1} \left\| z_{\tilde{\omega}(i)} - z_{\tilde{\omega}(i+1)} \right\|^\rho,$$

where $\rho > 1$, and study the limiting behavior of the minimum power weighted path, defined as,

$$\omega := \arg\min_{\tilde{\omega}\in\Omega} L(z, \tilde{\omega}). \tag{2.1}$$

Note that if we set ρ equal to one the solution to this optimization problem would simply be $\omega = \{\underline{i}, \bar{i}\}$, so we will focus on more interesting values of this parameter. Increasing ρ above this level can be viewed as placing a higher penalty on having large jumps between consecutive datapoints, or equivalently, as placing a lower penalty on having a longer Euclidean length.

We will make use of the following implication of a result provided by Hwang, Damelin, and Hero (2016) in the next section. The result provides the limiting value of the datapoints in $\{z_{\omega(i)}\}_i$. The graph of this path is given by $\{z \mid \exists\ i \in \omega \text{ and } \alpha \in [0,1] \text{ s.t. } z = \alpha z_{\omega(i)} + (1-\alpha) z_{\omega(i+1)}\}$, and it will be convenient to refer to this path by the function $(\hat{x}(t), \hat{y}(t)) : [0,1] \to \mathbb{R}^d$, or simply $\hat{z}(t)$.

Lemma 1 (Hwang, Damelin, and Hero, 2016): *Suppose $\{z_i\}_{i=1}^n$ is a dataset composed of i.i.d. observations of a random variable in Z with a continuous density function given by $f : Z \to \mathbb{R}$, and $f(z) > 0$ for all $z \in Z$. Then the minimum power weighted path through the complete graph G from $z_{\underline{i}} \in Z$ to $z_{\bar{i}} \in Z$, where $z_{\underline{i}} \neq z_{\bar{i}}$ converges in probability to*

$$\hat{z}(t) := \arg\min_{z(\cdot)\in\Gamma} \int_0^1 f(z(t))^{(1-\rho)/d} \left\| \nabla z(t) \right\| dt,$$

where Γ is the set of all piecewise differentiable curves $z : [0,1] \to Z$ such that $z(0) = z_{\underline{i}}$ and $z(1) = z_{\bar{i}}$.

Proof: See Theorem 1 in Hwang, Damelin, and Hero (2016).

□

Remark 5: Hwang, Damelin and Hero's (2016) result is considerably stronger than the theorem provided. This requires the additional assumptions that $\inf_z\ f(z) > 0$ and Z is compact. These are required because their result concerns the optimal paths between every pair of points in Z, say $w, v \in Z$, through a graph that is defined with vertices with locations $\{z_i\} \cup u \cup v$. To see why this is not necessary in our setting, let $\epsilon \in (0, \min_t f(z^\star(t)))$, $\tilde{Z}^{(\epsilon)} := \{z \mid f(z) \geq \epsilon\}$, and $\tilde{\Gamma}$ be the set of all piecewise differentiable curves $z : [0,1] \to \tilde{Z}$ such that $z(0) = z^\star(0)$ and $z(1) = z^\star(1)$. Note that this set is nonempty since $z^\star(\cdot) \in \tilde{\Gamma}$.[2] Also, using two results from the calculus of variations, along with the fact that $f(\cdot)$ is a

[2]Since $z_{\underline{i}}$ and $z_{\bar{i}}$ are consistent and $f(\cdot)$ is differentiable, $z_{\underline{i}}, z_{\bar{i}} \in \tilde{Z}^{(\epsilon)}$ with probability approaching one, so the probability that Γ is nonempty also approaches one for any such ϵ. We use $z^\star(0)$ and $z^\star(1)$ to avoid complications regarding the event that $f(z_{\underline{i}}), f(z_{\bar{i}}) = 0$, but the logic in the remark also applies to $z_{\underline{i}}$ and $z_{\bar{i}}$, with probability approaching one.

differentiable and proper density function, one can show the existence and differentiability of the minimizer,

$$z^{(\epsilon)}(t) := \underset{z(\cdot)\in\tilde{\Gamma}}{\arg\min} \int_0^1 f(z(t))^{(1-\rho)/d} \left\|\nabla z(t)\right\| dt;$$

see for example, Theorems 2 and 4 in chapter 8 of (Evans, 2010). Note that if $\lim_{\epsilon\to 0^+} f(z^{(\epsilon)}(t)) = 0$, we would have $\lim_{\epsilon\to 0^+} \int_0^1 f(z^{(\epsilon)}(t))^{(1-\rho)/d} \left\|\nabla z^{(\epsilon)}(t)\right\| dt = \infty$ since $z^{(\epsilon)}(t)$ is differentiable and $\rho > 1$. Thus, there exists $\epsilon > 0$ such that $\hat{z}(t)$ is in the interior of $\tilde{Z}^{(\epsilon)}$. Suppose $\epsilon > 0$ is fixed at such a value, and note that $\tilde{Z}^{(\epsilon)}$ satisfies the two additional assumptions of Hwang, Damelin and Hero's (2016) theorem. Since the minimizer is in the interior of $\tilde{Z}^{(\epsilon)}$ for every $t \in [0,1]$, we have $\hat{z}(\cdot) = z^{(\epsilon)}(\cdot)$.

□

Since this lemma is of central importance for our main result, we will briefly summarize the intuition of the proof when $Z = [0,1]^2$. To do this we will begin by considering a uniform density on $[0,1]^2$.

Suppose we approximate the number of points within a circle of radius $r > 0$, contained within $[0,1]^2$, by $\pi r^2 n$. After setting this equal to $k \in \mathbb{N}_{++}$ and solving for r, we have that the approximate distance between a given point and its k^{th} nearest neighbor is $\sqrt{k/(n\pi)} \in O(n^{-1/2})$. Thus, if the optimal path only passes from each point to one of the point's k nearest neighbors, the total number of points in the path would be approximately $\left\|z_{\underline{i}} - z_{\bar{i}}\right\| / \sqrt{k/(n\pi)} = cn^{1/2}$. Since moving a distance of $O(n^{-1/2})$ increases the objective function by $O(n^{-\rho/2})$, we can approximate the power weighted path length by multiplying the number of edges ($\approx cn^{1/2}$) by the path length ($\approx n^{-\rho/2}$), which is $cn^{(1-\rho)/2}$ in this case.

Now let's extend these results to the general density function $f(z)$, where $z \in Z \subseteq \mathbb{R}^2$ and Z is bounded. Suppose we partition $Z \subseteq \mathbb{R}^2$ into m^2 squares, denoted by $\{Q_i\}_{i=1}^{m^2}$, with centers denoted by $\{q_i\}_{i=1}^{m^2}$. We will approximate the density function $f(z)$ by $f(q_i)$ when $z \in Q_i$, and after rescaling the datapoints in Q_i to span the unit square, we can use the results from the uniform density case to approximate the increase in the objective function from passing through Q_i.

Since the number of data points in Q_i is approximately $m^{-2} f(q_i) n$, our analysis on the uniform density on $[0,1]^2$ implies that passing through Q_i contributes approximately $c\left(m^{-2} f(q_i) n\right)^{(1-\rho)/2} m^{-\rho}$ to the power weighted distance. The $m^{-\rho}$ term is due to rescaling and the fact that $L(\cdot)$ is homogeneous of degree ρ in $\{z_i\}_i$.

Now we can approximate the objective function of a path between the endpoints through $O(m)$ of the $\{Q_i\}_i$ partitions. We will denote the indices of the partitions that this path passes through by ψ. After simplifying and summing over all ψ, the approximate power

weighted distance is given by $n^{(1-\rho)/2}\sum_{i=1}^{|\psi|} c_i f(q_{\psi_i})^{(1-\rho)/2}/m$. After taking the limit as m and n diverge to infinity, we can see that the minimizer of this function converges to the minimizer of $\int_0^1 f(x(t),y(t))^{(1-\rho)/2}\sqrt{x'(t)^2+y'(t)^2}dt$, which is the objective function of the optimization problem given in Lemma 1 when $d=2$.

The estimator discussed in the next section can be viewed as a minimum power weighted path through the lifted graph of G, which we will denote by $\mathcal{G}$ (Cowlagi and Tsiotras, 2009). Before we construct $\mathcal{G}$, we will add two datapoints, denoted by z_0 and z_{n+1}, to the dataset, which will be defined to be equal to $z_{\underline{i}}$ and $z_{\bar{i}}$, respectively.

The vertices in $\mathcal{G}$ will be denoted by $\mathcal{V} := \{(i,j)\}_{i,j=1,j\neq i}^n \cup \{(0,\underline{i}),(\bar{i},n+1)\}$ so that each of the $n(n-1)$ vertices in $\{(i,j)\}_{i,j=1,j\neq i}^n$ corresponds to an edge in G. The vertices in $\{(0,\underline{i}),(\bar{i},n+1)\}$ will be viewed as the endpoints of the path. We will also add edges between the vertices (i,j) and (k,l) if and only if $j=k$ and $i\neq l$. The location of vertex (i,j) will be defined as $z_{i,j} := (x_i, x_j, c_{ij})$, where $c_{i,j} := y_i - y_j$. The next section uses this graph to define the estimator, and provides two results related to its limiting behavior.

2.4 The Nonparametric Modal Regression

In the context of our lifted graph, $\mathcal{G}$, we have

$$L(z,\tilde{\omega}) = \sum_{i=2}^{|\tilde{\omega}|-1} \left\| z_{\tilde{\omega}(i-1),\tilde{\omega}(i)} - z_{\tilde{\omega}(i),\tilde{\omega}(i+1)} \right\|^{\rho}.$$

The indices of the datapoints that define the estimator can be found by solving

$$\omega := \arg\min_{\tilde{\omega}\in\Omega} \sum_{i=2}^{|\tilde{\omega}|-1} \left\| z_{\tilde{\omega}(i-1),\tilde{\omega}(i)} - z_{\tilde{\omega}(i),\tilde{\omega}(i+1)} \right\|^{\rho}, \tag{2.2}$$

where Ω is the set of partial permutations of $\mathcal{V}$ such that for all $\tilde{\omega}\in\Omega$, $\tilde{\omega}(1) = (0,\underline{i})$ and $\tilde{\omega}(|\tilde{\omega}|) = (\bar{i},n+1)$. Note that when there is little risk of confusion, Ω and ω will also be used to denote the set of feasible paths in the unlifted graph G and the optimal path through G, respectively. In practice the proposed estimator appears to perform well when $\rho=2$, so we will use this value throughout the rest of the paper. After substituting this value for ρ and recalling $z_{i,j} := (x_i, x_j, y_i - y_j)$, we can show that (2) can also be written as,

$$\omega = \arg\min_{\tilde{\omega}\in\Omega} \sum_{i=2}^{|\tilde{\omega}|-1} 2\left\| x_{\tilde{\omega}(i)} - x_{\tilde{\omega}(i+1)} \right\|^2 + \left(y_{\tilde{\omega}(i-1)} - 2y_{\tilde{\omega}(i)} + y_{\tilde{\omega}(i+1)} \right)^2. \tag{2.3}$$

The first term in the objective function penalizes the distance between consecutive values of the covariates along the path $\omega\in\Omega$, and the last term places a penalty on paths that are less parsimonious.

Next we will provide a lemma that will be useful in the proof of our main convergence result. The result shows that asymptotically the distance between consecutive datapoints in $\{z_{\omega(i)}\}_i$, with locations in G, converges to zero with probability approaching one. Hwang, Damelin and Hero (2016) and Howard and Newman (1997) provide lemmas that show that the distance between consecutive points along the optimal path, which is analogous to the locations $\{(x_{\omega(i)}, x_{\omega(i+1)}, c_{\omega(i),\omega(i+1)})\}_i$ of the graph $\mathcal{G}$ in the context of this paper, goes to zero. In our setting, the result below can be viewed as stronger, since it implies their result and shows that the limiting value of each element of the set $\{c_{\omega(i)}\}_i$ is zero.

Lemma 2: *Suppose a dataset is composed of n i.i.d. observations of a random variable with a density function given by $f(z)$. For any $a > 0$ and $\gamma \in (0,1)$, there exists $b > 0$ such that $P\left(\left\|z_{\omega(i)} - z_{\omega(i+1)}\right\| > an^{(\gamma-1)/d}\right) < \exp(-bn^{\gamma})$ for every $i \in \{1, 2, ..., |\omega|\}$ asymptotically.*
Proof: We will assume throughout the proof that the Lebesgue measure zero event that three or more datapoints are collinear does not occur, which is sufficient to guarantee $L(z_\omega) > 0$.

Step 1: First we will show that $P\left(\left\|z_{\omega(i)} - z_{\omega(i+1)}\right\| > an^{(\alpha-1)/d}\right)$ decays exponentially for some non-empty subset of the datapoints. To do this we will compare the path, say ω, with an alternative, say $\breve{\omega}$, that includes all the points in ω, as well as additional points that will be added between every consecutive pair of points, other than the first and last pairs. In other words, if we denote the index of the point that we add between $\omega(i)$ and $\omega(i+1)$ as $\breve{\omega}(i+1/2)$, we can write $\breve{\omega}$ as

$$\{\omega(1), \omega(2), \breve{\omega}(3/2), \omega(3), ..., \omega(|\omega|-2), \breve{\omega}(|\omega|-3/2), \omega(|\omega|-1), \omega(|\omega|)\}.$$

We will also require that $z_{\breve{\omega}(i+1/2)} \in \Theta_i := B_\alpha\left((z_{\omega(i)} + z_{\omega(i+1)})/2\right)$, where $\alpha > 0$.

Let $h(z_\omega, z_{\breve{\omega}}) := L(z_\omega) - L(z_{\breve{\omega}})$. Note that when $h(z_\omega, z_{\breve{\omega}}) \geq 0$, the path $\breve{\omega}$ is shorter, in a power weighted sense, than the path ω. Let $g(z_\omega, \alpha) :=$arg $\min_{\{z_{\breve{\omega}(i+1/2)} \in \ \Theta_i\}} h(\cdot)$. We can find the limit of $g(\cdot)$ as $\alpha \to 0$ by simply evaluating $h(\cdot)$ at $z_{\breve{\omega}(i+1/2)} = (z_{\omega(i)} + z_{\omega(i+1)})/2$ for all $i \in \{2, ..., |\omega| - 2\}$. This limit is given by,

$$3/2(y_3 - y_2)^2 + (y_1 - (y_2 + y_3)/2)^2 - (y_3 - y_1)^2 + 3/4\left(\sum_{i=3}^{k-2}(y_{i+1} - 2y_i + y_{i-1})^2\right) +$$

$$\left(\sum_{i=2}^{k-2}(x_i - x_{i+1})^2\right) + ((y_{k-2} + y_{k-1})/2 - y_k)^2 + 3/2(y_{k-1} - y_{k-2})^2 - (y_k - y_{k-2})^2,$$

where $k = |\omega|$. Since $z_{\omega(1)} = z_{\omega(2)}$ and $z_{\omega(k)} = z_{\omega(k-1)}$ by construction, $g(z_\omega, 0)$ can be written as,

$$3/4\left((y_{\omega(3)} - y_{\omega(2)})^2 + (y_{\omega(k-1)} - y_{\omega(k-2)})^2 + \sum_{i=3}^{k-2}(y_{\omega(i+1)} - 2y_{\omega(i)} + y_{\omega(i-1)})^2\right)$$

$$+ \sum_{i=2}^{k-2}(x_{\omega(i)} - x_{\omega(i+1)})^2 > 0.$$

Since $g(z_\omega, \alpha)$ is continuous in α, by the intermediate value theorem we can increase α above zero such that $g(\cdot) > 0$ still holds. Suppose α is fixed at such a value.

Our definition of Θ_i implies that there exists $\beta > 0$ such that the volume of $\Theta_i \cap Z$ is at least $\beta \left\| z_{\omega(i)} - z_{\omega(i+1)} \right\|^d$. Suppose that $\left\| z_{\omega(i)} - z_{\omega(i+1)} \right\| > an^{(\gamma-1)/d}$ for every $i \in \{1, 2, ..., |\omega| - 1\}$. This implies that the volume of Θ_i is bounded below by $a^d \beta n^{(\gamma-1)}$. Let $f_i := \min_{z \in \Theta_i} f(z)$. For $k \in \{1, 2, ..., n\}$ this implies

$$P(z_k \notin \Theta_i) < 1 - a^d f_i \beta n^\gamma / n,$$

so

$$P(z_k \notin \Theta_i \; \forall k \in \{1, 2, ..., n\}) < (1 - a^d f_i \beta n^\gamma / n)^{n-2}.$$

Since the number of points in the path is bounded above by n,

$$P(\left\| z_{\omega(i)} - z_{\omega(i+1)} \right\| > an^{(\gamma-1)/d} \; \forall \; i \in \{2, ..., |\omega| - 2\})$$

$$< n(1 - a^d f_m \beta n^\gamma / n)^{n-2}$$

where $f_m = \min_i f_i$. This converges to $\exp(-bn^\gamma)$, where $b := a^d f_m \beta$. This establishes the desired property holds for at least one pair of consecutive points in $\{2, ..., |\omega| - 2\}$.

Step 2: Now let Γ be defined so that $\omega(i) \in \Gamma$ if and only if $z_{\omega(i)} - z_{\omega(i+1)}$ satisfies $P(\left\| z_{\omega(i)} - z_{\omega(i+1)} \right\| > an^{(\gamma-1)/d}) < \exp(-bn^\gamma)$ asymptotically. Also, redefine $\breve{\omega}$ so that $\breve{\omega}$ is a subpath of ω, $\breve{\omega}(i) \notin \Gamma$ for all $i \in \{2, ..., |\breve{\omega}| - 1\}$, and $\breve{\omega}(1), \breve{\omega}(|\breve{\omega}| - 1) \in \Gamma$. Sufficient conditions to guarantee that the first and last three terms of $g(z_\omega, 0)$ are positive, are $\left| y_{\breve{\omega}(1)} - y_{\breve{\omega}(2)} \right| < 3/4 \left| y_{\breve{\omega}(3)} - y_{\breve{\omega}(2)} \right|$ and $\left| y_{\breve{\omega}(k)} - y_{\breve{\omega}(k-1)} \right| < 3/4 \left| y_{\breve{\omega}(k-1)} - y_{\breve{\omega}(k-2)} \right|$. Suppose that the sample size is sufficiently large so that this property holds and also that $|\Gamma| < |\omega| - 2$. The same argument that was used above can be applied again using our new definition of $\breve{\omega}$ to show that there exists $b > 0$ such that the probability of this event converges to $\exp(-bn^\gamma)$.

Note that in the case in which only four points are in $\breve{\omega}$, $h(\cdot)$ is defined as,

$$(y_{\breve{\omega}(1)} - 2y_{\breve{\omega}(2)} + y_{\breve{\omega}(3)})^2 + (y_{\breve{\omega}(2)} - 2y_{\breve{\omega}(3)} + y_{\breve{\omega}(4)})^2 + \left\| x_{\breve{\omega}(2)} - x_{\breve{\omega}(3)} \right\|^2 -$$

$$(y_{\breve{\omega}(1)} - 2y_{\breve{\omega}(2)} + y_{\breve{\omega}(3/2)})^2 - (y_{\breve{\omega}(2)} - 2y_{\breve{\omega}(3/2)} + y_{\breve{\omega}(3)})^2 - (y_{\breve{\omega}(3/2)} - 2y_{\breve{\omega}(3)} + y_{\breve{\omega}(4)})^2$$

$$- \left\| x_{\breve{\omega}(2)} - x_{\breve{\omega}(3/2)} \right\|^2 - \left\| x_{\breve{\omega}(3/2)} - x_{\breve{\omega}(3)} \right\|^2.$$

Our definition of Θ_2 implies that this region is an ellipse. A sufficient condition to guarantee that $\left(z_{\breve{\omega}(2)} + z_{\breve{\omega}(3)} \right)/2$ is in this ellipse, as well as for the ellipse to be non-imaginary, is to have $\left| y_{\breve{\omega}(1)} - y_{\breve{\omega}(2)} \right|$ and $\left| y_{\breve{\omega}(3)} - y_{\breve{\omega}(4)} \right|$ be larger than $\left| y_{\breve{\omega}(2)} - y_{\breve{\omega}(1)} \right|$, so the argument above applies to this case as well since $\omega(2) \notin \Gamma$.

□

Remark 6: Note that the set Θ_i in the proof can sometimes be empty. For example, this would be the case for all Θ_i if all of the datapoints were approximately uniformly placed along a line with sufficiently high slope. Also, if $x_i = 0$ and $y_i = i$ for all $i \in \{1, ..., n\}$, we would have $L(z, \omega) = 0$. Since no points would be added to the path in this case, the elements in the set $\{c_{\omega(i),\omega(i+1)}\}_i$ would not converge to zero. Thus, if $c_{\omega(1),\omega(2)}$ and $c_{\omega(|\omega|-1),\omega(|\omega|)}$ were not equal to zero by construction, the argument above would not be able to rule out the case in which $c_{\omega(i),\omega(i+1)} \to c \neq 0$.

□

There are two assumptions in Lemma 1 that prevent us from applying it to the lifted graph $\mathcal{G}$. First, since $c_{i,j}$ is dependent on $c_{j,k}$, the location of the vertices in the lifted graph $\mathcal{G}$ are not iid. Second, since the vertex (i, j) and the vertex (k, l) are only connected if $j = k$, $\mathcal{G}$ is not a complete graph. Regardless, many of the arguments that are used in the proof of Lemma 1 can also be generalized to the graph $\mathcal{G}$, and we believe that it is possible to fully generalize their proof to the present setting, which is an area of active research. The next theorem shows that the path defined by (3) converges to the mode of $f(y \mid x)$ if this conjecture is true.

Theorem 3: *Suppose $f(z)$ is a proper density and $f(z)$ is a differentiable function. In addition, suppose that Lemma 1 holds on the lifted graph, $\mathcal{G}$. If $z_{\underline{i}}$ and $z_{\bar{i}}$ are consistent estimators of $z^\star(0)$ and $z^\star(1)$, then $\hat{z}(t) \xrightarrow{p} z^\star(t)$.*

Proof: We will denote its limiting value of $(\hat{x}(t), \hat{y}(t))$ by $(x(t), y(t))$, and the limiting value of $(\hat{x}_1(t), \hat{x}_2(t), \hat{c}(t))$ by $(x_1(t), x_2(t), c(t))$. Lemma 1 implies,

$(\hat{x}_1(t), \hat{x}_2(t), \hat{c}(t)) \xrightarrow{a.s.}$

$$\underset{(x_1(t),x_2(t),c(t))\in\Gamma}{\arg\min} \int_0^1 \frac{\sqrt{x_1'(t)^2 + x_2'(t)^2 + c'(t)^2}}{g(x_1(t), x_2(t), c(t))^{1/d}} dt, \tag{2.4}$$

where $g(x_i, x_j, c_{i,j})$ denote the density of the random variable $(x_i, x_j, c_{i,j})$ and Γ is the set of all differentiable curves such that $z(0) = z_{\underline{i}}$ and $z(1) = z_{\bar{i}}$. After finding the Euler-Lagrange equations and reparameterizing $(x_1(t), x_2(t), c(t))$ to be a function of arc length, denoted by s, so that $x_1'(s)^2 + x_2'(s)^2 + c'(s)^2 = 1$, we have,

$$dg(x_1(s), x_2(s), c(s))x_1''(s) + \frac{\partial g(x_1(s),x_2(s),c(s))}{\partial x_1(s)} = x_1'(s)\left(\frac{\partial g(x_1(s),x_2(s),c(s))}{\partial s}\right)$$

$$dg(x_1(s), x_2(s), c(s))x_2''(s) + \frac{\partial g(x_1(s),x_2(s),c(s))}{\partial x_2(s)} = x_2'(s)\left(\frac{\partial g(x_1(s),x_2(s),c(s))}{\partial s}\right)$$

$$dg(x_1(s), x_2(s), c(s))c''(s) + \frac{\partial g(x_1(s),x_2(s),c(s))}{\partial c(s)} = c'(s)\left(\frac{\partial g(x_1(s),x_2(s),c(s))}{\partial s}\right).$$

After a change of variables to find the functional form of $g(\cdot)$, we have $g(x_i, x_j, c_{ij}) := \int_{-\infty}^{\infty} f(x_i, y) f(x_j, y + c_{ij}) \partial y$. We can ensure that the Euler-Lagrange equations characterize the minimizer of (4), and uniqueness of this minimizer, using Jacobi's condition (Evans, 2010). This condition is given by,

$$\nabla_{w'(t),w'(t)}\left(\frac{\sqrt{x_1'(t)^2 + x_2'(t)^2 + c'(t)^2}}{g(x_1(t),x_2(t),c(t))^{1/d}}\right) > 0,$$

for $w'(t) \in \{x_1'(t), x_2'(t), c'(t)\}$, which is always positive by our assumption that $f(z) > 0$.

By Lemma 2, $\hat{c}(s) \to 0$, and $c(s) = 0$ implies $c'(s)$ and $c''(s)$ are also zero. These properties imply that the third Euler-Lagrange equation can be written as

$$\frac{\partial g(x_1(s), x_2(s), c(s))}{\partial c(s)} = 0. \tag{2.5}$$

After substituting in the definition of $g(\cdot)$, we have,

$$\frac{\partial \int_{-\infty}^{\infty} f(x_1(s),y(s)) f(x_2(s),y(s)+c(s)) \partial y}{\partial c(s)} = 0.$$

Dividing both sides of this equality by $\int_{-\infty}^{\infty} f(x_1(s), y(s)) \partial x_1(s)$, using the fact that $x_1(s) = x_2(s)$ and $c(s) = 0$ by Lemma 2, and Leibniz's rule implies,

$$E\left(\frac{\partial f(x(s),y(s))}{\partial y(s)} \mid x(s)\right) = 0.$$

Since, for a given value of s, $(\hat{x}(s), \hat{y}(s))$ converges, this expected value uniquely characterizes the limiting value of $\hat{y}(s)$.

□

In principal, it would be possible to derive conservative estimates of intervals containing the curve $z^\star(\cdot)$ with probability approaching one using several available bounds on $\{z_{\omega(i)}\}_i$. For example, Theorem 1 and Lemma 4 from Hwang, Damelin, and Hero (2016) provides an upper bound on $L(z, \omega)$ as well as an upper bound on $|\omega|$ respectively. These bounds, along with $z_{\omega(1)}$ and $z_{\omega(|\omega|)}$, may rule out some more extreme paths asymptotically. However, this approach does not appear to be promising because these bounds alone are not tight enough for the resulting intervals to converge; instead they are $O_p(1)$ as n diverges. There appear to be more convincing possibilities as well, and this is also an active area of research.

The literature on modal regressions places a large emphasis on prediction sets, and providing theoretical guarantees in this case is more straightforward; see for example (Chen et al., 2016; Zhou and Huang, 2018). To define this concept let $\mathcal{E}_{1-\alpha} = \{\epsilon \geq 0 \mid P(\inf_t \ \|\hat{y}(t) - \mathcal{Z}\| >$

$\epsilon) < \alpha\}$ and $\epsilon_{1-\alpha} := \inf_{\epsilon \in \mathcal{E}_{1-\alpha}} \epsilon$. Then, the unconditional $1-\alpha$ prediction set is defined as, $\mathcal{A}_{1-\alpha} := \{z \in Z \mid \|\hat{y}(t) - \mathcal{Z}\| \leq \epsilon\}$. Consistency of the coverage of this set follows from Kiefer and Wolfowitz's (1956) result that the empirical density of the errors is consistent. Prediction sets are particularly appealing in the context of nonparametric modal regressions because they can provide a large decrease in the volume of $\mathcal{A}_{1-\alpha}$ when $y^{\star}(\cdot)$ cannot be expressed as a function of x.

It is also possible to define a more general class of paths through $\mathcal{G}$ using by replacing our definition of the paths, given in (3), with,

$$\underset{\omega \in \Omega}{\arg\min} \ \sum_i \ \left\| x_{\omega(i)} - x_{\omega(i+1)} \right\|^2 + \gamma \kappa(\{z_{\omega(i)}\}_i, \omega, i),$$

where $\kappa(\cdot) \geq 0$ is a function that captures some notion of curvature and $\gamma > 0$ is a choice parameter.[3] One important advantage of defining $\kappa(\cdot)$ to be $(c_{\omega(i-1),\omega(i)} - c_{\omega(i),\omega(i+1)})^2$ is that $c_{\omega(i),\omega(i+1)}$ converges to zero. Not only does this play an important role in Theorem 3, but in practice, using definitions of $\kappa(\cdot)$ that do not have this property appear to increase the error of the estimator substantially. One alternative definition of $\kappa(\cdot)$ that does have this property is $\left\| z_{\omega(i-1)} - 2z_{\omega(i)} + z_{\omega(i+1)} \right\|^2$. This generally provides a slight gain in accuracy when estimating modal manifolds that cannot be represented by functions and a slight loss in accuracy otherwise.[4]

Multiplying $(c_{\omega(i-1),\omega(i)} - c_{\omega(i),\omega(i+1)})^2$ by a choice parameter that is greater than one does appear to provide a more parsimonious regression and a lower variance, but our preferred approach of lowering the variance is to estimate the conditional mode several times and then combining these paths to define the final estimator. The only exception to this recommendation is when the regression cannot be represented as a function of the covariates. In these cases, the distance between subsequent points is slower to converge to zero in regions in which the population modal manifold is parallel to the vertical axis, so there appears to be a larger gain in accuracy from specifying this choice parameter. This is shown in one of our simulations in the sixth section. The next section provides more detail on methods for solving (3) efficiently, and describes how to combine these paths to define the final estimator.

[3]Note that if there exists a function $g : Z \times Z \to \mathbb{R}$, so that $\kappa(\cdot) := (g(z_i, z_j) - g(z_k, z_l))^2$, generalizing Lemma 1 to the context of this paper would be straightforward; however, the resulting notion of curvature would be unconventional, since it would require the curvature of the path $i \to j \to k \to l$ be the same as the path $k \to l \to i \to j$.

[4]Note that all triangles with side lengths $a, b, c > 0$ satisfy the property $a^2 + b^2 \geq c^2/2$, with equality only in the case of the degenerate isosceles triangle that satisfies $a = b = 2c$. Since $(p - 2q + r)^2$, for $p, q, r \in \mathbb{R}^1$, can be written as $2(p-q)^2 + 2(q-r)^2 - (p-r)^2$, this definition of $\kappa(\cdot)$ has the intuitive interpretation of placing a larger penalty on consecutive triplets of $\{z_{\omega(i)}\}_i$ that do not approximately form this degenerate isosceles triangle.

2.5 Computation

Dijkstra's (1959) Algorithm

There is a large literature in computer science on calculating solutions to the shortest path problem efficiently, and Dijkstra's (1959) algorithm is an example of a particularly simple and efficient approach. This is most easily described by considering the computation of the shortest path length in an undirected graph from the source vertex, say 1, to every other vertex when the cost of moving from i to j is given by w_{ij}. Note that if all elements of $\{w_{ij}\}_{i=1,j>i}$ are unique, this is also equivalent to finding all of the optimal paths. To see this, suppose our path distances are stored in the vector s, so that s_j is the distance from j to 1. Now, starting at j, we can find the preceding point in the path by observing that k must satisfy $s_j - s_k = w_{jk}$. This is the case because whenever the shortest path from 1 to j passes through k, then the section of this path from k to j is the optimal path from k to j. This intuition is analogous to Bellman's principle of optimality, and is also the basis of Dijkstra's (1959) approach.

Given a graph with n vertices, the algorithm is initialized by defining v to be $\{1, 2, ..., n\}$ and defining the $n \times 1$ vector s so that $s_1 = 0$ and $s_{-1} = \infty$. At the beginning of each iteration we define $k := \arg\min_{i \in v} s_i$, and remove k from v. The value of s_k will not be changed in any of the subsequent iterations. With k defined, we can consider each of the elements $j \in v$, and see whether or not passing through k before going onto j is less than our current estimate of the path length from 1 to j. If passing through k in this way would decrease the path length, then we redefine s_j to be $s_k + w_{k,j}$. In other words, for each $j \in v$ we redefine s_j to be $\min(s_j, s_k + w_{k,j})$. Then the algorithm moves onto the next iteration until $v = \emptyset$. Algorithm 2 provides more details on this method.

Note that the worst-case computational cost of Dijkstra's algorithm is $O(|E|+|V|\log(|V|))$ Using a complete graph to model G, would imply that $\mathcal{G}$ would have $O(n^4)$ edges. Thus, the time complexity of this algorithm would be $O(n^4)$. However, Lemma 2 implies that we can ignore links between datapoints that are sufficiently far from one another.

Equivalently, we can define G to be a k_n-nearest neighbors graph, where $k_n := \lfloor n^\beta \rfloor$ for $\beta \in (0, 1]$. In practice setting β to approximately $1/2$ rarely has any impact on the resulting paths when the dataset contains at least a few hundred points. This also results in $\mathcal{G}$ having $O(nk_n) = O(n^{3/2})$ vertices and $O(nk_n^2) = O(n^2)$ edges, so the worst-case time complexity of Dijkstra's algorithm is $O(n^2)$. Since the time complexity of constructing a k_n-nearest neighbors graph is equal to the time complexity of finding the distance between each pair of points, which is $O(n^2)$, lowering β will not improve the time complexity of the estimator.

However, in large sample sizes lowering β does have a non-negligible impact on the wall

time of the estimator, and using lower values, say $\beta = 2/5$, generally produces paths that are similar or identical to the paths when $\beta = 1$ as long as $n \geq 250$ points, and, in light of Lemma 2, the same can be said for lower values as n increases beyond this point. When these lower values have a small impact on the estimator, it is often not obvious that this is detrimental to the estimator's accuracy; however, an increase in the error is apparent when β is set to a value that is too low.

Lastly, all of the remaining steps required to compute the estimator have a time complexity that scales at a rate of $O(n\log(n))$ or less in the sample size. Since the regression does not require that G be defined using the exact k_n-nearest neighbors graph, or even that each vertex have exactly k_n adjacent vertices, one could also approximate this step. One approach would be to sort each dimension, and adding edges between i and j if and only if they are near one another in each dimension. While this may seem nonstandard initially, this would provide more neighbors to vertices in regions with a higher mass. For a more standard approach that also appears promising, as well as a survey of related algorithms, see Andoni and Indyk (2006). Lastly, note that the algorithm is computationally efficient without using this generalization, so it is probably only worthwhile to do this in very large datasets. Section 5.3 provides more detail on the computation time of the algorithm and comparisons to other methods.

Algorithm 2 Dijkstra's (1959) Algorithm

function `Dijkstra's Algorithm`$(\underline{i}, \bar{i}, w)$
$t,\ s \leftarrow \infty_n$ # Initialize the distance vector, s, and t. Note t_i will contain the
vertex preceding i in the shortest path
$s_{\underline{i}} \leftarrow 0$
$v \leftarrow \{1, 2, ..., n\}$ # Initialize v
while $\bar{i} \in v$**:**
  $k \leftarrow \arg\min_{i \in v} s_i$
  $v \leftarrow v/k$
  for $i \in v$**:** # Consider all unvisited vertices
    temp $\leftarrow s_k + w_{i,k}$
    if temp $< s_i$**:** # If including k in the path to i shortens the path,
      $s_i \leftarrow$ temp # update the path length and update t_i.
      $t_i \leftarrow k$
$\omega^\star \leftarrow \{\bar{i}\}$, $u \leftarrow n + 1$
while $t_{\omega(1)} \neq \underline{i}$**:** # Retrace the path given in t from $\bar{i}$, to $\underline{i}$:
  $\omega \leftarrow \{t_{\omega(1)}, \omega\}$
return ω

Endpoints

When $d = 2$, finding values of x that the modal manifold passes through is generally easily achieved by plotting the datapoints. When this manifold cannot be described by a function of x, bounds on y may also need to be specified in the same way. After this has been done, the endpoints $z_{\bar{i}}$ and $z_{\underline{i}}$ can be specified, as described in the Notation section. This approach appears to be slightly more accurate than other approaches in the two dimensional case.

In higher dimensions, this approach may be more difficult to implement, and this difficulty is compounded by the fact that more endpoints are required when $d > 2$. In this case, it is best if the conditional mode can be expressed as a function of x, so we will suppose this is the case.[5] Next we will describe an automated process of defining these endpoints to be datapoints near the boundary of the X domain that are also close to the conditional mode.

To identify points near the boundary of X, we can use a technique called convex hull peeling (Small, 1990; Rousseeuw, Ruts, and Tukey, 1999). This involves estimating the convex hull of $\{x_i\}_i$, removing the points on the convex hull, and repeating until a sufficient number of points around the boundary are found. Let Γ denote this set of points.

Next we find the set of m points in Γ that are as far from one another as possible. Specifically, we would like to find the set $\Theta \subset \Gamma$, where $|\Theta| = m$, that solves $\arg\max_{\Theta} \min_{i,j \in \Theta} \|z_i - z_j\|$. This is analogous to a covering problem, so it is likely not possible to solve in polynomial time. However, for our purposes, an approximate solution will suffice. To achieve this, choose an arbitrary point in Γ to initialize Θ and then add the point to Θ that is the solution to $\arg\max_{j \in \Gamma} \min_{i \in \Theta} \|z_i - z_j\|$. This process is then continued until $|\Theta| = m$. Now we can calculate a kernel density estimate of $\{z_i\}_i$, and then, for each i in Θ, we can find its k_n nearest neighbors and define an endpoint to be the neighbor with the maximum density estimate. Our current implementation chooses $k_n = 2 \left\lfloor n^{1/3} \right\rfloor$.

In both the $d = 2$ and $d > 2$ cases, there are also two other options available when setting the endpoints. One alternative would be to define all of the k_n nearest neighbors of $i \in \Theta$ as a single endpoint. In this case, for each $i \in \Theta$ we add a vertex to our graph, say $n + i$, that is adjacent to each of the k_n nearest neighbors of i and define the distance from these neighbors to $n + i$ as zero. This is analogous to adding the endpoints to the variables being optimized over in (3). Second, one can also use Parzen's (1962) modal estimator to estimate the global mode of $\{z_i\}_i$ and require that all paths start at this point. This appears to work well when the global mode is near the center of the domain, X.

The method's speed and accuracy is fairly robust to these choices. Most of these methods

[5] It is also possible to maintain the manifold perspective of the conditional mode by embedding the manifold in a lower dimensional space before estimating the endpoints. The algorithms described by Tenenbaum, De Silva, and Langford (2000) and Belkin and Niyogi (2002) can be used to estimate these embeddings.

can also be combined with one another in reasonable way. For example, in $d = 2$, we can set the endpoints to be the k_n nearest neighbors of $i \in \Theta$ while using an estimate of the global mode as the starting point.

Combining the Paths

In higher dimensions, we recommend increasing the the number of endpoints as well as the number of times paths are estimated between each endpoint. To ensure each of these paths are distinct, the edges in the lifted graph $\mathcal{G}$ along ω should be removed before recomputing ω. In the unlifted graph G, this translates to requiring that the paths between iterations do not contain the same three consecutive points. When one has reason to believe that $z^\star(t)$ can be formulated as a function of x and $d = 2$, these estimates can simply be averaged using an arithmetic mean at each value of x. To estimate modal manifolds without assuming $f(y \mid x)$ is unimodal at each $x \in X$, we recommend using a regularized Wasserstein barycenter. More detail can be found on the third paper.

When $d > 2$, a better approach would be to estimate the regression over a mesh $\tilde{X} \subset X$. To do this, in a given iteration, let Φ be the collection of the datapoints that are in at least one path. After calculating the Delaney triangulation of $\{x_i\}_{i\in\Phi}$, the barycentric coordinates of each of the points in $\tilde{X}$ are found. At each point in the mesh, we then calculate a weighted mean of $\{y_i\}_{i\in\Phi}$, with weights given by the barycentric coordinates. After this is repeated over all of the iterations, we define the final regression as the average of the estimates from each iteration.

In terms of actual computation time, when there are approximately 1,000 observations, it generally takes approximately two seconds to find the paths. Combining the paths using an arithmetic mean takes a negligible amount of time; however, defining the final regression using the Wasserstein barycenter adds approximately 35 seconds to the computation time (regardless of the sample size). When using an arithmetic mean, the computation time of our approach compares favorably to the three publicly available implementations of estimators based on Parzen's (1962) approach that we have tested when $n \geq 1,000$ even when a reasonable bandwidth choice was already known (Einbeck and Tutz, 2006; Zhou and Huang, 2018; Chen et al., 2016). When using a Wasserstein barycenter to combine the paths, the required sample size to reach parity with the fastest of these approaches is approximately 2,000.

Figure 5 provides an example of averaging the paths using a Wasserstein barycenter when $d = 2$, and Figure 6 provides an example using the arithmetic mean when $d = 3$. The next section describes the data generating process used to generate Figure 5, and provides

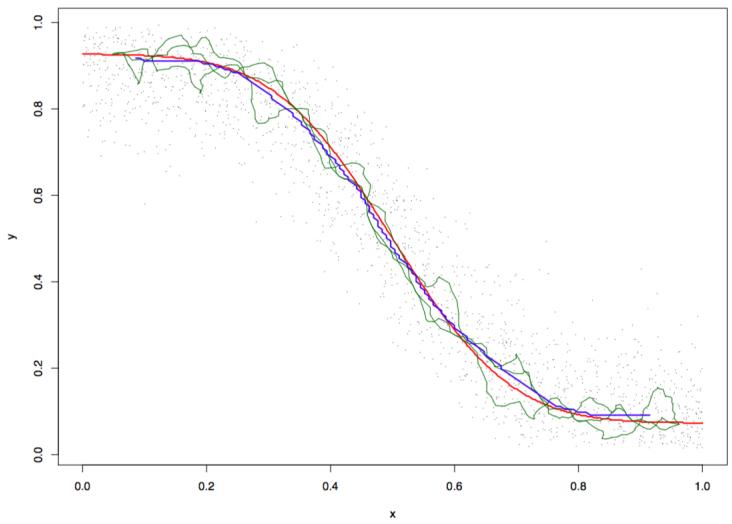


Figure 5: The true conditional mode is given by the red curve, the three paths defining the modal regression by the green curves, and the final estimate of the conditional mode is shown in blue.

simulations of this data generating process. Both the independent variables in Figure 6, say x and y, are uniformly distributed between -2 and 2, and the dependent variable is defined as $z_i = x_i^2 - y_y^2 + \epsilon_i$, where $\epsilon_i \sim N(0, 1/2)$. We will conclude this section with a brief outline of our proposed algorithm, which is given below.

Algorithm 3 A brief summary of the proposed algorithm

1) Define G using a $\lfloor n^{\beta} \rfloor$ −nearest neighbors graph.
2) Find the endpoints of the paths, say $\{\nu_i\}_i$.
3) Define $\mathcal{G}$ as the lifted graph of G with edge weights between vertices (i, j) and (j, k) equal to $\|z_{i,j} - z_{j,k}\|^2$.
4) Perform the following iterations.

for $i \in \{1, 2, ..m\}$**:**
  for $j, k \in \nu$ **and** $k \neq j$**:**
    $\underline{i} \leftarrow \nu_j \times \{1, ..., n\}$; $\overline{i} \leftarrow \nu_k \times \{1, ..., n\}$
    $\omega_{i,j,k} \leftarrow$`Dijkstra's Algorithm`$(\underline{i}, \overline{i}, \mathcal{G})$
  end
  $\mathcal{G} \leftarrow$`Remove Edges`$(\omega_{i,\cdot,\cdot}, \mathcal{G})$
  $\check{z}_i(t) \leftarrow$`Interpolate`$(z, \omega_{i,\cdot,\cdot}, \mathcal{G})$
end

5) Use the regularized Wasserstein barycenter or an average to combine $\{\check{z}_i(t)\}_{i=1}^m$.

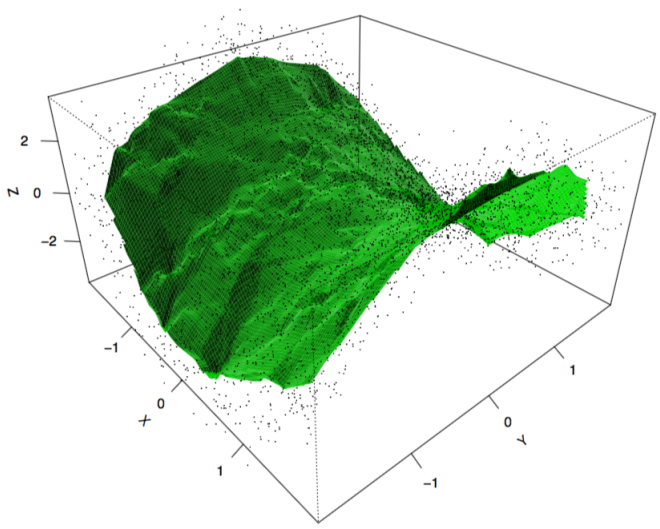


Figure 6: An estimate of the conditional mode in three dimensions using an average to combine the information contained in each path.

2.6 Simulations

Lemma 1 implies that the shortest power weighted path through the unlifted graph G results in a path that balances a tradeoff between paths with lower arc length and paths that pass through regions in which $f(x, y)$ is large asymptotically. Paths that maintain only one of these two objectives would converge to either a straight line or the path that follows the gradient of $f(x, y)$, respectively.[6] Our first set of simulations was designed to test the performance of the estimator when the distance between the path that follows the gradient of $f(x, y)$ is far from the path that follows the mode of $f(y \mid x)$. Since our endpoints are equal to $z^\star(0)$ asymptotically, these two paths are exactly identical if $z^\star(t)$ follows a straight line or if $f(x^\star(t), y^\star(t))$ is a constant for all $t \in [0, 1]$.

When $x_i \sim U(0, 1)$ and $y_i \sim Beta(\alpha(x_i), \beta(x_i))$, where $\alpha(x_i) := c_1 - (c_1 - c_2)\Phi(x_i)$, $\beta(x_i) := c_2 + (c_1 - c_2)\Phi(x_i)$, $c_1, c_2 > 1$, and $\Phi(\cdot)$ is a normal density function with mean $1/2$ and variance σ^2, the conditional mode can be made to be fairly far from the path of steepest ascent in the region near $1/2 - \sigma$ as well as $1/2 + \sigma$. This is due to the change in the slope of the path $z^\star(t)$ coinciding with a change in $f(x^\star(t), y^\star(t))$. For our simulation we chose $c_1 = 20$, $c_2 = 2.5$, and $\sigma = 0.15$. The regression in Figure 5 is an example of this data generating process with 2,000 datapoints.

The endpoints were found using Parzen's (1962) univariate modal estimator using the

[6]There are estimators of this later path, which is called the skeleton of the density. They are closely related to the kernel density ridge estimator, and in two dimensions, the two estimators are nearly identical (Chen et al., 2015).

lowest and highest $2\left\lfloor n^{1/3} \right\rfloor$ datapoints and a bandwidth equal to one half of Scott's (1992) rule of thumb. Afterward three paths were calculated and the Wasserstein barycenter was used to aggregate these three curves into a single estimate. These choices are maintained throughout this section, with one exception that we describe below. The addition of the smoothing parameter described in the Section 4 also appears to result in a gain in accuracy, so choosing this parameter with cross validation is advantageous when one is willing to pay the price of additional computational costs.

To compare the estimators to those based on Parzen's (1962) approach, we first needed to find a reasonable method to select the bandwidth. This is generally more difficult than setting bandwidth parameter of kernel density estimators for two reasons. First, in the case of kernel density estimators, reasonable results can often be found by normalizing each dimension of the data by the standard deviation and then setting a single bandwidth parameter. However, in the context of modal regressions, one can obtain a substantial gain in accuracy by setting the bandwidth parameter of the dependent variable independently of the parameter used for the covariates (Zhou and Huang, 2018). Second, in the modal regression setting, the optimal rate of convergence of these parameters is slower than the corresponding rate in the context of kernel density estimators (Chen et al., 2016). Chen (2018) provides a survey of available bandwidth selection approaches.

To choose this method, we estimated a lower bound on the accuracy of all approaches that use a single bandwidth. Specifically, we found the minimum of the mean integrated squared error (MISE) of all bandwidths, uniformly spaced in an interval, and compared these minimum MISE values to the MISE of estimates using the method described by Zhou and Huang (2018). We found that the method proposed by Zhou and Huang (2018) produced a MISE that was 15% above this minimum value at a sample size of 7,000 and 19% below this value at a sample size of 250. At least in this case, the method of Zhou and Huang (2018) appears to provide at least comparable accuracy to the methods that normalize the data by their standard deviation; see for example, (Chen et al., 2016).[7]

Additionally, we tested methods using conditional density estimation, such as the method proposed by Einbeck and Tutz (2006), but we choose to use Zhou and Huang's (2018) approach because it also compared favorably to these alternatives for this data generating process. Zhou and Huang (2018) also provide evidence through simulations of their approach's favorable accuracy over a range of sample sizes and data generating processes. Table 1 consists of the average MISE over 500 simulations for sample sizes of 250, 500, 1000, 2000, and

[7]Chen (2018) recommend the approach of Zhou and Huang (2018) over the method proposed by Chen et al. (2016), partly due to increased statistical efficiency. Regardless, we also tested both approaches for this data generating process, and Zhou and Huang's (2018) method does appear to be more accurate in this case.

DGP #1: MISE (×1,000,000)					
	Sample Size				
Method	*250*	*500*	*1000*	*2000*	*7000*
Path Reg.	1158	771	555	419	229
KDE Reg.	1060	779	596	458	280

Table 1: The mean integrated squared error for the first data generating process described in the text.

DGP #2: MISE (×1,000,000)					
	Sample Size				
Method	*250*	*500*	*1000*	*2000*	*7000*
Path Reg.	1595	1154	815	607	326
KDE Reg.	1134	683	413	279	147

Table 2: The mean integrated squared error for the second data generating process described in the text.

7000 from the data generating process described above. The estimator based on Parzen's (1962) approach is abbreviated *KDE Reg.* and our proposed method is abbreviated *Path Reg.*

The table shows that the proposed regression compares favorably to modal regressions based on kernel density estimators for this data generating process in sample sizes larger than 250. Although it was not our intention when choosing this data generating process, this may be caused by the fact that the local bias of the kernel density estimator is proportional to the second derivative of the population density. Since $z^\star(t)$ and $f(z^\star(t))$ are nonlinear in the same region in this data generating process, this bias may reduce the accuracy of estimators based on a kernel density estimator. Even though $\nabla_{s,s} z^\star(s)$ and $t \mapsto f(z^\star(t))$ are not constant in this data generating process, it is worth noting that their values are also not extreme. In particular, the ratio $\max_t \ f(z^\star(t)) / \min_t \ f(z^\star(t))$ is approximately equal to two, which is likely less extreme than applications in which the density of the covariates have tails.

Next we will explore a distribution in which $f(z^\star(t))$ is constant. Our next set of simulations defined the distribution of x_i as $U(-1, 1)$ and y_i as $x_i^2 + \epsilon_i$, where $\epsilon_i \sim N(0, 1/10)$. The average MISEs are provided in the following table. We can see that in this case, the proposed estimator performance falls relative to the accuracy of the estimate based on the kernel density estimator.

In our last set of simulations, x and y are distributed as $t_i \sin(t_i) + \epsilon_i$ and $t_i \cos(t_i) + \eta_i$, where t_i is a uniform discretization of $[\pi, 4\pi]$ and η_i, ϵ_i are iid $N(0, 1/10)$ random variables. Note that Figure 4 is similar to this distribution except we defined t_i to be a

DGP #3: L^2 ($\times$10,000)					
	Sample Size				
Method	*250*	*500*	*1000*	*2000*	*7000*
Path Reg.	1862754	3588	2989	2707	2817
KDE Reg.	12438	4919	2441	1451	829

Table 3: The mean integrated squared error for the third data generating process described in the text.

uniform discretization from $[\pi, 3\pi]$ in that case, for clarity of the individual paths. Since this is not a function, the MISE would not be a good choice. Instead, we define $g(t) := \min_u \|(u\sin(u), u\cos(u)) - \hat{z}(t)\|$, and integrate over t, $\int_0^{4\pi} g(t)^2 dt$. The results of the simulation are given in the table below.

There are two less than satisfying aspects of this set of simulations. First, there is a rather large error at a sample size of 250. This is caused by the regression occasionally skipping from $(2, 0)$ all the way to $(4, 0)$. Luckily, this appears to become more rare in larger sample size. Second, the error also appears to be fairly high in the largest two sample sizes, and the error actually increases when moving to a sample size of 7,000. Our working hypothesis is that this is due to a combination of using a mesh of only 150 vertices, which may be too course for manifolds this circuitous, and a slower rate of consistency when $z^\star(\cdot)$ cannot be expressed as a function of x when no smoothing parameter is used. We also ran this simulation again with sample sizes of 2,000 and 7,000, using a smoothing parameter of 10, and found the error metric was reduced to 2,273 and 1,754 respectively.[8]

2.7 UIUC Undergraduate Admissions

We will use the preceding methods to estimate the mode of the University of Illinois at Urbana-Champaign (UIUC) undergraduate students' cumulative GPA at the time of graduation conditional on their high school GPA and their ACT scores. The dataset was provided by the UIUC Office of Undergraduate Admissions, and after cleaning the dataset it consists of 21,650 undergraduate students who graduated from UIUC between 2000 and 2010. The dataset also includes the high school GPA and their ACT scores for non-graduates but not their UIUC GPA.

We will use the dataset to estimate the most typical value of UIUC GPA, conditional on ACT score and high school GPA. Specifically, if we define y as UIUC GPA, x_1, x_2 as high school GPA and ACT score respectively, and the joint density of these variables as

[8]We experimented with a few different values before choosing 10, but this value can likely be improved.

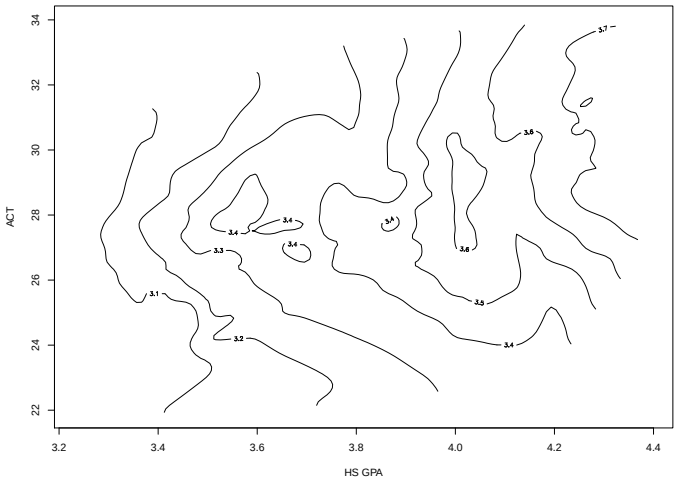


Figure 7: An estimate of the mode of undergraduate UIUC GPA conditional on ACT score and high school GPA.

$f(x_1, x_2, y)$, we will estimate, $m(x_1, x_2) := \arg\max_y f(y \mid x_1, x_2)$. We define the starting point of the paths as the global mode of the data. To specify the endpoints we find eight points near the boundary of the covariates. For each of these endpoints, we estimate the shortest path from the global mode to any of the 120 nearest neighbors of these endpoints. This process was repeated over eight iterations, and the final estimator took approximately one and a half minutes to calculate when β was set equal to 1/3. The contour plot of the resulting estimate is shown in Figure 7.

One interesting feature of the regression is that the estimate of the typical UIUC GPA appears to level off somewhat above a high school GPA of 4.0 and an ACT score of 26. This is likely due to the fact that many of the students with 4.0 GPAs at high schools that do not offer Advanced Placement or honors courses, which are typically graded on a 5.0 GPA scale, had the requisite abilities to go beyond this bound if these opportunities were available to them.

This estimate also appears to show that high school GPA has a larger influence on these typical values of UIUC GPA than ACT scores. Note that Rothstein (2004) and Geiser and Santelices (2007) show that a similar observation can be made using methods based on means at other large universities. To go beyond a describing the academic performance of past UIUC undergraduate students and discuss the counterfactual of the effects of a change in the university's admissions strategy, there are a few sample selection issues worth discussing. For example, we do not observe those that choose not to apply, that are not admitted, and those that choose not to enroll; however, admission decisions would not influence the

OLS Estimates: Graduation Model:		
	A	*B*
ACT	0.0013	0.0014
(s.e.)	0.0007	0.0007
HS GPA	0.1248	0.0909
(s.e.)	0.0071	0.0070

Table 4: Impact of ACT scores and high school GPA on cumulative UIUC GPA.

likelihood of attendance of non-applicants or non-enrollees. We will also frame the discussion around changes to the university's admissions strategies on the margin to avoid the sample not being representative after the new strategy is implemented. This of course assumes that there will not be a discrete change in the applicants' choices from this small change. The last group for which UIUC GPA, at the time of graduation, is not observable is the students that do not graduate, which will be a part of our discussion of the university's admissions strategy.

If increases in high school GPA is more strongly associated with increases in the typical UIUC GPA than ACT scores, it may be evidence that UIUC could adopt an alternative admissions strategy to increase the typical UIUC GPA by placing more weight on high school GPA. Alternatively, it is possible that it is difficult to find covariates that are correlated with the likelihood to graduate, and if all the students did graduate, Figure 7 would look differently. This seems unlikely for two reasons. First, these students only make up 15% of our sample, so they would have to be fairly tightly grouped on a manifold, that has a relatively smaller difference between the slope with respect to ACT and high school GPA, to make the shortest paths switch to this alternative.[9]

Second, it appears that it is straightforward to find covariates associated with the likelihood to graduate. The table below includes OLS estimates and standard errors for two simple models in which the dependent variable is a binary variable that is equal to one if the student graduated and zero otherwise. Model A only includes an intercept, high school GPA, and ACT score, and Model B adds controls for race and gender. The p-values for the coefficient on ACT score for these models were 0.069 and 0.054, respectively, while the corresponding values for high school GPA were on the order of $2 \cdot 10^{-16}$ in both models.

There are also a variety of alternative explanations as well. Including the more likely ex-

[9]An obvious sufficient condition for the modal manifolds of $\alpha f(x,y) + (1-\alpha)g(x,y)$, where $\alpha \in (0,1)$, to be equal to the modal manifolds of $f(x,y)$ is for $\alpha = 0$. We can derive a sufficient condition for equality of the slopes of the modal manifolds being equal using the definition of the modal manifolds from Second 2. Implicit differentiation implies this condition is $\nabla_{y,y} g(x,y) = 0$ and $\nabla_{x,y} g(x,y) = 0$.

OLS Estimates: UIUC GPA Model:		
	A	*B*
ACT	0.015	0.016
(s.e.)	0.001	0.001
HS GPA	0.346	0.299
(s.e.)	0.009	0.008

Table 5: Impact of ACT scores and high school GPA on cumulative UIUC GPA.

planation that the university is optimizing over variables other than probability of graduation and the most typical UIUC GPA. It could be that ACT scores are more strong associated with these other variables. More alternatives include that the university already places the optimal weight on ACT scores for each student, or that the difference in slopes in Figure 7 was not statistically significant.

In the interest of completeness, we also ran the same simple OLS regressions on UIUC GPA. It appears that ACT score does appear to influence UIUC cumulative GPA for graduating students; though the OLS coefficient is far smaller than the coefficient corresponding to high school GPA. The estimates for this model is given below. The nonparametric modal regression happens to describe a similar qualitative relationship to the OLS estimate in this case. Specifically, that ACT scores are weakly associated with future UIUC GPA but to a much smaller degree than high school GPA.

2.8 Conclusion

While most of the work on nonparametric modal regressions have used Parzen's (1962) univariate modal estimator as a starting point; see for example, (Scott, 1992; Einbeck and Tutz, 2006; Chen et al., 2016), in this paper we show that shortest paths through random geometric graphs also provide a promising starting point for nonparametric modal estimators. The advantage, as well as the disadvantage, of most modal estimators is that they are influenced by a decreasing proportion of the data. This allows them to have a breakdown point that converges to zero; however, it also results in an increasing proportion of the data having little, if any, influence on the estimator. Thus, they are particularly advantageous in large sample sizes; one of the central advantages of the proposed method is computational efficiency when the sample size is large.

Though there are some required choices, such as how to choose the endpoints, the number of endpoints, the number of paths to calculate from each endpoint, and how to combine

the paths into a final estimator, we believe these choices are more amenable to being set independently of the data generating process than bandwidth parameters when estimating a modal function. Even if this is not the case, each of these choices only has a small number of reasonable values, and generally there are several combinations of these choices that yield a similar estimate. Another central advantage of the proposed method is robustness to nonlinearities in $z^\star(t)$ and $f(z^\star(t))$ occurring in the same region. Ongoing areas of research include showing that Lemma 1 is applicable to $\mathcal{G}$, exploring the estimation of modal manifolds further, and deriving asymptotic confidence intervals for the estimator.

Chapter 3

Minimizing the Regularized Wasserstein Metric

Abstract

After Cuturi (2013) provided a method to approximate the Wasserstein metric in nearly linear time with the entropy regularized Wasserstein metric, there have been many novel applications that minimize this objective function; see also, (Altschuler et al., 2017). This is often achieved with a first order optimization method. In this paper, we provide a method to evaluate products with the Hessian as well as a method for evaluating products with its inverse. We show the time complexity and memory requirements of both methods are $O(m \log(m))$ when the input densities are supported in $\mathbb{R}^d$. Thus, Newton steps can be evaluated with the same nearly linear time complexity that is required to evaluate the regularized Wasserstein metric. After providing these methods, we discuss their use in existing optimization implementations. In the linear constraint case, we show that matrix decompositions can be avoided, and, in certain cases, that these optimization problems can be solved with second order methods in $O(m \log(m))$ time. To test the method in the nonlinear constraint case, we apply the methods proposed here to the computation of shape constrained density estimators.

3.1 Introduction

One could measure the discrepancy between the distributions $F : \mathcal{A} \to \mathbb{R}^1_+$ and $G : \mathcal{B} \to \mathbb{R}^1_+$, where $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^d$, in a variety of ways, including the Hellinger distance, Kullback-Leibler divergence, and the $p-$Wasserstein metric, which is given by,

$$\inf_{\varphi} \; E_{\varphi}(\|x - y\|_p^p)^{1/p} \text{ subject to:}$$

$$\text{for all } A \subset \mathcal{A} \text{ and } B \subset \mathcal{B},\ \varphi(\mathcal{B}, A) = F(A) \text{ and } \varphi(B, \mathcal{A}) = G(B).$$

Measures in the last class have a few attractive properties, including that they satisfy the triangle inequality; see for example, (Villani, 2003). They are also natural measures of discrepancy when the support of the input densities do not overlap. For example, measuring the discrepancy between two Dirac delta functions, centered at $z_1, z_2 \in \mathbb{R}^d$, with any of the other measures of discrepancy listed above, will only result in one of two values: zero when $z_1 = z_2$ and infinity otherwise. In contrast, the $p-$Wasserstein metric is equal to $\|z_1 - z_2\|_p$.

When the densities are supported on $\mathbb{R}^d$, or a Riemannian manifold, and are uniformly continuous, these measures of discrepancy are approximated after discretizing. In the generic case, evaluating the $p-$Wasserstein metric requires a higher computational cost compared to its counterparts, having a time complexity of $O(m^3)$ when the mesh consists of m points. In addition, attempting to minimize the $p-$Wasserstein metric over one of the input densities subject to a set of constraints often requires m to be prohibitively large to avoid oscillations in the minimizing density, particularly when $d \geq 2$; see for example, (Carlier et al., 2015; Cuturi and Peyré, 2016).

Cuturi (2013) proposed an approximation of the $2-$Wasserstein metric, known as the regularized Wasserstein distance, and Altschuler et al. (2017) showed that this approximation has a time complexity that is proportional to approximating a convolution with a normal density, which can be evaluated in $O(m \log(m))$ time using a fast Fourier transform. Afterward, Solomon et al. (2015) generalized this approach to Riemannian manifolds. These new computational approaches paved the way to novel computational methods that rely on minimizing the regularized Wasserstein distance in statistics (Bernton et al., 2019), machine learning (Cuturi and Doucet, 2014; Cuturi, 2013), object interpolation (Solomon et al., 2015), and partial differential equations (Cuturi and Peyré, 2016); for a textbook treatment, see also, (Peyré and Cuturi, 2019).

When minimizing the Wasserstein metric, it is desirable for the optimization algorithm to have a few properties in particular. Since the regularized Wasserstein distance and its derivatives are only defined when the mass of both of the input densities are equal, algorithms that minimize this function must satisfy this mass normalization constraint in each iteration. Also, since these optimization problems are often high dimensional, it is advantageous for the algorithm to have a time complexity that is no worse than that of finding the regularized Wasserstein distance.

There are two commonly used algorithms that achieve these goals. First, after showing the gradient of the regularized Wasserstein metric is only unique up to an additive constant, Cuturi and Doucet (2014) proposed defining this additive constant to ensure mass normalization and using gradient descent. Second, Benamou et al. (2015) provided an algorithm

using alternating Bregman projections. This is a particularly efficient method when the Bregman projections (the constrained minimum of the Kullback-Leibler divergence) have a closed form solution. In these cases, the method has the advantage of not requiring the regularized Wasserstein metric to be computed before each iteration. Peyré and Cuturi (2019) provide more detail on these algorithms, extensions, and alternative approaches.

While some methods use L-BFGS or automatic differentiation, fully second order approaches are not the norm. This is the case, in part, because the Hessian is dense, generally poorly scaled, and only unique up to a choice of generalized inverse. These difficulties are compounded by the fact that, for most applications of interest, the input densities are multi-dimensional, so m is prohibitively large to efficiently evaluate any choice of generalized inverse used to define this Hessian, since the time complexity and memory requirements of these methods scale at a rate of $O(m^3)$.

Rather than focusing on calculating the Hessian itself, in this paper we provide methods to evaluate products with the Hessian and its generalized inverse. Our requirement that the mass normalization constraint holds in every iteration restores uniqueness of both of these functions. We show that evaluating products with the generalized inverse of the Hessian simply requires two convolutions, as well as elementwise operations on vectors in $\mathbb{R}^m$, so this operation is particularly efficient. Then we provide a simple preconditioner for this matrix and show that the number of iterations required by the preconditioned conjugate gradient method, to evaluate a product with the Hessian, is bounded by a constant as m diverges. Thus, the time complexity and memory requirements of the methods are both $O(m \log(m))$.

Afterward, we describe how these methods can be used in standard second order optimization algorithms to minimize the regularized Wasserstein metric subject to constraints. The linear constraint case is particularly computationally efficient because solving the dual problem has the effect of substituting products with the Hessian for (less computationally expensive) products with the generalized inverse of the Hessian, as well as avoiding matrix factorizations. This provides a few circumstances in which the memory requirements and time complexity of second order methods scale at a rate of $O(m \log(m))$.

There are also options to increase the efficiency of second order methods in the nonlinear constraint case, which we discuss in the context of a primal-dual interior point method. Next, we provide numerical experiments in which shape constrained density estimators are defined by minimizing the regularized Wasserstein metric subject to a set of nonlinear constraints. The dominant cost when solving these optimization problems were convolutions, and, even in this non-linear constraint case, the ratio of convolutions required to evaluate Hessian products to those required to find the objective function ranged from 0.97 to 1.71, and this metric of performance appears to scale well as m increases and as the regulariza-

tion parameter decreases. Given the advantages of second order methods in terms of the typical total number of iterations required, and the attainable level of accuracy resulting from local quadratic convergence, this appears to be a promising approach for minimizing the regularized Wasserstein metric.

Since the regularized Wasserstein metric is calculated using an iterative procedure, that is often terminated prior to its iterates fully converging, one may be concerned whether Newton steps are less accurate than gradient based steps. For this reason, we also do not assume that this iterative process has converged. Instead, we define the Hessian so that it is consistent with the gradient, in the sense that it is the Jacobian of the gradient, regardless of the number of iterations used. This may be advantageous to those who wish to trade the accuracy of the approximation to the unregularized Wasserstein metric for lower computational costs offered by methods that often converge within a few dozen iterations.

The next section introduces notation and key concepts, while the third section provides results on the Hessian and gradient. Other results that may be of independent interest are also provided in the third section, including recommendations for applying the methods here to the regularized Wasserstein barycenter. The fourth section discusses our proposed approaches to evaluate products with the Hessian and its generalized inverse. After a transformation of variables to ensure these matrices are well scaled, they can be passed directly to many existing optimization implementations that support the conjugate gradient method, such as those provided by `LANCELOT` or `KNITRO` (Conn et al., 1988; Byrd et al., 2006). However, there are a few ways to increase computational efficiency further in both the linear as well as the nonlinear constraint cases, which are also discussed in the fourth section. The fifth section provides the numerical experiments, and the sixth concludes.

3.2 Preliminaries

Suppose $\mu_i : \mathcal{A} \to \mathbb{R}^1_{++}$, where $\mathcal{A} \subset \mathbb{R}^d$ or is a $d-$dimensional Riemannian manifold. We will denote the points contained in a mesh over $\mathcal{A}$ by $\{\boldsymbol{a}_i\}_{i=1}^m$. While most of our discussion will be focused on triangular and uniform meshes, other options are also provided by Crane et al. (2013).

Let $\mathbf{1}$ denote an $m \times 1$ vector of ones. We will define the vectors $\boldsymbol{\mu_1}, \boldsymbol{\mu_2} \in \mathbb{R}^m$ so that $\boldsymbol{\mu_1}_i = \mu_1(\boldsymbol{a}_i)$ and $\boldsymbol{\mu_2}_i = \mu_2(\boldsymbol{a}_i)$ for $i \in \{1, ..., m\}$. We will suppose throughout that all densities, say $\boldsymbol{\mu} \in \mathbb{R}^m$, have been multiplied by the area weights associated with the mesh so that $\int_{\mathcal{A}} \mu(x)\mathrm{d}x = \mathbf{1}^\top \boldsymbol{\mu} = 1$.

Also, for $A \in \mathbb{R}^{m\times m}$, we will denote the matrix with $(i,j)^{th}$ element given by $\exp(A_{ij})$ as $\exp(A)$ and a similar convention will be used for $\log(A)$. Also, a diagonal matrix with a main

diagonal given by the vector x will be denoted by D_x, element-wise division of the vectors x and y by $x \oslash y$, and element-wise multiplication by $x \otimes y$.

The $p-$Wasserstein Metric

Optimal transport began with the work of Gaspard Monge in the 18^{th} century when he posed the problem of finding the optimal method of moving a pile of sand to a nearby hole, each of the same volume. Specifically, suppose the amount of mass in the pile over $B \subset \{\boldsymbol{a}_i\}_i$, as well as the amount of mass required to fill in the hole over $B \subset \{\boldsymbol{a}_i\}_i$, is given by $\sum_{i \in B} \boldsymbol{\mu_1}_i$ and $\sum_{i \in B} \boldsymbol{\mu_2}_i$ respectively. Also, suppose the cost of moving one unit of sand from $\boldsymbol{a}_i$ to $\boldsymbol{a}_j$ is equal to $M_{ij} \in \mathbb{R}_1^+$.

To solve this problem in its general form, Kantorovich (1958) introduced the concept of a transportation plan, or coupling, which in our discrete case is defined as $\varphi \in \mathbb{R}_+^{m \times m}$, where φ_{ij} is the amount of mass transported from $\boldsymbol{a}_i$ to $\boldsymbol{a}_j$. Feasibility of this transportation plan requires that $\boldsymbol{\mu_1}$ is the density of the sand before it is relocated, so any feasible φ must satisfy $\varphi \mathbf{1} = \boldsymbol{\mu_1}$. Also, after the mass is relocated, it must have a density given by $\boldsymbol{\mu_2}$, so $\varphi^\top \mathbf{1} = \boldsymbol{\mu_2}$. Note that this also implies that any element in the set of feasible couplings can be viewed as a discretized density with support given by $\mathcal{A} \times \mathcal{A}$. In summary, Kantorovich (1958) defined the optimal coupling as the solution to the optimization problem,

$$\min_{\varphi} \sum_{i,j=1} M_{ij} \varphi_{ij} \text{ such that:}$$
$$\varphi \mathbf{1} = \boldsymbol{\mu_1},\ \varphi^\top \mathbf{1} = \boldsymbol{\mu_2}.$$

Despite the initial specificity of the motivation for optimal transport, there are a wide variety of questions that can be answered with reformulations of this concept. For example, the $p-$Wasserstein metric can be defined as,

$$\mathcal{W}_{0,p}(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) := \min_{\varphi} \left(\sum_{i,j=1} d(\boldsymbol{a}_i, \boldsymbol{a}_j)^p \varphi_{ij} \right)^{1/p} \text{ such that:}$$
$$\varphi \mathbf{1} = \boldsymbol{\mu_1},\ \varphi^\top \mathbf{1} = \boldsymbol{\mu_2},$$

where $d(\boldsymbol{a}_i, \boldsymbol{a}_j)$ is the distance between $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ over $\mathcal{A}$, which is $\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_p^p$ when $\mathcal{A} \subset \mathbb{R}^d$. The prototypical value of p used in applications is two, and since an approximation of the $2-$Wasserstein Metric is our primary focus, in the subsequent sections we will denote $M \in \mathbb{R}^{m \times m}$ as the matrix with elements given by $M_{ij} = d(\boldsymbol{a}_i, \boldsymbol{a}_j)^2$.

Regularized Wasserstein Distance

Cuturi (2013) proposed the following approximation of $\mathcal{W}_{0,2}(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})^2$.

$$\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) := \min_{\varphi} \sum_{i,j=1} \varphi_{ij} \left(M_{ij} + \gamma \log(\varphi_{ij}) - \gamma\right) \text{ such that:} \tag{3.1}$$

$$\varphi \mathbf{1} = \boldsymbol{\mu_1},\ \varphi^\top \mathbf{1} = \boldsymbol{\mu_2} \tag{3.2}$$

The addition of the negative entropy term, given by $\gamma\varphi_{ij}\log(\varphi_{ij})$, penalizes couplings with large changes in nearby values of φ_{ij}, which has the effect of damping the oscillations in the minimizer of $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$. The additional $-\gamma\varphi_{ij}$ term is less critical, as it will not influence the optimal coupling. Its purpose is to provide a slight simplification in the dual problem.

To see how the addition of negative entropy also provides a large gain in computational efficiency, we can define the corresponding Lagrangian,

$$\mathcal{L} = \textstyle\sum_{i,j=1} \varphi_{ij} \left(M_{ij} + \gamma \log(\varphi_{ij}) - \gamma\right) - \boldsymbol{\lambda}_1^\top \left(\varphi \mathbf{1} - \boldsymbol{\mu_1}\right) - \boldsymbol{\lambda}_2^\top \left(\varphi^\top \mathbf{1} - \boldsymbol{\mu_2}\right).$$

The first order conditions imply that the optimal coupling is

$$\varphi_{ij} = \exp\left(\boldsymbol{\lambda}_{1i}/\gamma\right) \exp\left(-M_{ij}/\gamma\right) \exp\left(\boldsymbol{\lambda}_{2j}/\gamma\right).$$

In other words, there exists $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}_+^m$ such that the optimal coupling is given by $\varphi_{ij} = K_{ij}\boldsymbol{w}_i\boldsymbol{v}_j$, where $K_{ij} := \exp\left(-d(\boldsymbol{a}_i, \boldsymbol{a}_j)^2/\gamma\right)$. This can also be written as,

$$\varphi = D_{\boldsymbol{w}} K D_{\boldsymbol{v}},$$

so adding the entropy term to the objective function reduces the dimensionality of the optimization problem from m^2 to $2m$. Sinkhorn (1967) shows that φ is uniquely defined by the constraints on its row and column sums. However, clearly φ can also be written as $D_{\boldsymbol{w}c} K D_{\boldsymbol{v}/c}$, for any $c \in \mathbb{R}_{++}^1$, so $\boldsymbol{w}$ and $\boldsymbol{v}$ are unique up to c. Also, $\boldsymbol{w}$ and $\boldsymbol{v}$ can be found efficiently using the iterative proportional fitting procedure (IPFP); see for example, (Sinkhorn, 1967; Krupp, 1979). After initializing $\boldsymbol{w}$ to be $\mathbf{1}$, this method iteratively redefines $\boldsymbol{v}$ so that $D_{\boldsymbol{v}} K \boldsymbol{w} = \boldsymbol{\mu_2}$, and subsequently redefines $\boldsymbol{w}$ so that $D_{\boldsymbol{w}} K \boldsymbol{v} = \boldsymbol{\mu_1}$, which is summarized in Algorithm 4. Combining the constraints on the row and column sums of φ shows that $\boldsymbol{w}$ must satisfy,

$$D_{\boldsymbol{w}} K \left(\boldsymbol{\mu_2} \oslash (K\boldsymbol{w})\right) = \boldsymbol{\mu_1}, \tag{3.3}$$

which will be helpful in the subsequent sections.

Algorithm 4 The iterative proportional fitting procedure.

function IPFP$(K, \boldsymbol{\mu_1}, \boldsymbol{\mu_2})$
$\boldsymbol{w} \leftarrow \mathbf{1}$
until convergence:
  $\boldsymbol{v} \leftarrow \boldsymbol{\mu_2} \oslash (K\boldsymbol{w})$
  $\boldsymbol{w} \leftarrow \boldsymbol{\mu_1} \oslash (K\boldsymbol{v})$
return $\boldsymbol{w}, \boldsymbol{v}$

In the rest of the paper, we will make substantial use of the dual of (1)-(2), which Cuturi and Doucet (2014) show is given by the unconstrained optimization problem,

$$\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) := \max_{(x,y)\in\mathbb{R}^{2m}} \boldsymbol{x}^\top \boldsymbol{\mu_1} + \boldsymbol{y}^\top \boldsymbol{\mu_2} - \gamma \sum_{i,j} \exp\left((\boldsymbol{x}_i + \boldsymbol{y}_j - M_{ij})/\gamma\right). \tag{3.4}$$

The first order conditions of (4) can be written as,

$$\boldsymbol{\mu_1} \oslash (K \exp(\boldsymbol{y}/\gamma)) = \exp(\boldsymbol{x}/\gamma), \tag{3.5}$$

$$\boldsymbol{\mu_2} \oslash (K \exp(\boldsymbol{x}/\gamma)) = \exp(\boldsymbol{y}/\gamma), \tag{3.6}$$

which are analogous to the updates of $\boldsymbol{w}$ and $\boldsymbol{v}$ given in Algorithm 4, with $\boldsymbol{w} = \exp(\boldsymbol{x}/\gamma)$ and $\boldsymbol{v} = \exp(\boldsymbol{y}/\gamma)$.

A few comments regarding the effect of γ on the optimal coupling will also be useful in subsequent sections. Higher values of γ correspond to placing a higher penalty on the negative entropy of the coupling, so the optimal coupling becomes more dispersed as this parameter is increased. Also, in the limit as γ converges to zero, $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$ converges to $\mathcal{W}_{0,2}(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})^2$, and φ converges to the optimal unregularized coupling, both at a rate of $O(\exp(-1/\gamma))$; see Benamou et al. (2015) and Cuturi (2013).

Note that our description of Algorithm 4 requires matrix multiplications with K, so its time complexity is $O(m^2)$. The next section shows how this can be improved.

Convolutional Wasserstein Distance

Reducing the time complexity of Algorithm 4 when $\mathcal{A} \subset \mathbb{R}^d$ is straightforward, since $(K\boldsymbol{z})_i$ is proportional to $\sum_j \boldsymbol{z}_j \exp\left(-\|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2^2/\gamma\right)$, a convolution of $\boldsymbol{z}$ with a Gaussian density. Since this can be carried out in $O(m\log(m))$ time using a fast Fourier transform, $K\boldsymbol{z}$ can simply be replaced by this convolution operation.

Solomon et al. (2015) generalized this approach to the case of Riemannian manifolds. The method can be described using the heat equation, which is given by

$$\triangle_x f_t(x) = \tfrac{1}{2} \cdot \tfrac{\partial f_t(x)}{\partial t} \text{ and } f_0(x) = h(x),$$

where $\triangle_x$ is the Laplacian operator, $f : \mathbb{R}^1_+ \times \mathcal{A} \to \mathbb{R}^1_+$ is the density of heat at time t and $h(x)$ is the initial density of heat. The solution can be found using $f_t(x) = \int_{\mathcal{A}} \phi_t(x, y)h(y)dy$, where $\phi : \mathbb{R}^1_+ \times \mathcal{A} \times \mathcal{A} \to \mathbb{R}^1_+$ is a heat kernel. For example, when $\mathcal{A} = \mathbb{R}^d$, $\phi_t(x, y) = (2\pi t)^{-d/2} \exp\left(-\left\|x - y\right\|_2^2 / (2t)\right)$.

Although $\phi_t(x, y)$ cannot be expressed in a closed form in the more general case in which $\mathcal{A} \neq \mathbb{R}^d$, numerical solutions are still feasible. In the Riemannian manifold setting this amounts to solving

$$(I + \epsilon L)\, \boldsymbol{f}_\epsilon = \boldsymbol{f}_0$$

$$\implies \boldsymbol{f}_\epsilon = \Phi \boldsymbol{f}_0,$$

where $\Phi := (I + \epsilon L)^{-1}$ is a discretized heat kernel and L is a discrete Laplacian matrix associated with the mesh $\{\boldsymbol{a}_i\}_i$. L can be defined as the normalized cotangent Laplacian when a triangular mesh is used, which is described in more detail by Crane et al. (2013).

The link between $\mathcal{W}^2_\gamma(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$ and the heat kernel was provided by Varadhan (1967),

$$d(\boldsymbol{a}_i, \boldsymbol{a}_j)^2 = \lim_{t \to 0} -2t \log(\phi(\boldsymbol{a}_i, \boldsymbol{a}_j, t)).$$

Thus, in the limit as γ go to zero (and temporarily setting aside discretization errors), K converges to Φ, after renormalizing. With this in mind, Solomon et al. (2015) proposed approximating products in Algorithm 4 of the form $\boldsymbol{y} := K\boldsymbol{z}$ by solutions of the linear system, $(I + \gamma/2L)\, \boldsymbol{y} = \boldsymbol{z}$. There are two primary advantages of this approach. First, it does not require the calculation of distances between every pair of elements in $\{\boldsymbol{a}_i\}_i$. Second, a sparse Cholesky factorization of the matrix $I + \gamma/2L$ can be found in an efficient manner, and each solve step after this is computed is computationally inexpensive. The lowest currently known upper bound on this time complexity is $O(m^2)$, but it appears to be nearly linear in practice (Schmitz and Ying, 2010; Crane et al., 2013).

Since $M_{ij} = -\gamma \log(K_{ij})$, the objective function (1) can also be written as,

$$\gamma \left(1 + \textstyle\sum_{i,j=1} \varphi_{ij} \log(\varphi_{ij}/K_{ij})\right),$$

so when K is replaced by $\Phi \in \mathbb{R}^{m \times m}_{++}$ in Algorithm 4, $\varphi := D_w \Phi D_v$ is the minimizer of

$$\min_{\varphi}\ \gamma \left(1 + \sum_{i,j=1} \varphi_{ij} \log(\varphi_{ij}/\Phi_{ij})\right) \text{ such that:} \tag{3.7}$$

$$\varphi \mathbf{1} = \boldsymbol{\mu_1},\ \varphi^\top \mathbf{1} = \boldsymbol{\mu_2}. \tag{3.8}$$

The next section provides a few properties of the function $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$ that will be useful when discussing optimization methods in our setting.

3.3 Properties of The Derivatives

The results in this section concern the gradient and Hessian of $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$ when the kernel is given by $H \in \{K, \Phi\}$. However, to allow for other possibilities, when $H = \Phi$ our only assumptions will be that Φ is invertible and it consists of strictly positive and real elements. In the cotangent Laplacian case, this can be achieved by ensuring the mesh is Delaunay (Spielman, 2010).

As mentioned in the introduction, for higher dimensional applications, which generally require m to be large, Algorithm 4 is often terminated after a fixed number of iterations or when the error of the iterates satisfy a given tolerance level. Afterward, the common approach is to use $\boldsymbol{w}$ and $\boldsymbol{v}$ to perform the assignments $\boldsymbol{x} \leftarrow \log(\boldsymbol{w})\gamma$ and $\boldsymbol{y} \leftarrow \log(\boldsymbol{v})\gamma$, and then define the value of the objective function using (4). Specifically, this value is defined as,

$$\boldsymbol{x}^\top \boldsymbol{\mu_1} + \boldsymbol{y}^\top \boldsymbol{\mu_2}.$$

Note that the final term in (4), $\gamma \sum_{i,j} \exp\left((\boldsymbol{x}_i + \boldsymbol{y}_j - M_{ij})/\gamma\right)$, is generally omitted because it is a constant in every iteration; specifically, $\gamma \sum_{i,j} \exp\left((\boldsymbol{x}_i + \boldsymbol{y}_j - M_{ij})/\gamma\right) = \gamma \sum_{ij} \varphi_{ij} = \gamma m$.[1]

This can lead to inefficiencies in optimization methods because the commonly used gradient does not correspond to this objective function. As will be clear below, this derivative requires that $\boldsymbol{w}$ and $\boldsymbol{v}$ satisfy the first order conditions of (4), which are $\boldsymbol{w} \otimes (H\boldsymbol{v}) = \boldsymbol{\mu_1}$ and $\boldsymbol{v} \otimes (H\boldsymbol{w}) = \boldsymbol{\mu_2}$. Since each iteration of Algorithm 4 ends with the assignment $\boldsymbol{w} \leftarrow \boldsymbol{\mu_1} \oslash (H\boldsymbol{v})$, the equality $\boldsymbol{w} \otimes (H\boldsymbol{v}) = \boldsymbol{\mu_1}$ holds up to numerical precision, but the equality $\boldsymbol{v} \otimes (H\boldsymbol{w}) = \boldsymbol{\mu_2}$ generally does not. These inconsistencies can be resolved by leaving the inputs of Algorithm 4 unchanged, defining $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}} := \boldsymbol{v} \otimes (H\boldsymbol{w})$, and then defining the objective function as,

$$\boldsymbol{x}^\top \boldsymbol{\mu_1} + \boldsymbol{y}^\top \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}.$$

There are two primary advantages of this modification. First, $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$ converges to $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \boldsymbol{\mu_2})$ as the number of iterations in Algorithm 4 diverges in a manner that is transparent to users because $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}} - \boldsymbol{\mu_2}$ is generally used to define the stopping criteria of Algorithm

[1] This property is also independent of the stopping criteria of Algorithm 4, since it only requires that φ is a coupling between two arbitrary densities.

4. Second, this ensures the gradient, Hessian and objective function are consistent with one another. Note that we will not require any modifications of the gradient to bring about this consistency, which ensures that these modifications will not impact the critical points of optimization problems.

We will make a similar modification for the Hessian, which, in light of the gradient remaining unchanged, is somewhat notational; it is simply the accurate Jacobian of the normally used gradient. However, this accurate definition ensures that the Hessian is positive semidefinite numerically, which is a requirement of the proposed method for calculating products with the Hessian that will be described in the next section. For these reasons, we will make a distinction between $\boldsymbol{\mu_2}$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}$.

The next Lemma expresses the derivative of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$; a similar result is provided by Cuturi and Doucet (2014). The Lemma also provides the gradient of $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$, which will be used below to derive the Hessian of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$.

Lemma 1: *For $c \in \mathbb{R}^1$ and $H \in \{\Phi, K\}$, the gradients of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$ and $(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}) \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$ are given by*

$$\nabla_{\boldsymbol{\mu_1}} \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}) = \boldsymbol{x} + c \tag{3.9}$$

$$\nabla_{(\boldsymbol{\mu_1}^\top, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^\top)^\top} \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}) = (\boldsymbol{x}^\top + c, \boldsymbol{y}^\top - c)^\top. \tag{3.10}$$

Proof: Since the objective function in the dual problem, given by (4), is differentiable, we can apply the envelope theorem. This implies $\nabla_{(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})} W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}) = (\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top$. (5) and (6) imply $\boldsymbol{x} := \gamma \log(\boldsymbol{w})$ and $\boldsymbol{y} := \gamma \log(\boldsymbol{v})$. Since $\boldsymbol{w}$ and $\boldsymbol{v}$ yield the same coupling as $\tilde{c}\boldsymbol{w}$ and $\boldsymbol{v}/\tilde{c}$, where $\tilde{c} \in \mathbb{R}^1_{++}$, (9) and (10) follow from the definition $c := \gamma \log(\tilde{c})$.

□

The constant c in Lemma 1 is a result of one of the $2m$ constraints in (1) being redundant because both densities also have the same sum. As mentioned previously, Cuturi and Doucet (2014) minimize the regularized Wasserstein metric using gradient descent and define c as $-\sum_i \boldsymbol{x}_i/m$ to ensure the mass normalization constraint holds in each iteration. This results in the gradient flows of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$ being contained to the domain of $\mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}})$, the unit simplex.

The next theorem uses (10) to derive the Hessian of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_\mathbf{2})$.

Theorem 2: *For $H \in \{\Phi, K\}$, let*

$$B := \left(D_{\boldsymbol{\mu_1}} - \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_\mathbf{2}} \varphi^\top\right) / \gamma.$$

The Hessian of $\boldsymbol{\mu_1} \mapsto \mathcal{W}_\gamma^2(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_\mathbf{2})$ is given by

$$\nabla^2_{\boldsymbol{\mu_1}} W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_\mathbf{2}) = B^\sim, \tag{3.11}$$

where $(\cdot)^\sim$ is a g-inverse.

Proof: We will find the Hessian by finding the Jacobian of $\nabla_{(\boldsymbol{\mu}_\mathbf{1}^\top, \tilde{\boldsymbol{\mu}}_\mathbf{2}^\top)^\top} W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_\mathbf{2})$ from Lemma 1. Recall that $\boldsymbol{w} := \exp(\boldsymbol{x}/\gamma)$, $\boldsymbol{v} := \exp(\boldsymbol{y}/\gamma)$, and the first order conditions of the dual, given by (7) and (8), can be written as

$$\begin{bmatrix} D_{\exp(\boldsymbol{x}/\gamma)} H \exp(\boldsymbol{y}/\gamma) \\ D_{\exp(\boldsymbol{y}/\gamma)} H \exp(\boldsymbol{x}/\gamma) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \tilde{\boldsymbol{\mu}}_\mathbf{2} \end{bmatrix} \tag{3.12}$$

Thus,

$$\begin{aligned} \nabla_{(\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top} (\boldsymbol{\mu}_\mathbf{1}^\top, \tilde{\boldsymbol{\mu}}_\mathbf{2}^\top)^\top &= \begin{bmatrix} D_{\exp(\boldsymbol{x}/\gamma) \otimes (H \exp(\boldsymbol{y}/\gamma))} & D_{\exp(\boldsymbol{x}/\gamma)} K D_{\exp(\boldsymbol{y}/\gamma)} \\ D_{\exp(\boldsymbol{y}/\gamma)} H D_{\exp(\boldsymbol{x}/\gamma)} & D_{\exp(\boldsymbol{y}/\gamma) \otimes (K \exp(\boldsymbol{x}/\gamma))} \end{bmatrix} / \gamma \\ &= \begin{bmatrix} D_{\boldsymbol{\mu_1}} & \varphi \\ \varphi^\top & D_{\tilde{\boldsymbol{\mu}}_\mathbf{2}} \end{bmatrix} / \gamma, \end{aligned}$$

where the second equality follows from (12) and $\varphi := D_{\exp(\boldsymbol{x}/\gamma)} H D_{\exp(\boldsymbol{y}/\gamma)}$.

Thus, $\nabla_{(\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top} (\boldsymbol{\mu}_\mathbf{1}^\top, \tilde{\boldsymbol{\mu}}_\mathbf{2}^\top)^\top$ is also singular because

$$\begin{bmatrix} D_{\boldsymbol{\mu_1}} & \varphi \\ \varphi^\top & D_{\tilde{\boldsymbol{\mu}}_\mathbf{2}} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu_1} - \varphi \mathbf{1} \\ -\varphi^\top \mathbf{1} + \tilde{\boldsymbol{\mu}}_\mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

By a result provided by Ouellette (1981), all g-inverses of $\nabla_{(\boldsymbol{x}^\top, \boldsymbol{y}^\top)^\top} (\boldsymbol{\mu}_\mathbf{1}^\top, \tilde{\boldsymbol{\mu}}_\mathbf{2}^\top)^\top$ have an upper left block that is given by

$$\nabla^2_{\boldsymbol{\mu_1}} W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_\mathbf{2}) = \left((D_{\boldsymbol{\mu_1}} - \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_\mathbf{2}} \varphi^\top)/\gamma\right)^\sim.$$

□

The following theorem provides the first result on the Hessian, which we will use below to discuss the non-uniqueness of the Hessian.

Theorem 3: *For $H \in \{\Phi, K\}$, B is positive semidefinite and has a null space that is spanned by* $\mathbf{1}$. *When $(\cdot)^{\sim} := (\cdot)^{+}$, the Moore-Penrose pseudoinverse, $\nabla^2_{\boldsymbol{\mu_1}} W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\mathbf{2}})$ is also positive semidefinite and has a null space that is spanned by* $\mathbf{1}$.

Proof: Since B is a real symmetric matrix, it admits an orthonormal eigenbasis. Also, note that establishing the required properties of this eigenspace for B will imply that they also hold for B^+, since, for diagonalizable matrices, the pseudoinverse can be defined by inverting each nonzero eigenvalue (Penrose, 1955).

Let $A_1 := D_{\mathbf{1} \oslash \boldsymbol{\mu_1}} \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top$, and note that A_1 also admits an eigendecomposition by a similarity transformation to a symmetric matrix. Feasibility of the optimal coupling implies,

$$D_{\mathbf{1} \oslash \boldsymbol{\mu_1}} \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top \mathbf{1} = D_{\mathbf{1} \oslash \boldsymbol{\mu_1}} \varphi \mathbf{1} = \mathbf{1}.$$

Since $A_1 \mathbf{1} = \mathbf{1}$ and $A_{1,ij} > 0$, the Perron-Frobenius theorem implies that A_1 has one eigenvalue that is 1, with a corresponding eigenvector of $\mathbf{1}/\sqrt{m}$, and $m-1$ eigenvalues that are strictly less than one. Let $A_2 := I - D_{\mathbf{1} \oslash \boldsymbol{\mu_1}} \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top$. If λ and v are an eigenvalue and eigenvector of A_1 respectively, then $A_2 v = (1-\lambda)v$, so A_2 has one eigenvalue that is 0, with a corresponding eigenvector of $\mathbf{1}/\sqrt{m}$, and $m-1$ eigenvalues that are strictly positive. Thus, this property also holds for $D_{\boldsymbol{\mu_1}} A_2 = \gamma B$ since each element of $\boldsymbol{\mu_1}$ is strictly positive.

□

Remark 7: Note that there are more straightforward ways of establishing convexity of $\boldsymbol{\mu_1} \mapsto W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\mathbf{2}})$. For example, this follows from positivity and weak diagonal dominance of B. Alternatively, the objective function in (1) can be viewed as a linear function, plus a strictly convex function, in φ, and $\boldsymbol{\mu_1}$ is linear in φ. Also we can show that $\mathbf{1}$ is in the null space of B by simply multiplying, or by noting that $\boldsymbol{w}$ is the root of $F(\boldsymbol{w}) = \boldsymbol{\mu} - D_{\boldsymbol{w}} H(\tilde{\boldsymbol{\mu}}_{\mathbf{2}} \oslash (H\boldsymbol{w}))$, which is homogeneous of degree zero, so $\nabla_{\boldsymbol{w}} F(\boldsymbol{w}) = \gamma B D_{\mathbf{1} \oslash \boldsymbol{w}}$. These approaches were not used because we require $\boldsymbol{\mu_1} \mapsto W_\gamma(\boldsymbol{\mu_1}, \tilde{\boldsymbol{\mu}}_{\mathbf{2}})$ to be strictly convex over the set $\{\boldsymbol{\mu_1} \mid \mathbf{1}^\top \boldsymbol{\mu_1} = 1\}$.

□

Although the Hessian is not unique, this is not problematic from an optimization standpoint. To see why, we will consider the example of a Newton step $\boldsymbol{d}$, which is defined by the solution to the system $B^{\sim} \boldsymbol{d} = -\boldsymbol{x}$. Since the regularized Wasserstein metric is not defined when $\boldsymbol{\mu_1}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{2}}$ do not have equal mass, we also require the density to have the same mass

in every iteration of the optimization methods; in other words, we need to ensure that the constraint $\boldsymbol{d}^\top \mathbf{1} = 0$ holds in each iteration.

This can be achieved by solving,

$$PB^{\sim}P\boldsymbol{d} = -P\boldsymbol{x},$$

where $P = I - \mathbf{1}\mathbf{1}^\top/m$, the projection matrix onto the set of mean zero vectors. This is because Theorem 3 implies that all eigenvalues of $PB^{\sim}P$ are the same as those of $B^{\sim}$, other than the eigenvalue corresponding to the eigenvector of $\mathbf{1}/\sqrt{m}$. In other words, the addition of these projection operations can be viewed as an implicit choice of a g-inverse and c. The former is equivalent to the Moore-Penrose pseudoinverse because the pseudoinverse of a diagonalizable matrix can be defined by taking the reciprocal of the non-zero eigenvalues while leaving the eigenvalues that are zero unchanged, and clearly $\boldsymbol{x} \leftarrow P\boldsymbol{x}$ ensures that c is zero (Penrose, 1955).[2] The theorem implies that this system can also be written as, $B^{+}\boldsymbol{d} = -\boldsymbol{x}$, as long as $c = 0$, or simply as $\boldsymbol{d} = -B\boldsymbol{x}$, for any $c \in \mathbb{R}^1$. For this reason, it is sometimes convenient to view the Hessian as B^{+}.

Also, note that Theorem 3 implies that the alternative formulation for the Hessian,

$$B^{\sim} = \hat{B}^{-1},$$

where $\hat{B} := B + \mathbf{1}\mathbf{1}^\top/(\gamma m)$, is also a valid approach. We will continue introducing full rank and normalized hessians as well as their generalized inverses throughout this section, with the intention of providing more options when optimizing the regularized Wasserstein metric. All of the full rank hessians, and their inverses, can be derived from their singular counterparts by changing the unique eigenvalue that is zero to $1/\gamma$. In other words, if $\boldsymbol{z}$ is a norm one eigenvector corresponding to the eigenvalue of zero, by adding $\boldsymbol{z}\boldsymbol{z}^\top/\gamma$ to these matrices. Likewise, these full rank matrices can be multiplied by the projection matrix, $P = I - \boldsymbol{z}\boldsymbol{z}^\top$, as above, to recover their singular counterparts. The methods described in the next section, as well as optimization algorithms, will sometimes be more straightforward to implement when nonsingular matrices are used, which will be discussed in more detail below.

An immediate corollary of Theorem 2, is that B satisfies all of the required conditions of a graph Laplacian, which is defined as a symmetric M-matrix with $\mathbf{1}$ in the null space (Spielman, 2010).

[2]Note that this holds for any $B^{\sim}$ because the definition of a g-inverse, given by $BB^{\sim}B = B$, uniquely defines the mapping $\boldsymbol{z} \mapsto B^{\sim}\boldsymbol{z}$ for $\boldsymbol{z}$ is in the range space of B.

Corollary 4: *For $H \in \{\Phi, K\}$, let G denote the complete graph with vertices $V := \{1, ..., m\}$ and weight assigned to the edge connecting $i, j \in V$, where $i \neq j$, given by $\left(\varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top\right)_{ij}$. The Laplacian of G is γB.*

Proof: B is positive semidefinite and has a null space spanned by $\mathbf{1}$ by Theorem 3. Since $\gamma, \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top, \tilde{\boldsymbol{\mu}}_{\mathbf{2}}$ and $\boldsymbol{\mu}_{\mathbf{1}}$ are composed of strictly positive elements, we have $B_{ii} > 0$ and $B_{ij} \leq 0$ for all $i, j \in \{1, ..., m\}$ and $i \neq j$, so B is also a (singular) M-matrix.

□

Many iterative methods for solving the linear system $\boldsymbol{z} = A\boldsymbol{y}$, where $A \in \mathbb{R}^{m \times m}$ is nonsingular, have a rate of convergence that can be expressed in terms of the condition number of A, which is defined as $\lambda_1(A)/\lambda_m(A)$, where $\lambda_1(A) \geq \lambda_2(A) \geq ... \geq \lambda_m(A)$ are the eigenvalues of A (Saad, 2003). When solving $\boldsymbol{z} = B\boldsymbol{y}$ where B is a Laplacian matrix, or a singular matrix more generally, we can express the rate of convergence of iterative methods using the *finite condition number* of the normalized Laplacian, given by,

$$B_{\mathrm{N}} := D_{\mathbf{1} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}} B D_{\mathbf{1} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}},$$

which is equal to $\kappa_f(B_{\mathrm{N}}) := \lambda_1(B_{\mathrm{N}})/\lambda_k(B_{\mathrm{N}})$, where $k = \min_{\lambda_i(B_{\mathrm{N}}) \neq 0} i = m - 1$ (Spielman, 2010). Since $B_{\mathrm{N}} \sqrt{\boldsymbol{\mu}_{\mathbf{1}}} = \mathbf{0}$, its full rank counterpart is given by $\hat{B}_{\mathrm{N}} := B_{\mathrm{N}} + \sqrt{\boldsymbol{\mu}_{\mathbf{1}}} \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}^\top / (\gamma m)$. The next theorem provides a bound on this condition number in the euclidean case, $H = K$. We will discuss in the next section how this implies that using a transformation of variables so that the Hessian is normalized, or using $D_{\boldsymbol{\mu}_{\mathbf{1}}}$ as a preconditioner, provides a significant gain in computational efficiency.

Theorem 5: *Suppose $H = K$, that γ is not dependent on m, and $\boldsymbol{\mu}_{\mathbf{1}i}, \boldsymbol{\mu}_{\mathbf{2}i} > 0$ for all i. Then, there exists $c \in \mathbb{R}^1$, that is independent of m, such that $\lim_{m \to \infty} \kappa_f(B_N) \leq c$.*

Proof: The proof of Theorem 2 defines $A_1 := D_{\mathbf{1} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}} \varphi D_{\mathbf{1} \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}} \varphi^\top D_{\mathbf{1} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}}$ and shows that the eigenvalue corresponding to the eigenvector of $\mathbf{1}/\sqrt{m}$ is $\lambda_1(A_1) = 1$. Since K is a Gaussian kernel matrix (and all elements of $\{\boldsymbol{a}_i\}_i$ are unique) it is positive definite, so $\lambda_m(A_1) \geq 0$.

To find a bound on $\lambda_2(A_1)$, let $A_2 := D_{\mathbf{1} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}} \varphi D_{\mathbf{1} \oslash \sqrt{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}}}$, and note that $A_1 = A_2 A_2^\top$. Also,

$$A_2 = D_{\boldsymbol{w} \oslash \sqrt{\boldsymbol{\mu}_{\mathbf{1}}}} K D_{\boldsymbol{v} \oslash \sqrt{\tilde{\boldsymbol{\mu}}_{\mathbf{2}}}} = D_{\boldsymbol{w} \oslash \sqrt{\boldsymbol{w}\bar{\boldsymbol{v}}}} K D_{\boldsymbol{v} \oslash \sqrt{\boldsymbol{v}\bar{\boldsymbol{w}}}} = D_{\sqrt{\boldsymbol{w} \oslash \bar{\boldsymbol{v}}}} K D_{\sqrt{\boldsymbol{v} \oslash \bar{\boldsymbol{w}}}},$$

which is clearly similar to a symmetric matrix, so it admits an orthonormal eigendecomposition. A_2 is also similar to $D_{\sqrt{\boldsymbol{w}\boldsymbol{v} \oslash \bar{\boldsymbol{w}}\bar{\boldsymbol{v}}}} K$, so we will use the bound,

$$\lambda_2(A_2) \leq \max(\sqrt{\boldsymbol{w} \otimes \boldsymbol{v} \oslash (\bar{\boldsymbol{w}} \otimes \bar{\boldsymbol{v}})}) \lambda_2(K).$$

The eigenvalues and eigenfunctions of the unnormalized Gaussian kernel function, $(x, y) \mapsto \exp(-\|x-y\|^2/\gamma)$, can be expressed in a closed form (Williams and Rasmussen, 2006). The limiting behavior of the k^{th} eigenvalue is $O(\gamma^{d/2}/(1+\gamma+\sqrt{\gamma(1+\gamma)})^{d(k+1/2)})$, as k diverges or as γ converges to zero. In other words, given the sequence $\{\gamma_m\}_{m=1}^{\infty}$ such that $\gamma_m > 0$ and $\lim_{m\to\infty}\gamma_m = 0$, the k^{th} eigenvalue converges to zero at a rate of $O(\gamma_m^{d/2})$. Using the asymptotic relationship between the eigenvalues of K and the eigenvalues of the Gaussian kernel function, as described in more detail by Shawe-Taylor et al. (2006), we have that $\lambda_k(K) = O(m\gamma^{d/2})$ for fixed $k \in \{1, ..., m\}$. To shed more light on the effect of both m and γ on the condition number, we will use $\lambda_2(K) = O(m\gamma^{d/2})$.

By construction, we have $\boldsymbol{\mu_1}_i = O(1/m)$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}i} = O(1/m)$ for all $i \in \{1, ..., m\}$, which are also independent of the limiting behavior of γ. Since $\boldsymbol{\mu_1} = \boldsymbol{w} \otimes \bar{\boldsymbol{v}}$ and $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}} = \boldsymbol{v} \otimes \bar{\boldsymbol{w}}$, we have

$$\boldsymbol{w} \otimes \bar{\boldsymbol{v}} = O(1/m) \text{ and } \boldsymbol{v} \otimes \bar{\boldsymbol{w}} = O(1/m)$$

The effect of increasing m on the magnitude of the elements of $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{v}}$ can be decomposed into a three parts using the eigendecomposition of K/m, which we will denote $U\Lambda U^\top$, where Λ is a diagonal matrix with $\Lambda_{ii} = \lambda_i(K/m)$. First, m has a direct effect on $\bar{\boldsymbol{w}}_i$ because K is not normalized by m. Second, changing the magnitude of $\|\boldsymbol{w}\|$ has a multiplicative impact on $\bar{\boldsymbol{w}}_i$. Third, the inner product of $\boldsymbol{w}/\|\boldsymbol{w}\|$ and U may change the influence that each eigenvalue Λ_{ii} has on $\bar{\boldsymbol{w}}_i$. Combining these three effects, we have $\bar{\boldsymbol{w}} = O(m\|\boldsymbol{w}\|\zeta_{\boldsymbol{w}})$ and $\bar{\boldsymbol{v}} = O(m\|\boldsymbol{v}\|\zeta_{\boldsymbol{v}})$, where $\zeta_{\boldsymbol{w}} := \sum_i \lambda_i(K/m)\boldsymbol{w}_i U_{ij}^2/\|\boldsymbol{w}\|$ and similarly for $\zeta_{\boldsymbol{v}}$. Note that $\zeta_{\boldsymbol{w}}$ and $\zeta_{\boldsymbol{v}}$ can be viewed as weighted means of $\{\lambda_j(K/m)\}_j \cup \{0\}$ with weights given by $\{\boldsymbol{w}_j U_{ij}^2/\|\boldsymbol{w}\|\}_{j=1}^m \cup \{1 - \sum_j \boldsymbol{w}_j U_{ij}^2/\|\boldsymbol{w}\|\}$. Since the elements of $\{\lambda_j(K/m)\}_j$ are only functions of γ, and not m, $\zeta_{\boldsymbol{w}}$ and $\zeta_{\boldsymbol{v}}$ also only depend on γ.

Using these rates in our system of equations implies,

$$O(\boldsymbol{w}) = O(1/\left(m\sqrt{\zeta_{\boldsymbol{v}}}\right)) \text{ and } O(\boldsymbol{v}) = O(1/(m\sqrt{\zeta_{\boldsymbol{w}}})).$$

Since $\bar{\boldsymbol{w}} = O(\zeta_{\boldsymbol{w}}/\sqrt{\zeta_{\boldsymbol{v}}})$, we have $\bar{\boldsymbol{v}} = O(\zeta_{\boldsymbol{v}}/\sqrt{\zeta_{\boldsymbol{w}}})$, so we have,

$$O(\max(\sqrt{\boldsymbol{w} \otimes \boldsymbol{v} \oslash (\bar{\boldsymbol{w}} \otimes \bar{\boldsymbol{v}})}) = O(1/\left(m\sqrt{\zeta_{\boldsymbol{v}}\zeta_{\boldsymbol{w}}}\right)).$$

Using the fact that $O(m\gamma^{d/2})$ is an upper bound on $\lambda_1(K)$, we have

$$\lambda_2(A_2) = \lambda_2(D_{\sqrt{\boldsymbol{w}\otimes\boldsymbol{v}\oslash(\bar{\boldsymbol{w}}\otimes\bar{\boldsymbol{v}})}}K) \leq O(\gamma^{d/2}/\sqrt{\zeta_{\boldsymbol{v}}\zeta_{\boldsymbol{w}}}).$$

Thus,

$$\lambda_2(A_1) = \lambda_2(A_2 A_2^\top) \leq O(\gamma^d/\left(\zeta_{\boldsymbol{v}}\zeta_{\boldsymbol{w}}\right)).$$

Since A_1 is positive definite,

$$\lambda_1(I - A_1) \leq 1,$$

and the upper bound on $\lambda_2(A_1)$ corresponds to the lower bound,

$$\lambda_{m-1}(I - A_1) \geq O(1 - \gamma^d / (\zeta_{\boldsymbol{v}} \zeta_{\boldsymbol{w}})).$$

Thus, the finite condition number of B_{N} , which is equal to that of $I - A_1$, satisfies the asymptotic bound,

$$\kappa_f(I - A_1) := \lambda_1(I - A_1)/\lambda_{m-1}(I - A_1) \leq c = O(1/\left(1 - \gamma^d / (\zeta_{\boldsymbol{v}} \zeta_{\boldsymbol{w}})\right)).$$

□

Remark 8: An extension to the case in which $H = \Phi$ would also be useful. The underlying property that the result requires is that $\lambda_1(H/m)$ does not converge to $\lambda_2(H/m)$ as m diverges, and extending the argument to the case in which $H = \Phi$ is straightforward if one is able to bound this eigenvalue gap for a given mesh. However, it is likely to be difficult to ensure that this condition holds for a general mesh.

□

To test how large m must typically be for the bound on the finite condition number to be evident, we also calculated $\kappa_f(B_{\mathrm{N}})$ for $m \in \{50, 100, 150, ..., 700\}$ and $\gamma \in \{0.0025, 0.01, 0.0175, 0.025\}$ for $\boldsymbol{\mu_1} = \mathrm{Beta}(8, 5)$ and $\boldsymbol{\mu_2} = \mathrm{Beta}(3, 9)/2 + \mathrm{Beta}(9, 3)/2$. The lowest value of γ was chosen so that $\|\varphi \mathbf{1} - \boldsymbol{\mu_1}\|_\infty$ was approximately 10^{-10} after $2,000$ iterations of the IPFP method for each value of m. Figure 8 provides these input densities and the resulting condition numbers. The maximum difference between the condition numbers over all values of m (for fixed γ) was 7×10^{-5}.

Before moving on, it is worth noting that the results and methods provided here can also be generalized in a straightforward way to the case of the regularized Wasserstein barycenter, which, in the two input density case, is given by, $\arg\min_\eta \xi \mathcal{W}_\gamma^2(\eta, \boldsymbol{\mu_1}) + (1 - \xi)\mathcal{W}_\gamma^2(\eta, \boldsymbol{\mu_2})$ for $\xi \in (0, 1)$. As stated in the introduction, the alternating Bregman projections method provided by Benamou et al. (2015) is particularly efficient in this setting, and it is likely difficult to improve on this approach with a second order method in early iterations. Instead, our recommendation is to to start with this approach in order to initialize a second order method. This has the advantage increasing the likelihood that the second order method will exhibit a locally quadratic rate of convergence. The approach provided by Benamou et al. (2015) also has the advantage of not requiring Algorithm 4 to be run in each iteration,

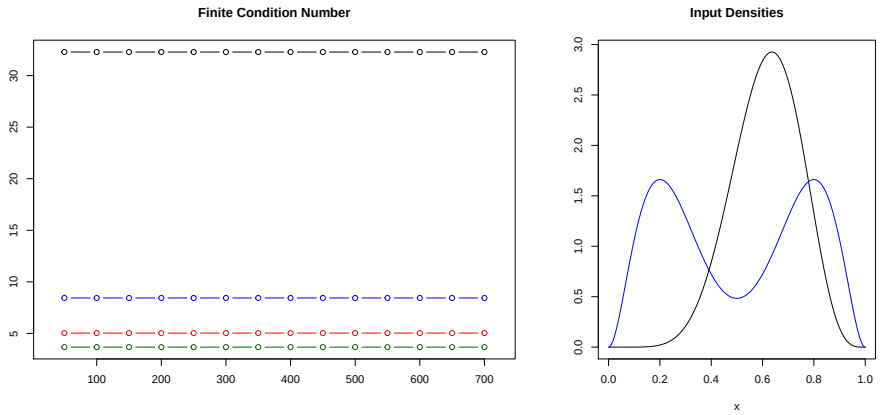


Figure 8: On the left plot, the black, blue, red, and green curves provide the finite condition number of B_{N} when γ is equal to $0.0025, 0.01, 0.0175$, and 0.025 respectively. The horizontal axis corresponds to the values of m. The input densities are provided on the right.

as it is itself a generalization of Algorithm 4. Thus, it is also capable of providing the couplings, $\varphi_1 = D_{\boldsymbol{w_1}} H D_{\boldsymbol{v_1}}$ and $\varphi_2 = D_{\boldsymbol{w_2}} H D_{\boldsymbol{v_2}}$, from $\boldsymbol{\eta_1} := \boldsymbol{w_1} \otimes (H\boldsymbol{v_1})$ to $\boldsymbol{\mu_1}$ and from $\boldsymbol{\eta_2} := \boldsymbol{w_2} \otimes (H\boldsymbol{v_2})$ to $\boldsymbol{\mu_2}$, respectively.

The first order conditions for the barycenter can be found using Theorem 1. In terms of $\boldsymbol{w_1}$ and $\boldsymbol{w_2}$, these can be written as $\xi\gamma \log(\boldsymbol{w_1}) + (1-\xi)\gamma \log(\boldsymbol{w_2}) = \mathbf{0}$. After combining these equations with the definitions of $\boldsymbol{v_i}$ and $\boldsymbol{w_i}$ for $i \in \{1, 2\}$, we can show the barycenter is defined, by the solution the system of equations $\boldsymbol{\eta_1} = \boldsymbol{\eta_2}$, and,

$$\boldsymbol{v_1} = \boldsymbol{\mu_1} \oslash (H\boldsymbol{w_1}),\ \boldsymbol{w_1} = \boldsymbol{\eta_1} \oslash (K\boldsymbol{v_1}),$$

$$\boldsymbol{w_2} = \boldsymbol{\eta_2} \oslash (K\boldsymbol{v_2}), \boldsymbol{v}_2 = \boldsymbol{\mu_2} \oslash (K\boldsymbol{w_2}), \text{ and } \boldsymbol{w_1}^{\xi} \boldsymbol{w_2}^{1-\xi} = \mathbf{1}.$$

While deriving the Newton step direction, the only equation that we will not require to hold in each iteration is $\boldsymbol{\eta_1} = \boldsymbol{\eta_2}$. Since, after leaving this equation out of the system, the rest of the variables can be recovered from $\boldsymbol{w_2}$, we will view these other variables as functions of $\boldsymbol{w_2}$. In this case, the system can be reduced to,

$$\boldsymbol{w_2} \otimes (H(\boldsymbol{\mu_2} \oslash (H\boldsymbol{w_2}))) - \boldsymbol{w_2}^{(\xi-1)/\xi} \otimes \left(H\left(\boldsymbol{\mu_2} \oslash (H\boldsymbol{w_2}^{(\xi-1)/\xi})\right)\right) = 0,$$

which can be solved using Newton's method. As noted in Remark 7, $\nabla_{\boldsymbol{w_i}}(\boldsymbol{\eta}_i - D_{\boldsymbol{w_i}} H(\boldsymbol{\mu_i} \oslash (H\boldsymbol{w_i}))) = \gamma B_i D_{\mathbf{1} \oslash \boldsymbol{w_i}}$, where $B_i := (D_{\boldsymbol{\eta_i}} - \varphi_i D_{\mathbf{1} \oslash \boldsymbol{\mu_i}} \varphi_i^{\top})/\gamma$ for $i \in \{1, 2\}$. If we denote the residual of the system in the current iteration as $\boldsymbol{b}$, this implies the search direction, $\boldsymbol{p}$, can be found by solving,

$$\left[\gamma \ \left(B_1 + (1-\xi)/\xi B_2\right) \ D_{\mathbf{1} \oslash \boldsymbol{\eta_1}}\right] (\boldsymbol{p} \otimes H\boldsymbol{v_1}) = -\boldsymbol{b}.$$

Since a matrices eigenvalues are continuous functions of the elements, Theorem 5 implies that there exists $\epsilon > 0$ such that if $\|\boldsymbol{\eta_1} - \boldsymbol{\eta_2}\| < \epsilon$, then $\kappa_f(\ \gamma B_2\ D_{\mathbf{1} \oslash \boldsymbol{\eta_1}})$ is $O(1)$ as m diverges. Thus, if $\|\boldsymbol{\eta_1} - \boldsymbol{\eta_2}\| < \epsilon$, we have $\kappa_f(\gamma\ (B_1 + (1-\xi)/\xi B_2)\ D_{\mathbf{1} \oslash \boldsymbol{\eta_1}})$ is also $O(1)$ in m. This implies that the techniques discussed in the next section are also capable of calculating $\tilde{\boldsymbol{p}} := -(\gamma\ (B_1 + (1-\xi)/\xi B_2)\ D_{\mathbf{1} \oslash \boldsymbol{\eta}_1})^{-1} \boldsymbol{b}$ with a provably nearly linear time complexity when $\|\boldsymbol{\eta_1} - \boldsymbol{\eta_2}\| < \epsilon$. Afterward, one can define the search direction as $\boldsymbol{p} \leftarrow \tilde{\boldsymbol{p}} \oslash (H \boldsymbol{v_1})$. This method can be generalized to more than two input density, but the resulting system of equations becomes coupled in this case.

The next section will use the results shown above to provide methods to evaluate products with $\hat{B}$ and $\hat{B}^{-1}$ efficiently.

3.4 Numerical Methods

In this section we will outline our proposed approaches for evaluating the matrix products required for optimization. We will begin by describing our approach for products with $\hat{B}$, which can be carried out in a particularly efficient manner, as this only requires two convolutions and element-wise operations of vectors in $\mathbb{R}^m$. The method follows directly from the definition of this matrix, and is summarized in Algorithm 5. We denote the convolution with a Gaussian with variance $\gamma/2$, or, in the case of non-Euclidean meshes, the solution of the prefactored linear system, by $H \circledast \boldsymbol{z}$ to emphasize that matrix products are not a requirement.

Algorithm 5 Evaluates a product with the (full rank) inverse Hessian.

$\hat{B}(\boldsymbol{p})$**:**
$\boldsymbol{q} \leftarrow (\boldsymbol{v}^2 \oslash \tilde{\boldsymbol{\mu}}_{\mathbf{2}}) \otimes (H \circledast (\boldsymbol{p} \otimes \boldsymbol{w}))$
$\boldsymbol{r} \leftarrow \boldsymbol{p} - \boldsymbol{w} \otimes (H \circledast \boldsymbol{q}) + \mathbf{1}^\top \boldsymbol{p}/m$
return $\boldsymbol{r}/\gamma$

One approach to solve systems of the form $\boldsymbol{z} = A\boldsymbol{y}$, where $A \in \mathbb{R}^{m \times m}$ is a symmetric and positive definite matrix, is the conjugate gradient method; see for example (Saad, 2003; Nocedal and Wright; 2006). The solution produced by this method can be viewed as a linear combination of an $A-$orthogonal basis in the Krylov space, which, in our formulation, can be defined as the span of $\{\boldsymbol{z}, A\boldsymbol{z}, A^2\boldsymbol{z}, ...\}$.[3] When A is positive definite, the conjugate gradient method will converge within m iterations, but when the spectrum of A is favorable,

[3]For the purpose of a concise exposition, we formulate the algorithm with the initialization $\boldsymbol{y_0} \leftarrow \mathbf{0}$. If this is not done, the iterates are in the Krylov space $\{\boldsymbol{r_0}, A\boldsymbol{r_0}, A^2\boldsymbol{r_0}, ...\}$, where $\boldsymbol{r_0}$ is the initial residual, $\boldsymbol{z} - A\boldsymbol{y_0}$.

this is a conservative upper bound. For example, if A has only r distinct eigenvalues, the method converges within r iterations (Nocedal and Wright, 2006). More generally, the method converges to $\tilde{\boldsymbol{z}}$, such that $\|\boldsymbol{z} - \tilde{\boldsymbol{z}}\|_A \leq \epsilon \|\boldsymbol{z}\|_A$, where $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^\top A \boldsymbol{x}}$, within $O(\sqrt{\kappa(A)} \log(1/\epsilon))$ iterations (Saad, 2003). This result also extends to singular matrices, as long as the system is consistent, and the bound in this case is $O(\sqrt{\kappa_f(A)} \log(1/\epsilon))$. Although, in practice it is best to ensure the mean of $\boldsymbol{r_k}$ is zero in each iteration by demeaning this vector directly; Kaasschieter (1988) outlines this approach and alternative modifications for stability of the conjugate gradient method in the singular case.

The conjugate gradient method is also used in optimization implementations to calculate Newton steps. In this context, it is sometimes stopped early or initialized in different ways. While this is not related to evaluating products with the inverse Hessian, we will refer back to it in the next two sections. In this setting, A can be viewed as a Hessian and $\boldsymbol{z}$ can be viewed as a negative gradient, and the conjugate gradient method produces iterates that are monotonically decreasing in the second order Taylor series expansion,

$$q(\boldsymbol{y_k}) := \boldsymbol{y_k}^\top A \boldsymbol{y_k}/2 - \boldsymbol{z}^\top \boldsymbol{y_k}.$$

In this sense, the iterates improve on the initial value of $\boldsymbol{y_0}$. This allows for the method to be warm started when an initial approximation of $\boldsymbol{y}$ is known, such as the last Newton step. Steihaug (1983) also proposed a simple way to ensure the method produces steps that satisfy trust region constraints, of the form $\|\boldsymbol{y}\| \leq d$; this is done by initializing $\boldsymbol{y_0}$ at $\mathbf{0}$, and, if $\|\boldsymbol{y_{k+1}}\| > d$, the method returns $\boldsymbol{y_k} + \alpha \boldsymbol{y_{k+1}}$, where α satisfies, $\|\boldsymbol{y_k} + \alpha \boldsymbol{y_{k+1}}\| = d$. Steihaug (1983) showed that this initialization implies the iterates will satisfy $\|\boldsymbol{y_{k+1}}\| \geq \|\boldsymbol{y_k}\|$, which, combined with the fact that $q(\boldsymbol{y_k})$ is monotonically decreasing, implies that this strategy provides the lowest possible value of $q(\boldsymbol{y_k})$ along the path of convergence.

The efficiency of the conjugate gradient method can also be improved with the use of a preconditioner. Given a preconditioner for A, say $\tilde{A}$, and factoring this matrix as $\tilde{A} = CC^\top$, this is analogous to solving the system,

$$C^{-1} A C^{-1} (C\boldsymbol{y}) = C^{-1} \boldsymbol{z}$$

in two steps. First, we solve $C^{-1} A C^{-1} \tilde{\boldsymbol{y}} = C^{-1} \boldsymbol{z}$ using the conjugate gradient method and then define the final solution as $\boldsymbol{y} = C^{-1} \tilde{\boldsymbol{y}}$. Note that the actual preconditioned conjugate gradient method uses $\tilde{A}$ rather than C explicitly; Algorithm 6 presents the method in the context of evaluating $\hat{B}^{-1} \boldsymbol{z}$. As discussed in the previous section, projections and rescalings can be applied to the input and output of Algorithm 5 and 6 to evaluate products with their normalized or singular counterparts.

Algorithm 6 Evaluates products with the (full rank) Hessian.

$\hat{B}^{-1}(\boldsymbol{z})$**:**
$\boldsymbol{y_0} \leftarrow \boldsymbol{0}$; $\boldsymbol{r_0} \leftarrow -\boldsymbol{z}$
$\boldsymbol{s_0} \leftarrow \boldsymbol{r_0} \oslash \boldsymbol{\mu_1}$; $\boldsymbol{p_0} \leftarrow -\boldsymbol{s}$
$k \leftarrow 0$
While $\boldsymbol{r_k} \neq \boldsymbol{0}$**:**
  $\alpha_k \leftarrow \boldsymbol{r_k}^\top \boldsymbol{y_k} / \boldsymbol{p_k}^\top \hat{B}(\boldsymbol{s_k})$
  $\boldsymbol{y_{k+1}} \leftarrow \boldsymbol{y_k} + \alpha_k \boldsymbol{p_k}$
  $\boldsymbol{r_{k+1}} \leftarrow \boldsymbol{r_k} + \alpha_k \hat{B}(\boldsymbol{s_k})$
  $\boldsymbol{s_{k+1}} \leftarrow \boldsymbol{r_{k+1}} \oslash \boldsymbol{\mu_1}$
  $\beta_k \leftarrow \boldsymbol{r_{k+1}}^\top \boldsymbol{s_{k+1}} / \boldsymbol{r_k}^\top \boldsymbol{s_k}$
  $\boldsymbol{p_{k+1}} \leftarrow \beta \boldsymbol{p_k} - \boldsymbol{s_{k+1}}$
  $k \leftarrow k + 1$
End
Return $\boldsymbol{y_{k+1}}$

When using the preconditioned conjugate gradient method, the total iteration count is $O(\sqrt{\kappa(\tilde{A}^{-1}A)}\log(1/\epsilon))$ (Saad, 2003). Since $\kappa(D_{\boldsymbol{1} \oslash \boldsymbol{\mu_1}} \hat{B}) = \kappa_f(\hat{B}_N)$ and Theorem 5 implies $\kappa_f(\hat{B}_N) = O(1)$ as m diverges, an immediate implication is that products with the Hessian can be found with a time complexity of $O(\log(1/\epsilon))$ times the time complexity of a convolution, or a suitable approximation, in the Euclidean case, which is summarized in the following Theorem.

Theorem 6: *Suppose the conditions of Theorem 5 hold, and let L denote the time complexity of either a Gaussian convolution or an approximation of a Gaussian convolution with $\lambda_1(H/m) - \lambda_2(H/m) \nrightarrow 0$. Then, Algorithm 6 can be used to evaluate $\hat{B}^{-1}\boldsymbol{z}$ to an accuracy of $\|\boldsymbol{z} - \tilde{\boldsymbol{z}}\|_{D_{\boldsymbol{1} \oslash \boldsymbol{\mu_1}}} \leq \epsilon \|\boldsymbol{z}\|_{D_{\boldsymbol{1} \oslash \boldsymbol{\mu_1}}}$ in $O(L\log(1/\epsilon))$ time. This time complexity is $O(m\log(m)\log(1/\epsilon))$ if the convolution is approximated with a fast Fourier transform.*

Proof: The preceding discussion, and Theorem 5, imply that the preconditioned conjugate gradient method requires $O(\log(1/\epsilon))$ iterations as m diverges as long as the eigenvalue gap is bounded as m diverges (Saad, 2003). The dominant cost of each iteration is evaluating products with $\hat{B}$. Each product is $O(L)$, since it requires two convolutions, as well as element wise algebraic operations on vectors of length m, so the total cost of evaluating $\hat{B}^{-1}\boldsymbol{z}$ is $O(L\log(1/\epsilon))$.

□

Remark 9: To preserve generality, the eigenvalue gap property of the Gaussian kernel that was used in Theorem 5 is stated as an assumption. Note that as long as approximations

Hessian Product Times

$\gamma = 0.0025$				
m	500	1,000	1,500	2,000
CG	0.012	0.047	0.124	0.205
LU	0.036	0.278	0.871	2.01
MPI	0.555	5.389	17.875	-
$\gamma = 0.01$				
m	500	1,000	1,500	2,000
CG	0.007	0.028	0.067	0.106
LU	0.047	0.349	0.855	1.994
MPI	0.549	5.350	17.872	-

Table 6: Each figure is the average time, in seconds, over five evaluations of the Hessian product. LU refers to the LU decomposition of $\hat{B}$ and MPI to the Moore-Penrose pseudoinverse of B; the time required to form these matrices was not included in these figures. The MPI decomposition for $m = 2,000$ timed out after 30 seconds for both values of γ.

are mean preserving, $\lambda_1(H/m) = 1$ by the Perron-Frobenius theorem, so this is primarily an assumption on $\lambda_2(H/m)$.

Since our goal is in part optimization to a high level of accuracy, in this paper we use a convolution with a Gaussian that is truncated to four standard deviations from its mean, which has a time complexity of $O(m^2)$. There are also a variety of alternatives that provide approximations of the convolution that are available with a time complexity of $O(m \log(m))$ or even $O(m)$ (Getreuer, 2013). While the fast Fourier transform is not always ideal when γ is small relative to m, since it can become unstable or produce negative output, it does converge to the desired result as m diverges for fixed γ.

□

The timings for the method proposed here, along with those for the Moore-Penrose pseudoinverse of B and an LU decomposition of $\hat{B}$, can be found in Table 6. The conjugate gradient method was written in `C++`. In addition, references to timings throughout this paper refer to timings on a MacBook Air, circa 2015. The input densities were the same as those described in Figure 8, and the product of the Hessian with the negative gradient was evaluated in each case. Since it is common to use fairly low values of γ, that still allow for convergence of Algorithm 4, we only report the two lowest values here. The behavior shown in Table 6 continued as γ increased; with times for the proposed approach ranging between 0.004 and 0.073 when γ was 0.025.

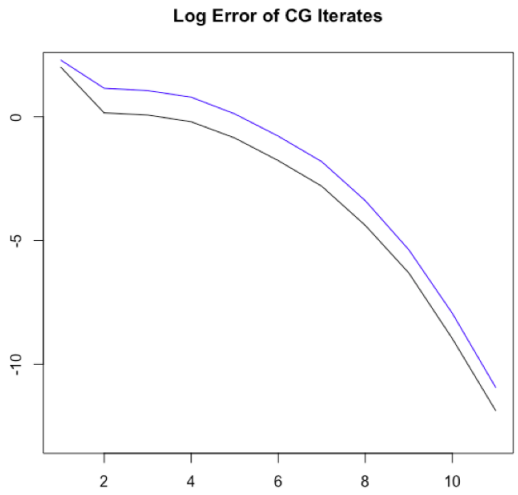


Figure 9: Provides the absolute error values, defined as $\log_{10}(\|B\boldsymbol{y}_k - \boldsymbol{z}\|)$, in each iteration of Algorithm 6 for the input densities from Figure 8 and $\gamma = 0.01$. The black and blue lines correspond to the case in which $m = 100$ and $m = 10,000$, respectively.

Algorithm 6 was terminated when $\|\boldsymbol{r_k}\| \leq \|\boldsymbol{y_k}\| \, 10^{-6}$. When $\gamma = 0.0025$ and $\gamma = 0.01$, the algorithm required 14 and 8 iterations, respectively, for all values of m. To provide a better idea of the error in each iteration, Figure 9 provides $\log_{10}(\|B\boldsymbol{y}_k - \boldsymbol{z}\|)$ with $\gamma = 0.01$ and m equal to 100 and $10,000$, which shows the method converged to a high level of accuracy around the eighth iteration in both cases. The reason this performance is possible is because of the favorable spectrum of $\hat{B}_N$; for example, when $m = 500$ and $\gamma = 0.0025$, $\hat{B}_N$ has 477 eigenvalues that are within 10^{-4} of 1. For comparison, there are only 25 eigenvalues of B that are within 10^{-4} of any other eigenvalue, and $\kappa_f(B)$ is of the order 10^6.

As discussed in the introduction, Algorithms 5 and 6 can be used directly in a variety of optimization software; however, some properties that are specific to the objective function provide a few opportunities to increase the efficiency of these methods. This will be discussed in the next two subsections, in the context of linear and nonlinear constraints respectively.

Linear Constraints: The Projected Newton Method

Since Newton's method can be viewed as gradient descent on a basis that is dictated by the Hessian, performing a projection onto the feasible set in a different basis (such as an orthogonal projection) will generally not result in iterates that converge. Bertsekas (1982) provided a simple solution to this problem by proposing the removal of the nondiagonal elements in the rows and columns of the Hessian that correspond to the active set.

Many of the current approaches of projected Newton methods can still be viewed as separating out the variables in the active set in an analogous way. Thus, these methods

are particularly efficient when this orthogonal projection can be computed efficiently, so they are often described in the context of optimization subject to bound constraints; see for example, (Conn et al., 1988; Moré and Toraldo, 1991; Lin and Moré, 1999). In our setting, the ability to take products with the inverse Hessian efficiently allows for a much wider array of possibilities. Similar implementations to the one described here include `TRON` and `LANCELOT` (Conn et al., 1988; Lin and Moré, 1999). The primary advantages of the method in our setting is that they allow for the use of the conjugate gradient method to avoid matrix factorizations. The iterates also often converge quickly, even when compared to other second order active set approaches, because an arbitrary number of constraints can be added or removed from the active set in each iteration.

We will consider solving the optimization problem given by,

$$\min_{\boldsymbol{\mu}} \mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2}) \text{ subject to:}$$

$$A\boldsymbol{\mu} \geq b,$$

where $A \in \mathbb{R}^{n \times m}$ is full row rank, with a sequentially quadratic programming (SQP) method, but before continuing, it is worth discussing how the mass normalization constraint should be imposed. Suppose $B^\sim = \hat{B}^{-1}$; in this case the Karush-Kuhn-Tucker (KKT) system corresponding to this optimization problem is,

$$\begin{bmatrix} \hat{B}^{-1} & -A^\top \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{x} \\ \boldsymbol{b} \end{bmatrix}.$$

This matrix is non-singular because $\hat{B}$ is full rank, so in this case, c can be defined to be $\min_i \boldsymbol{x}_i$ so that mass normalization can be ensured using the inequality constraint $\mathbf{1}^\top \boldsymbol{\mu} \geq 1$. Or, it can also be ensured by simply adding an equality constraint to the formulation above.

However, considering the case in which $B^\sim = B^+$ sheds light on a potential problem that may arise when the Hessian is not full rank in the null space of A. If this is not the case, so that $A\mathbf{1} = \mathbf{0}$, then

$$\begin{bmatrix} B^+ & -A^\top \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

Thus, the KKT system is singular, and if a solution exists, it is not unique. For this reason, when the Hessian is defined as B^+ we will assume $A\mathbf{1} \neq \mathbf{0}$. If this holds, the mass normalization constraint is automatically ensured because the Newton step takes the form, $B\left(A^\top \boldsymbol{\lambda} - \boldsymbol{x}\right) = \boldsymbol{p}$. Also, c can be set equal to zero in this case. We will focus on applying the SQP method in the full rank case in the interest of generality.

In each iteration, this method finds a primal search direction $\boldsymbol{p} := \boldsymbol{\mu} - \boldsymbol{\mu}^{(i)}$, by solving a quadratic program given by,

$$\min_{\boldsymbol{p}} \boldsymbol{r}^\top \boldsymbol{p} + \boldsymbol{p}^\top \hat{B}^{-1} \boldsymbol{p}/2 \text{ subject to:}$$

$$A(\boldsymbol{\mu}^{(i)} + \boldsymbol{p}) \leq \boldsymbol{b}.$$

Since the objective function is differentiable and strictly convex, this can be solved using its dual, which is advantageous for a few reasons. First, this only requires bound constraints, which allows for efficient algorithms like the projected Newton method. Second, we have a closed form solution for $\hat{B}$, which will allow for line searches to be carried out efficiently. Third, if $n < m$, the dimension of the problem will be smaller. For these reasons, solving the dual problem will be our primary focus.

If the current iterates estimate of the dual variables is denoted by $\boldsymbol{\lambda}^{(i)}$, the dual problem is,

$$\min_{\boldsymbol{d}} q^{(i)}(\boldsymbol{d}) := \boldsymbol{t}^\top \boldsymbol{d} + \boldsymbol{d}^\top Q \boldsymbol{d}/2 \text{ subject to:}$$

$$\boldsymbol{\lambda}^{(i)} + \boldsymbol{d} \geq 0,$$

where $Q := A\hat{B}A^\top$ and $t := \boldsymbol{b} + A\hat{B}\boldsymbol{r}$. The solution of the dual is related to the solution of the primal by the equality $\boldsymbol{p}^\star = -\hat{B}^{-1}(\boldsymbol{r} + A^\top \boldsymbol{d}^\star)$; for more detail, see (Bertsekas, 1997; Nocedal and Wright, 2006). Though we will continue describing the projected gradient method below, note that implementing this optimization algorithm is not actually necessary, since it is already in a form that can be passed to an existing implementation that uses the conjugate gradient method. If this is done, a change of variables in the initial primal objective function can be used so that the Hessian of the dual is $A\hat{B}_N A^\top$. The performance of the implementation in this case is dependent on the condition number of this matrix, and, when it has a bounded condition number, the same results on the number of iterations required by the conjugate gradient method that were discussed in the previous section are applicable in this case as well. Thus, if A is sufficiently sparse, so that products with $A\hat{B}_N A^\top$ can be computed in $O(m \log(m))$ time, and $A\hat{B}_N A^\top$ is well conditioned, this procedure has a time complexity of $O(m \log(m))$. This can also be achieved by using a suitable preconditioner for $A\hat{B}_N A^\top$. Although, this may require implementing this optimization algorithm, since we are not aware of an implementation that supports user defined preconditioners.

Projected Newton methods solve the problem above using two stages in each iteration. First, a projected gradient step is taken to estimate the active set, which is given by

$$\boldsymbol{d}_G^{(i+1)} \leftarrow \max(\boldsymbol{\lambda}^{(i)} - \omega \boldsymbol{t}, 0) - \boldsymbol{\lambda}^{(i)},$$

where ω, is a step length parameter that can be defined using a line search to find a point that satisfies the Armijo condition,

$$l := \min_{k \in \{0,\ldots\}} \{k \mid q^{(i)}(\boldsymbol{d}_G^{(i+1)}) \leq 10^{-3}\omega \boldsymbol{t}\}$$

$$\text{and } \omega \leftarrow (1/2)^l.$$

This step is also used to fix the active set, which is given by $\mathcal{A} \leftarrow \{k \mid \boldsymbol{\lambda}_k^{(i)} + \boldsymbol{d}_G^{(i+1)} = 0\}$.

The next stage of iteration is a Newton step on the free variables, $\mathcal{A}^c$, which can be computed using the conjugate gradient method to avoid factoring Q. Specifically, if we let $Z \in \mathbb{R}^{n \times |\mathcal{A}^c|} := I_{\mathcal{A}^c,\cdot}$, the columns of an $m \times m$ identity matrix with indices in $\mathcal{A}^c$, the Newton step is given by,

$$\boldsymbol{d}_{\mathcal{A}^c}^{(i+1)} \leftarrow Z^\top \boldsymbol{d}_G^{(i+1)} - \tilde{\omega}(Z^\top Q Z)^{-1} Z^\top \left(\boldsymbol{t} + Q\boldsymbol{d}_G^{(i+1)}\right),$$

where $\tilde{\omega}$ is chosen in a similar manner as ω.

In this stage, we also need to ensure that only a single index is either added to or removed from the active, to avoid the problems discussed at the beginning of this section. This is straightforward when using the conjugate gradient method, as it simply involves backtracking to the point in which the constraint was added and then terminating the algorithm.

At this point the method moves onto the next iteration and repeats this process until the iterates converge. A variety of details were omitted from this description in the interest of a concise exposition, such as trust regions and repeating each of these two stages multiple times before proceeding. In particular, Moré and Toraldo (1991) propose continuing the projected gradient step multiple times, which is particularly advantageous in early iterations. A similar strategy is also proposed for the Newton steps, along with effective stopping criteria for each stage to ensure the algorithm switches between these methods efficiently.

Non-Linear Constraints: A Primal-Dual Interior Point Method

It is more difficult to avoid evaluating products with the inverse Hessian in the non-linear constraint case, but there are opportunities in our setting to increase the efficiency when using existing approaches. We will describe these modifications in the context of a primal-dual interior point method, which is a simplified form of the approach implemented by Waltz et al. (2006) in `KNITRO`; see also, (Nocedal and Wright, 2006). Note that a change of variables can be used so that the Hessian of the transformed problem becomes $\hat{B}_N^{-1}$ in order to use `KNITRO` directly; although, in this case, the performance of the method requires that the Hessian of the Lagrangian is well conditioned, which we will describe below in more detail. We implemented the method described here for this reason; we are not aware of a non-linear

optimization method that allows for user-defined preconditioners in a conjugate gradient method.

We will consider the optimization problem,

$$\min_{\boldsymbol{\mu}} \mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2}) \text{ subject to:}$$

$$c(\boldsymbol{\mu}) \geq 0 \text{ and } A\boldsymbol{\mu} = \boldsymbol{b},$$

where $A \in \mathbb{R}^{n_1 \times m}$ and $c : \mathbb{R}^m \to \mathbb{R}^{n_2}$ is a twice differentiable and concave function. The same notes regarding mass normalization constraint also hold in this case; we will still focus on the case in which the Hessian is $\hat{B}^{-1}$.

In the approach presented here, we gradually increase the iterations used in Algorithms 4 and 6. This has the direct advantage of saving convolutions for the later iterations where higher accuracy is required. Also, the values of $\boldsymbol{x}$ resulting from Algorithm 4 have a lower magnitude in earlier iterations, which generally results in smaller changes in $\boldsymbol{\mu}$ and thus increased stability when $\boldsymbol{\mu}$ is far from its optimal value. This could also be done with the use of a trust region, as in Waltz et al. (2006), likely with a gain in robustness, but we attempt to achieve this goal while avoiding additional convolutions.

Primal-dual interior point methods have the advantage of not requiring the initial density to be feasible. To do this, the slack variables, $\boldsymbol{s} \in \mathbb{R}^{n_2}_{++}$, are introduced, so the optimization problem becomes,

$$\min_{\boldsymbol{\mu},\boldsymbol{s}} \mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2}) - t\log(\boldsymbol{s}) \text{ subject to:}$$

$$c(\boldsymbol{\mu}) = \boldsymbol{s},\ A\boldsymbol{\mu} = \boldsymbol{b}, \text{ and } \boldsymbol{s} \geq 0.$$

The Lagrangian corresponding to this optimization problem is,

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{\lambda}, \boldsymbol{\eta}) := \mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_1}) - t\log(\boldsymbol{s}) - \boldsymbol{\lambda}^\top(c(\boldsymbol{\mu}) - \boldsymbol{s}) - \boldsymbol{\eta}^\top\left(A\boldsymbol{\mu} - \boldsymbol{b}\right),$$

and the search directions for the primal and dual variables can be defined using the corresponding KKT system,

$$\begin{bmatrix} \hat{B}^{-1} - C & 0 & A^\top & \nabla c(\boldsymbol{\mu})^\top \\ 0 & D_{\boldsymbol{\lambda} \oslash \boldsymbol{s}} & 0 & -I \\ A & 0 & 0 & 0 \\ \nabla c(\boldsymbol{\mu}) & -I & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{p_\mu} \\ \boldsymbol{p_s} \\ -\boldsymbol{p_\eta} \\ -\boldsymbol{p_\lambda} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{x} - A^\top\boldsymbol{\eta} - \nabla c(\boldsymbol{\mu})^\top\boldsymbol{\lambda} \\ \boldsymbol{\lambda} - t\mathbf{1} \oslash \boldsymbol{s} \\ A\boldsymbol{\mu} - \boldsymbol{b} \\ c(\boldsymbol{\mu}) - \boldsymbol{s} \end{bmatrix}, \tag{3.13}$$

where $C := \sum_i \boldsymbol{\lambda}_i \nabla^2_{\boldsymbol{\mu}} c_i(\boldsymbol{\mu})$. The second block of equalities is a relaxation of the complementary slackness condition; it ensures that the iterates converge to a point that satisfies

$\boldsymbol{\lambda}_i \boldsymbol{s}_i = t$. Thus, we ensure complementary slackness holds at the optimum by decreasing the barrier parameter as the iterates converge. While feasibility of $\boldsymbol{\mu}$ is not constrained to be feasible in the initialization, we do require that $\boldsymbol{s}_i$ and $\boldsymbol{\lambda}_i$ satisfy $\boldsymbol{s}_i > 0$ and $\boldsymbol{\lambda}_i > 0$ for $i \in \{1, 2, ..., n_2\}$, which is maintained throughout all of the iterations.

After the direction of the step is found we can use this requirement as the first step in our line search. Specifically, we define $\alpha_\lambda := \max_{\alpha_\lambda}\{\alpha_\lambda \mid \boldsymbol{p}_{\boldsymbol{\lambda}} + \alpha_\lambda \boldsymbol{\lambda} \geq 0.005\boldsymbol{\lambda}\}$ and $\alpha_s := \max\{\alpha_s \mid \boldsymbol{p}_{\boldsymbol{s}} + \alpha_s \boldsymbol{s} \geq 0.005\boldsymbol{s}\}$. Afterward, we follow the approach of Waltz et al. (2006) and find α using a line search with an Armijo condition, in a similar manner as described in the previous section. This step has the potential to be a significant computational cost because Algorithm 4 must be reevaluated at each point in the line search. However, our approach for limiting the iterations in Algorithms 4 and 6 in early iterations resulted in the largest step possible being taken in all iterations of the numerical experiments presented in the next section. After the line search is complete, the primal and dual variables are updated with the assignments,

$$\boldsymbol{s} \leftarrow \boldsymbol{s} + \alpha\alpha_{\boldsymbol{s}}\boldsymbol{p}_{\boldsymbol{s}}$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \alpha\alpha_{\boldsymbol{s}}\boldsymbol{p}_{\boldsymbol{\mu}}$$

$$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \alpha\alpha_{\boldsymbol{\eta}}\boldsymbol{p}_{\boldsymbol{\eta}}$$

$$\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} + \alpha\alpha_{\boldsymbol{\lambda}}\boldsymbol{p}_{\boldsymbol{\lambda}}.$$

Rather than approximately solving the optimization for fixed values of t, as described by Waltz et al. (2006), we update t in every iteration until it reaches a suitable lower bound. We use the expression that is implemented in `LOQO` by Vanderbei and Shanno (1999), which is given by,

$$t \leftarrow \sigma \boldsymbol{s}^\top \boldsymbol{\lambda}/n_2,$$

where

$$\sigma = 0.1 \min(0.05(1-\xi)/\xi, 2)^3, \text{ and } \xi = \min_i \ \boldsymbol{s}_i\boldsymbol{\lambda}_i / \left(\boldsymbol{s}^\top\boldsymbol{\lambda}/n_2\right).$$

After t is updated, the algorithm moves onto the next iteration. Most of the remaining aspects of the method closely follow the approach provided by Waltz et al. (2006). For example, we use the same termination criteria, the relative error in the l^∞ norm of the KKT residuals, and the same merit function. We also follow the approach they provide for resetting the slack variables. There are aspects of solving the system given in (13) that allow

for a gain in efficiency, and, since this is a significant cost relative the remaining aspects of the algorithm, we will end the section with a discussion of solving this system.

Since $\hat{B}$ is dense, solving (13) directly would not be an efficient approach. The conjugate gradient method also cannot be used directly because (13) is not positive definite. However, this method could be used to solve (13) if we were to eliminate the constrained primal variables beforehand, since the Hessian is positive definite over the null-space of the constraints. Also, this approach would be equivalent to first finding a feasible point and then constraining any subsequent changes in the iterates to the nullspace of the constraints. Keller et al. (2000) described a method known as constraint preconditioning, which can be viewed as applying this strategy. Specifically, in cases in which the derivatives of the nonlinear constraints and A are sparse, the steps described above can be achieved with solutions to the system given by,

$$\begin{bmatrix} D_{\mathbf{1}\oslash\boldsymbol{\mu}} + C & 0 & A^\top & \nabla c(\boldsymbol{\mu})^\top \\ 0 & D_{\boldsymbol{\lambda}\oslash\boldsymbol{s}} & 0 & -I \\ A & 0 & 0 & 0 \\ \nabla c(\boldsymbol{\mu}) & -I & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_{\boldsymbol{\mu},\boldsymbol{s}} \\ \boldsymbol{q}_{\boldsymbol{\eta},\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g} \\ \boldsymbol{b} \end{bmatrix}, \tag{3.14}$$

If we let $\mathcal{P}(\boldsymbol{g}, \boldsymbol{b})$ denote the linear operator defined by the solution to this system, $\mathcal{P}_1(\boldsymbol{g}, \boldsymbol{b})$ as the first block of the solution, $\boldsymbol{q}_{\boldsymbol{\mu},\boldsymbol{s}}$, and $\mathcal{P}_2(\boldsymbol{g}, \boldsymbol{b})$ as the second block, $\boldsymbol{q}_{\boldsymbol{\eta},\boldsymbol{\lambda}}$, then in each iteration $\mathcal{P}_1(\boldsymbol{r}, \mathbf{0})$ corresponds to the projection of $\boldsymbol{r}$ onto the null space of the constraints in our outline of the preconditioned conjugate gradient method above. The remaining modification of the classic conjugate gradient method is that the initial iterate must be feasible. We can use the initialization $\boldsymbol{y} \leftarrow \mathcal{P}_1(\boldsymbol{g}, \boldsymbol{b})$ for any $\boldsymbol{g} \in \mathbb{R}^{m+n_2}$ when using a constraint preconditioner. For more detail on constraint preconditioners, see also (Nocedal and Wright, 2006).

Though, the method does require a decomposition of the matrix in (14), this is at least less costly than decomposing the dense matrix $\hat{B}$, and there are also specialized methods for constraint preconditioning KKT systems that can be used. Our implementation used the `MA57` solver from HAL (Duff, 2004); timing of this decomposition will be discussed further in the next section. Also, an additional benefit of having $\mathcal{P}(\boldsymbol{g}, \boldsymbol{b})$ is that it can also be used after the algorithm terminates to define the dual variables, $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$.

Note that the convergence behavior of the conjugate gradient method when using constraint preconditioners depends on how well the matrix in (14) preconditions the matrix in (13). However, Keller et al. (2000) showed that the rate of convergence is only dependent on the spectrum of,

$$(D_{\mathbf{1}\oslash\boldsymbol{\mu}} - C)^{-1} \left(\hat{B}^{-1} - C \right)$$

because all of the other $m + n_2 + n_1$ eigenvalues in the system are equal to one after preconditioning. This is a particularly attractive property because inside of this constraint preconditioned conjugate gradient method, there is also an inner loop, the conjugate gradient method used to compute products with $\hat{B}^{-1}$, defined in Algorithm 6. For this reason, we limit the number of iterations in the outer loop to $2 + i$, where i is the iteration number of the optimization algorithm. Even in the early iterations, this strategy produces search directions that are rarely rejected in practice, which can also be explained by a bound on the error of the k^{th} iterate of the conjugate gradient method, $\boldsymbol{y_k}$, given by Luenberger (1973),

$$\|\boldsymbol{y_k} - \boldsymbol{y}^\star\|_A \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \|\boldsymbol{y_0} - \boldsymbol{y}^\star\|_A ,$$

where A is the inverse of the matrix in (14), multiplied by the matrix in (13), λ_i denotes the i^{th} largest eigenvalue of A, and $\boldsymbol{y}^\star$ is the Newton step. Since $D_{\boldsymbol{\mu}}\hat{B}^{-1}$ has most of its eigenvalues near one, C is an ideal preconditioner for itself, and the remaining $m + n_2 + n_1$ of A are one, it is not too surprising that a few iterations suffice to provide a significant improvement over gradient descent. However, these iterations are still a significant proportion of the computational cost of the method, which will be described in the next section in more detail.

Using a similar notation as Algorithm 6, the outer conjugate gradient method is also terminated when $\left\|\boldsymbol{s_k}^\top \boldsymbol{r_k}\right\|_2 \leq \epsilon(1 + \left\|\boldsymbol{s_0}^\top \boldsymbol{r_0}\right\|_2)$, where $\epsilon \leftarrow \max(10^{-4-i/10}, 10^{-9})$ and i denotes the iteration of the optimization method. Given the low value of the maximum iterations, this tolerance level is sometimes only achieved around the 15^{th} to 20^{th} iteration. In contrast, the inner conjugate gradient method, Algorithm 6, is terminated only when it reaches a tolerance of $\epsilon/100$. This high tolerance is meant to avoid the residuals in the outer conjugate gradient method losing orthogonality. In addition, Algorithm 4 terminates when $\|\boldsymbol{\mu_2} - \tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}\|_\infty \leq \tilde{\epsilon}$, where $\tilde{\epsilon} \leftarrow \max(10^{-3-i/10}, 10^{-7})$. These choices are meant to avoid convolutions in the early iterations, while ensuring that the objective function, derivatives, and Newton steps are sufficiently accurate by the twentieth iteration. Since the objective function is actually dependent on the number of iterations used in Algorithm 4, we avoid changing the number of iterations by only checking for convergence every thirty iterations.

In the next section we will discuss the computational performance of the method with an application related to estimating the density function of a random variable. This application will also be used to describe how the method described in this section can be used to solve optimization problems of the form,

$$\min_{\boldsymbol{\mu}} \mathcal{W}_\gamma^2(f(\boldsymbol{\mu}), \boldsymbol{\mu_2}) \text{ subject to:}$$

$$A\boldsymbol{\mu} \geq \boldsymbol{b},$$

where $f : \mathbb{R}^m \to \mathbb{R}^m$ is convex and differentiable.

3.5 Numerical Experiments

This application constructs an estimator of the density function, $\mu^\star : \Omega \to \in \mathbb{R}^1_{++}$ with $\Omega := [0,1]^2$, using a dataset consisting of n iid observations, $\{(y_i, z_i)\}_{i=1}^n$, with $y_i, z_i \in \mathbb{R}^1$ for all i, drawn from $\mu^\star(y, z)$. Specifically, we will describe a shape constrained density estimator. In terms of the assumptions on the functional form of $\mu^\star(y, z)$, these estimators can be viewed as a middle ground between fully nonparametric estimators and fully parametric estimators. In cases in which the true population density satisfies the shape constraint, these density estimators often compare favorably to their fully nonparametric counterparts; see for example, (Cule et al., 2010; Koenker and Mizera, 2010).

To do this, we will define $\mu_2(y, z)$ as kernel density estimator, with bandwidth $\sigma \in \mathbb{R}^1_{++}$, and then find the density, $\mu(y, z)$, that minimizes the regularized Wasserstein metric from $\mu_2(y, z)$, such that for all $z \in [0, 1]$, the function $y \mapsto \log(\mu(y \mid z))$ is concave, where $\mu(y \mid z) := \mu(y, z) / \left(\int \mu(y, z) dz\right)$ is the density estimate of y conditional on z. This is known as $\log-$concave shape constraint, and this particular form of the constraint could be used to estimate the most typical value of y conditional on z, or the mode the conditional density $z \mapsto \mu^\star(y \mid z)$. For related work on estimates of the conditional mode, which is also known as modal regressions; see for example, (Lee, 1989; Yao et al., 2012; Chen et al., 2016).

After discretizing on a uniform mesh in $[0, 1] \times [0, 1]$, and letting $\{\boldsymbol{\mu_2}_{ij}\}_{i,j=1}^{\sqrt{m}}$ denote the values of the kernel density estimator at these points, we can define the estimator by solving,

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^m_{++}} \mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2}) \text{ subject to:}$$

$$\log(\boldsymbol{\mu}_{i,j}) \geq (\log(\boldsymbol{\mu}_{i-1,j}) + \log(\boldsymbol{\mu}_{i+1,j}))/2 \text{ for } i \in \{2, 3, ..., \sqrt{m} - 1\},\ j \in \{1, 2, ..., \sqrt{m}\} \text{ and}$$

$$\mathbf{1}^\top \boldsymbol{\mu} = 1.$$

Unfortunately this is not a convex optimization problem. Before outlining how we transform this problem into a convex optimization problem, we will describe a third option to ensure mass normalization. Note that mass normalization, as well as the non-uniqueness of both of the derivatives of $\mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2})$, is simply a consequence of not accounting for mass normalization in the optimization problem defining $\mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2})$. For example, for fixed $(l, k) \in \{(i, j)\}_{i,j=1}^{\sqrt{m}}$, $\mathcal{W}_\gamma^2(\boldsymbol{\mu}, \boldsymbol{\mu_2})$ can be written as,

$$\max_{(x,y) \in \mathbb{R}^{2m}} \left(\sum_{(i,j) \neq (l,k)} \boldsymbol{x}_{ij}^\top \boldsymbol{\mu_1}_{ij} + \boldsymbol{x}_{lk}(1 - \boldsymbol{\mu_1}_{ij}) \right) +$$

$$\sum_{i,j} \boldsymbol{y}_{ij}\boldsymbol{\mu_2}_{ij} - \gamma \exp\left((\boldsymbol{x}_{ij} + \boldsymbol{y}_{ij} - M_{ij})/\gamma\right),$$

which has the effect of defining $\boldsymbol{\mu_1}_{lk}$ with the mass normalization constraint. In a slight abuse of notation, we will let $\mathcal{W}_\gamma^2(\exp(\boldsymbol{\nu}), \boldsymbol{\mu_2})$, with $\boldsymbol{\nu} \in \mathbb{R}^{m-1}$, denote the regularized Wasserstein distance between $\boldsymbol{\mu}$ and $\boldsymbol{\mu_2}$, where $\boldsymbol{\mu}_{ij}$ is defined as $\exp(\boldsymbol{\nu}_{ij})$ for all $(i,j) \neq (l,k)$ and $\boldsymbol{\mu}_{lk} := 1 - \sum_{(i,j)\neq(l,k)} \exp(\boldsymbol{\nu}_{ij})$. We will also choose (l,k) to be $(l,k) = \arg\min_{(i,j)} \boldsymbol{x}_{ij}$, for reasons that we will describe below.

Eliminating the mass normalization constraint in this way is our recommended approach when the density is defined as a non-linear function of the arguments being optimized. After adding bound constraints to improve the robustness of the method, the final form of the optimization problem is,

$$\min_{\boldsymbol{\nu}} \mathcal{W}_\gamma^2(\exp(\boldsymbol{\nu}), \boldsymbol{\mu_2}) \text{ subject to:}$$

$$\text{For all } i \in \{2, 3, ..., \sqrt{m} - 1\},\ j \in \{1, 2, ..., \sqrt{m}\}, \text{ and } (i,j) \neq k:$$

$$2\boldsymbol{\nu}_{ij} \geq \boldsymbol{\nu}_{i-1,j} + \boldsymbol{\nu}_{i+1,j}) \text{ and}$$

$$\boldsymbol{\nu}_{i,j} \geq \log(5 \cdot 10^{-6}).$$

Note that if we express $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = f(\boldsymbol{\nu})$, then we can write the Hessian of $\mathcal{W}_\gamma^2(f(\boldsymbol{\nu}), \boldsymbol{\mu_2})$ in the general form,

$$(\nabla_{\boldsymbol{\nu}} f)^\top \hat{B}_N^{-1} (\nabla_{\boldsymbol{\nu}} f) + \sum_i \boldsymbol{x}_i \nabla_{\boldsymbol{\nu}}^2 f_i,$$

and thus, in the absence of closed form expression of the inverse of this matrix, we must perform Algorithm 6 inside of another conjugate gradient method, as in the last section. In our case the Hessian is given by

$$E^\top D_{\boldsymbol{\mu}} \left(\hat{B}^{-1} + D_{\boldsymbol{x} \oslash \boldsymbol{\mu}}\right) D_{\boldsymbol{\mu}} E,$$

where, for $\tilde{k}$ defined as the index corresponding to the subscript (l,k), E is defined by binding together the first $\tilde{k} - 1$ rows of an $m-1 \times m-1$ identity matrix, the row $-\mathbf{1}_{m-1}^\top$, and then the last $m - \tilde{k}$ rows of an $m-1 \times m-1$ identity matrix. For clarity, if $k = (2,1)$, so that $\tilde{k} = 2$ if we vectorize columnwise, then $E \in \mathbb{R}^{m \times m-1}$ is given by,

$$E := \begin{bmatrix} 1 & & & \\ -1 & -1 & -1 & \cdots \\ & 1 & & \\ & & \ddots & \end{bmatrix}.$$

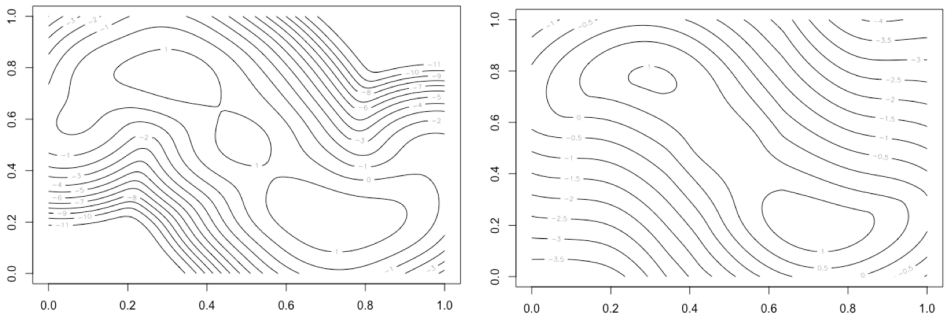


Figure 10: The natural logarithm of the input density and the shape constrained density estimator are shown on the left and right respectively for $m = 125^2$ and $\gamma = 0.01$.

Note that the term $E^\top D_{\boldsymbol{x}\otimes\boldsymbol{\mu}} E$ of the Hessian can also be written as,

$$E^\top D_{\boldsymbol{x}\otimes\boldsymbol{\mu}} E = D_{\exp(\boldsymbol{\nu})\otimes(\boldsymbol{x}_{-(l,k)} - \boldsymbol{x}_{lk})},$$

so choosing (l, k) to be $(l, k) = \arg\min_{(i,j)} \boldsymbol{x}_{ij}$ ensures the optimization problem is convex.

A gain in robustness and generality can be achieved by considering the direction of the Newton step in all m dimensions. Given that the definition of the gradient is $D_{\boldsymbol{\mu}\otimes\boldsymbol{x}} E$, this step can be defined as $\boldsymbol{p}_{\tilde{\boldsymbol{\nu}}} = E\boldsymbol{p}_{\boldsymbol{\nu}} \in \mathbb{R}^m$. In other words, the Newton step in part accounts for $\boldsymbol{p}_{\tilde{\boldsymbol{\nu}}lk}$ using the overly local approximation, $\mathbf{1}_m^\top \boldsymbol{p}_{\tilde{\boldsymbol{\nu}}} = 0$. The accuracy of this prediction can be improved by using this equality directly, so that the update of $\boldsymbol{\mu}_{lk}$ becomes,

$$\boldsymbol{\mu}_{lk}^+ \leftarrow \boldsymbol{\mu}_{lk} \exp(-\textstyle\sum_{ij} \boldsymbol{p}_{\boldsymbol{\nu}ij}).$$

Afterward, $\boldsymbol{\mu}$ can be renormalized to ensure mass normalization, which can be viewed as averaging out the error used in this local approximation over all m dimensions. Note that an alternative approach is to optimize over all m dimensions, with the addition of the constraint $\mathbf{1}_m^\top \boldsymbol{p}_{\tilde{\boldsymbol{\nu}}} = 0$, which has the additional advantage of allowing for the shape constraint to be imposed over all points in the mesh.

To provide an idea of the performance of the methods proposed here, we solved the optimization problem given above with $m \in \{25^2, 75^2, 125^2\}$ and $\gamma \in \{0.01, 0.02\}$ at a tolerance of 10^{-6}. We drew a random dataset of 100 points and used this same dataset in each case. The $\{x_i\}_i$ was drawn from a $Beta(1/2, 1/2)$ distribution and then each $y_i \in \{y_i\}_i$ was defined as $y_i = 1/2 + \sin(2x_i)/3 + \epsilon_i$, where $\epsilon \sim N(0, 1/15)$. The bandwidth used to define $\boldsymbol{\mu_2}$ was set equal to $\sqrt{\gamma/2}$ in each case. Note that this results in the global unconstrained minimum being equal to a kernel density estimator with bandwidth $\sqrt{\gamma}$. The resulting density estimator when $m = 125^2$ and $\gamma = 0.01$ is provided in Figure 10.

We also provide the number of convolutions and iterations at each combination of these values of m and γ in Table 7. Note that the number of matrix factorizations required is equal

to the number of iterations; although, this cost was actually less significant, in relative terms, to the cost of convolutions. For example, for these values of m, the solver `MA57` required 0.006, 0.08, and 0.44 seconds to factor the constraint preconditioners with dimensions of the order $5m$, which appears to scale at a similar rate as our implementation of Algorithm 4, and is approximately 4.5 and 1.6 times faster on average with a tolerance of 10^{-6} when $\gamma = 0.01$ and $\gamma = 0.02$ respectively.

Given the simplicity of the optimization algorithm described here, it is likely possible to decrease the number of iterations in a few cases. For example, note that the case in which $m = 25^2$ and $\gamma = 0.01$ required significantly more iterations, and thus also convolutions, than the rest of the cases. This is likely a result of inefficiencies in the simplified algorithm implemented here, since the barrier parameter reached its lower bound after only nine iterations in this case. For this reason, the primary value of the table is to provide an idea of the relative costs of Algorithm 4, and of Algorithm 6 inside of constraint preconditioned conjugate gradient method. As mentioned above, it is worth noting that the results are dependent on whether linear or non-linear constraints are implemented, as the linear constraint case generally requires fewer evaluations of Algorithms 5 and 6.

One interesting feature in Table 4 is that the proportion of convolutions evaluated in Algorithm 5 to the total convolutions evaluated is fairly stable over all values of m and γ. Specifically, it ranges from 0.63 to 0.49, which correspond to the $m = 25^2$ and $\gamma = 0.01$ case and the $m = 125$ and $\gamma = 0.01$ case respectively. This proportion also generally scales well as m is increased or γ is decreased.

In each of the examples shown, when averaged over all iterations, the additional computational costs arising from convolutions is no more than 1.71 times what it would be in a method that did not require Algorithm 6. If this were the only additional cost, this second order method would compare favorably to its first order counterparts, since far fewer iterations are required. However, the method also requires a factorization of (14) in each iteration. The computational cost of these operations are application dependent, but, at least for these cases, in which the constraint matrix is highly sparse, these decompositions add less than 16 seconds to the total time when using an ordinary laptop. Thus, even in the context of nonlinear constraints, we expect the efficiency of this approach in similar cases will compare favorably with first order methods. Of course, an additional advantage is local quadratic convergence, which allows the regularized Wasserstein metric to be applicable in settings that require a high degree of accuracy.

Convolutions Evaluated			
	$\gamma = 0.01$		
m	25^2	75^2	125^2
N_1	$41,460$	$29,280$	$21,540$
N_2	$70,846$ (0.63)	$35,474$ (0.55)	$20,960$ (0.49)
Iterations	61	45	35
	$\gamma = 0.02$		
m :	25^2	75^2	125^2
N_1	$6,600$	$4,080$	$5,760$
N_2	$10,930$ (0.62)	$6,386$ (0.61)	$9,034$ (0.61)
Iterations	31	23	27

Table 7: The number of convolutions evaluated in Algorithms 4 and 5 are given by N_1 and N_2 respectively. The values in parenthesis corresponds to $N_2/(N_1 + N_2)$, the proportion of the total convolutions that were required to evaluate products with $\hat{B}$ and $\hat{B}^{-1}$. The number of iterations is also equal to the number of matrix decompositions.

3.6 Discussion

This paper proposes methods for evaluating products with the Hessian and its inverse to allow for the use of efficient second order optimization algorithms to minimize the regularized Wasserstein metric. When $\mathcal{A} \subset \mathbb{R}^d$, we find bounds on the finite condition number of the Hessian, and a direct implication is that the conjugate gradient method can be used to find products with the Hessian in a bounded number of iterations. We also show the required number of iterations is generally small in practice. Thus, this method, when combined with replacing products with approximations of convolutions, allows for the evaluation of products with the Hessian in $O(m \log(m))$ time. We also provide a method to evaluate products with the inverse or pseudoinverse of the Hessian that only requires two convolutions and elementwise operations of a vectors in $\mathbb{R}^m$.

Also, we discuss existing implementations of optimization algorithms that can use these results when minimizing the regularized Wasserstein metric subject to either linear or nonlinear constraints. These implementations can be used to find solutions in an efficient manner and with a high level of accuracy. In the linear constraint case, solving the dual problem provides a further gain in computational efficiency, since this has the effect of avoiding matrix factorizations and substituting some products with the Hessian for products with the inverse or pseudoinverse. In favorable circumstances, meaning when the constraint matrix, A, has a number of nonzero elements that is $O(m)$ and $AA^\top$ has a condition number that is bounded asymptotically, or a suitable preconditioner is available for $ABA^\top$, the time complexity of the optimization problems in the linear constraint case is provably $O(m \log(m))$.

When implementing nonlinear constraints, it is generally more difficult to avoid matrix factorizations and products with the Hessian; although, the efficiency can be improved further with the use of preconditioner for the Hessian of the Lagrangian. After discussing these topics, we provide numerical experiments. In the cases we consider the number of additional convolutions required by the method scales well as γ goes to zero and as m diverges. We hope that the approaches provided here will allow second order methods to be applicable in the large scale problems that are commonly encountered in the regularized Wasserstein metric setting.

References

Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review*, 568-598.

Altschuler, J., Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems* (pp. 1964-1974).

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 193-212.

Andoni, A., & Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006.*

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica, 68*(2), 399-405.

Bagnoli, M., & Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory, 26*(2), 445-469.

Bajari, P., Houghton, S., & Tadelis, S. (2006). Bidding for incomplete contracts: An empirical analysis. *National Bureau of Economic Research.*

Baldauf, M., & Silva, J. S. (2012). On the use of robust regression in econometrics. *Economics Letters, 114*(1), 124-127.

Bauschke, H. H., Borwein, J. M., & Combettes, P. L. (2003). Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization, 42*(2), 596-636.

Bauschke, H. H., & Lewis, A. S. (2000). Dykstras algorithm with bregman projections: A convergence proof. *Optimization, 48*(4), 409-427.

Beardwood, J., Halton, J. H., & Hammersley, J. M. (1959). The shortest path through many points. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 55, No. 04, pp. 299-327). Cambridge University Press.

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques

for embedding and clustering. *Advances in neural information processing systems* (pp. 585-591).

Benamou, J. D., Carlier, G., Cuturi, M., Nenna, L., & Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing, 37*(2), A1111-A1138.

Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology).

Bertsekas, D. P. (1982). Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on control and Optimization, 20*(2), 221-246.

Bertsekas, D. P. (1997). Nonlinear programming. J*ournal of the Operational Research Society, 48*(3), 334-334.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics, 7*(3), 200-217.

Byrd, R. H., Nocedal, J., & Waltz, R. A. (2006). Knitro: An integrated package for nonlinear optimization. In *Large-scale nonlinear optimization*. Springer, Boston, MA.

Cardoso, A. R., & Portugal, P. (2005). Contractual wages and the wage cushion under different bargaining settings. *Journal of Labor Economics, 23*(4), 875-902.

Carlier, G., Oberman, A., & Oudet, E. (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis, 49*(6), 1621-1642.

Carroll, R. J., Delaigle, A., & Hall, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association, 106*(493), 191-202.

Chen, Y. C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics, 10*(4), e1431.

Chen, Y. C., Ho, S., Freeman, P. E., Genovese, C. R., & Wasserman, L. (2015). Cosmic web reconstruction through density ridges: method and algorithm. *Monthly Notices of the Royal Astronomical Society, 454*(1), 1140-1156.

Chen, Y. C., Genovese, C. R., Tibshirani, R. J., & Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics, 44*(2), 489-514.

Chen, Y.C., & Samworth, R. J. (2013). Smoothed log −concave maximum likelihood estimation with applications. *Statistica Sinica*, 1373-1398.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1), 31-41.

Conn, A. R., Gould, G. I. M., & Toint, P. L. (2013). *LANCELOT: a Fortran package for large-scale nonlinear optimization*. Springer Science & Business Media.

Cowlagi, R. V., & Tsiotras, P. (2009). Shortest distance problems in graphs using history-dependent transition costs with application to kinodynamic path planning. In *American Control Conference, 2009. ACC'09*. IEEE.

Cule, M., Samworth, R., & Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(5), 545-607.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 2292-2300.

Cuturi, M., & Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning* (pp. 685-693).

Cuturi, M., & Peyré, G. (2016). A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences, 9*(1), 320-343.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*(1), 269-271.

Doss, C. R., & Wellner, J. A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *The Annals of Statistics, 44*(3), 954-981.

Duff, I. S. (2004). MA57---a code for the solution of sparse symmetric definite and indefinite systems. *ACM Transactions on Mathematical Software (TOMS), 30*(2), 118-144.

Dümbgen, L., & Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli, 15*(1), 40-68.

Dümbgen, L., Samworth, R., & Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *The Annals of Statistics, 39*(2), 702-730.

Einbeck, J., & Tutz, G. (2006). Modeling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society, 55*(4), 461-475.

Einmahl, U., & Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics, 33*(3), 1380-1403.

Evans, L. C. (2010). *Partial Differential Equations.* American Mathematical Society.

Ewerhart, C. (2013). Regular type distributions in mechanism design and $\rho-$concavity. *Economic Theory, 53*(3), 591-603.

Fan, J. Y., & Yuan, Y. X. (2001). A new trust region algorithm with trust region radius converging to zero. In *Proceeding of the 5th International Conference on Optimization: Techiniques and Applications* (pp. 786-794). ICOTA, Hong Kong.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Geiser, Saul, and Maria Veronica Santelices. Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High- School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. *Center for Studies in Higher Education* (2007).

Getreuer, P. (2013). A survey of Gaussian convolution algorithms. *Image Processing On Line*, 2013, 286-310.

Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal, 1956*(2), 125-153.

Guerre, E., Perrigne, I., & Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica, 68*(3), 525-574.

Hickman, B. R., & Hubbard, T. P. (2015). Replacing sample trimming with boundary correction in nonparametric estimation of first-price auctions. *Journal of Applied Econometrics, 30*(5), 739-762.

Ho, C., Damien, P., & Walker, S. (2017). Bayesian mode regression using mixtures of triangular densities. *Journal of Econometrics, 197*(2), 273-283.

Hoffleit, E. D., & Warren Jr, W. H. (1991). Yale Bright Star Catalog. *New Haven: Yale Univ. Obs.*

Horn, R. A., & Johnson, C. R. (1990). *Matrix analysis*. Cambridge university press.

Horowitz, J. L., & Lee, S. (2017). Nonparametric estimation and inference under shape restrictions. *Journal of Econometrics, 201*(1), 108-126.

Howard, C. D., & Newman, C. M. (1997). Euclidean models of first-passage percolation. *Probability Theory and Related Fields, 108*(2), 153-170.

Howard, C. D., & Newman, C. M. (2001). Geodesics and spanning trees for Euclidean first-passage percolation. *Annals of Probability*, 577-623.

Hwang, S. J., Damelin, S. B., & Hero III, A. O. (2016). Shortest path through random points. *The Annals of Applied Probability, 26*(5), 2791-2823.

Kantorovitch, L. (1958). On the translocation of masses. *Management Science, 5*(1), 1-4.

Karlin, S. (1968). *Total positivity*. Stanford: Stanford University Press.

Karunamuni, R. J., & Mehra, K. L. (1991). Optimal convergence properties of kernel density estimators without differentiability conditions. *Annals of the Institute of Statistical Mathematics, 43*(2), 327-346.

Keller, C., Gould, N. I., & Wathen, A. J. (2000). Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1300-1317.

Kemp, G. C., & Silva, J. S. (2012). Regression towards the mode. *Journal of Econometrics, 170*(1), 92-101.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887-906.

Kim, A. K., & Samworth, R. J. (2016). Global rates of convergence in log −concave density estimation. *The Annals of Statistics, 44*(6), 2756-2779.

Koenker, R., & Mizera, I. (2010). Quasi-concave density estimation. *The Annals of Statistics*, 2998-3027.

Krupp, R. S. (1979). Properties of Kruithof's projection method. *Bell Labs Technical Journal, 58*(2), 517-538.

Laffont, J. J., Ossard, H., & Vuong, Q. (1995). Econometrics of first-price auctions. *Econometrica: Journal of the Econometric Society*, 953-980.

Lee, M. J. (1989). Mode regression. *Journal of Econometrics, 42*(3), 337-349.

Lee, M. J. (1993). Quadratic mode regression. *Journal of Econometrics, 57*(1-3), 1-19.

Lin, C. J., & Moré, J. J. (1999). Newton's method for large bound-constrained optimization problems. *SIAM Journal on Optimization, 9*(4), 1100-1127.

Luenberger, D. G. (1973). *Introduction to linear and nonlinear programming.*

Macdonell, W. R. (1902). On criminal anthropometry and the identification of criminals. *Biometrika, 1*(2), 177-227.

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 508-515.

McAfee, R. P., & McMillan, J. (1987). Auctions and bidding. *Journal of economic literature, 25*(2), 699-738.

Moré, J. J., & Toraldo, G. (1991). On the solution of large quadratic programming problems with bound constraints. *SIAM Journal on Optimization, 1*(1), 93-113.

Mumford, D., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics, 42*(5), 577-685.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research, 6*(1), 58-73.

Myerson, R. B., & Satterthwaite, M. A. (1983). Efficient mechanisms for bilateral trading. *Journal of economic theory, 29*(2), 265-281.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Ouellette, D. V. (1981). Schur complements and statistics. *Linear Algebra and its Applications, (36)*, 187-295.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065-1076.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical proceedings of the Cambridge philosophical society* (Vol. 51, No. 3, pp. 406-413). Cambridge University Press.

Peyré, G., & Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning, 11*(5-6), 355-607.

Rothstein, Jesse M. "College performance predictions and the SAT." *Journal of Econometrics*. (2004): 297-317.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician, 53*(4), 382-387.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (Vol. 82). SIAM.

Schmitz, P. G., & Ying, L. (2012). A fast direct solver for elliptic problems on general meshes in 2D. *Journal of Computational Physics, 231*(4), 1314-1338.

Scott, D. W. (1992). *Multivariate Density Estimation*. Wiley.

Seregin, A., & Wellner, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Annals of statistics, 38*(6), 3751.

Shawe-Taylor, J., Williams, C. K., Cristianini, N., & Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory, 51*(7), 2510-2522.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society,* 683-690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC press.

Sinkhorn, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly, 74*(4), 402-405.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions.

The annals of mathematical statistics, 19(2), 279-281.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., & Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG), 34*(4), 66.

Spielman, D. A. (2010). Algorithms, graph theory, and linear equations in Laplacian matrices. *Proceedings of the International Congress of Mathematicians* (pp. 2698-2722).

Steihaug, T. (1983). The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis, 20*(3), 626-637.

Student. (1908). The probable error of a mean. *Biometrika,* 1-25.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319-2323.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Vanderbei, R. J., & Shanno, D. F. (1999). An interior-point algorithm for nonconvex nonlinear programming. *Computational Optimization and Applications, 13*(1-3), 231-252.

Varadhan, S. R. S. (1967). On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics, 20*(2), 431-455.

Venter, J. H. (1967). On estimation of the mode. *The Annals of Mathematical Statistics,* 1446-1455.

Villani, C. (2003). *Topics in optimal transportation* (No. 58). American Mathematical Society.

Waltz, R. A., Morales, J. L., Nocedal, J., & Orban, D. (2006). An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical programming*, 107(3), 391-408.

Wellner, J. A. & Laha, N. (2017). Bi-s-concave distributions. *arXiv preprint.*

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT Press.

Yao, W., Lindsay, B. G., & Li, R. (2012). Local modal regression. *Journal of nonparametric statistics, 24*(3), 647-663.

Zhou, H., & Huang, X. (2018). Bandwidth selection for nonparametric modal regression. *Communications in Statistics-Simulation and Computation*, 1-17.