Shape-Constrained Density Estimation Via Optimal Transport

Ryan Cumings*

October 26, 2017

Abstract

Optimal transportation is used to define a nonparametric density estimator as the solution of a convex optimization problem. The framework allows for density estimation subject to a variety of shape constraints, including \( \rho \)-concavity and Myerson's (1981) regularity condition. The mean integrated squared error for the density estimator of a random variable in \( \mathbb{R}^d \) achieves an asymptotic rate of convergence of \( O_p(N^{-4/(d+4)}) \). After deriving algorithms for finding the density estimate, the framework is applied to data from the California Department of Transportation to explore whether their choice of awarding construction contracts using a first price auction is cost minimizing.

JEL Classification: C14

Keywords: Nonparametric density estimation, Kernel density estimation, Optimal transport, Log-concavity, \( \rho \)-concavity, s-concavity

1 Introduction

Nonparametric density estimation has the distinct 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) showed that we can extricate ourselves from this “Dirac catastrophe” by constraining the density to be monotonic.

A great deal of the progress was made in the subsequent decades by adding penalty terms to the maximum likelihood objective function to restore the parsimony of the

*Department of Economics, University of Illinois at Urbana-Champaign, Address: 214 David Kinley Hall, 1407 West Gregory Drive, Urbana, IL 61801, USA. E-mail address: cumings2@illinois.edu.
density estimator; for example, see (Silverman, 1982). Parzen (1962) also showed that kernel density estimators resulted in consistent density estimators and derived the rates of convergence, and a large literature developed around this topic. Unlike Grenander’s (1956) approach, the performance of maximum penalized likelihood estimators and kernel density estimators is highly dependent the specification of penalty terms and bandwidths, respectively, which can be difficult to choose.

Recently there has been a renewed interest in ensuring the parsimony of the maximum likelihood density estimate through conditioning on the information provided by the shape of the underlying density. For example, Cule, Samworth, and Stewart (2010) studied maximum likelihood density estimation subject to a constraint that the logarithm of the density was concave, which is known as a log-concavity constraint. Bagnoli and Bergstrom (2005) showed that log-concavity implies various functions of the density are monotonic and that these properties can be used to prove the existence and uniqueness of equilibria in models throughout economics. One early pioneer on the advantages of log-concavity for both statistical testing as well as estimation was Karlin (1968). Some examples of log-concavity’s many implications include, products and convolutions between log-concave densities are log-concave, the density $f(x - \theta)$ has a monotonic likelihood ratio if and only if $f(\cdot)$ is log-concave, and the hazard function of the log-concave density $f(x)$, defined by $f(x)/(1 - F(x))$, where $F(x)$ is the density function of $f(x)$, is increasing.

Koenker and Mizera (2010) extended these methods by maximizing Rényi entropy of order $\rho \in \mathbb{R}$ subject to a shape constraint of the form

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

for all $\alpha \in [0, 1]$, which is equivalent to maximum likelihood subject to a log-concavity constraint as $\rho \to 0$. Decreasing $\rho$ corresponds to a relaxation of the shape constraint. In other words, if $f(x)$ satisfies the constraint for some $\rho$, then it also satisfies the constraint for all $\rho' < \rho$. This constraint is equivalent to concavity when $\rho$ 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.

In this paper we use optimal transportation to formulate a density estimator that is the solution of a convex optimization problem subject to the same $\rho$-concavity constraint given above. Galichon (2016) provides a comprehensive overview of the optimal transport literature as well as its many applications in economics. The next section briefly surveys the parts of this literature that are required to formulate our estimator, and the third section defines the estimator and shows that the asymptotic rate of convergence of the estimator is equivalent to the rate of convergence of kernel

\footnote{Maximum likelihood is equivalent to maximizing Shanon entropy, and Shanon entropy is equivalent to the limit of Rényi entropy of order $\rho$ as $\rho \to 0$. Maximum likelihood does not provide a convex optimization problem for values of $\rho < 0$, and the objective function is not finite when $\rho < -1$ (Doss and Wellner, 2016).}
density estimators. The fourth and fifth sections provide algorithms to solve for the estimator.

After exploring the estimator resulting from conditioning on a $\rho$–concavity constraint, the sixth section generalizes the framework presented here to allow for density estimation subject to a much larger class of shape constraints, with a particular focus on shape constraints that arise in the economics literature. For example, Myerson (1981) proves the revenue equivalence theorem using a regularity condition that is now common in mechanism design. Myerson defined the virtual valuations as $x - (1 - F(x))/f(x)$, where $f(x)$ is the density function of the agents’s valuations and $F(x)$ is its distribution function, and requires this function be strictly increasing. The seventh section applies this generalization to data on bids for construction contracts from the California Department of Transportation to explore whether or not their choice of a first price auction to award construction contracts is cost minimizing. We find that a kernel density estimator of the valuations does not satisfy this condition. However, the shape-constrained density estimator follows the data closely, and this view is corroborated with an Anderson-Darling test. The final section concludes the paper.

In the subsequent sections, a few notational conventions are followed that are worth noting here. The application of a function to a matrix or vector refers to the application of the function to each element of that matrix or vector. Throughout the paper, all vectors are expressed as column vectors. Also, a diagonal matrix with a diagonal equal to the vector $x$ will be denoted by $D_x$. An $n \times 1$ vector of ones will be denoted by $1_n$, and we will also use $\odot$ for element-wise division of two vectors. Lastly, $\text{sgn}(x)$ will be used to denote a function that is $1$ when $x \geq 0$ and $-1$ when $x < 0$.

2 Optimal Transportation

The theory of optimal transportation began with Monge in the 18th century. Given a mound of earth and a nearby hole, both with the same volume, Monge explored how to efficiently transport the earth from the mound to the hole. For example, if we model our mound of earth with the density $\mu_0 : X \to \mathbb{R}^+_+$ and the hole using another density, $-\mu_1 : Y \to \mathbb{R}^-$, Monge sought to find a transportation plan, $T : X \to Y$, that minimizes a function describing the transportation cost while ensuring that all mass from $\mu_0$ is transported and that the hole described by $-\mu_1$ is completely filled. This problem does not actually have a solution for general densities $\mu_0$ and $\mu_1$ due to the fact that the transportation plans are required to be functions; in other words, all the mass at each $x \in X$ must be transported to a single location $y \in Y$.

Kantorovich (1942) showed how to find a solution to a more general problem by writing the transportation plan as another density, say $\pi : X \times Y \to \mathbb{R}^+_+$. Given $x \in X$ and $y \in Y$, $\pi(x, y)$ can be interpreted as the amount of earth (or mass) moved from $x$
to \( y \) under the transportation plan \( \pi(\cdot) \). Obviously after moving all the earth, the hole must be filled completely, which implies \( \int_X \pi(x, y) \delta x = \mu_1(y) \) for all \( y \in Y \). Likewise, the entire mound, \( \mu_0 \), is used to fill this hole, so \( \int_Y \pi(x, y) \delta y = \mu_0(x) \) for all \( x \in X \). The set of densities, or couplings, that satisfy these two constraints will be denoted by \( \Pi(\mu_0, \mu_1) \).

The most common cost function in optimal transport is simply squared Euclidean distance. In other words, the cost of moving one unit of earth from \( x \in X \) to \( y \in Y \) is proportional to \( \|x - y\|^2 \). The resulting minimization problem is then given by

\[
\min_{\pi \in \Pi(\mu_0, \mu_1)} \int_{X \times Y} \pi(x, y) \|x - y\|^2 \delta x \delta y,
\]

which we will call the Wasserstein distance. The Wasserstein distance has many desirable properties, one being that it satisfies all of the usual properties of a distance measure. The regularized Wasserstein distance is a generalization of this distance measure. It is defined by

\[
\min_{\pi \in \Pi(\mu_0, \mu_1)} \int_{X \times Y} \pi(x, y) \|x - y\|^2 \delta x \delta y - \gamma H(\pi(x, y)),
\]

where \( \gamma \geq 0 \) and \( H(\pi) := -\int_{X \times Y} \pi(x, y) \log \pi(x, y) \delta x \delta y \) is the Shannon entropy of \( \pi \) (Cuturi, 2013; Cuturi and Doucet, 2014). This objective function is strictly convex when \( \gamma > 0 \); however, the primary advantage of adding this entropy term is that it will allow us to efficiently compute the optimal coupling, say \( \pi^*_\gamma \), as we will show below.

Rather than using continuous functions for \( \pi, \mu_1, \) and \( \mu_0 \), we will approximate these functions by discretizing them on a uniform mesh. The coupling \( \pi \) will be stored as an \( m \times n \) matrix and \( \mu_0 \) and \( \mu_1 \) will be stored as \( n \times 1 \) and \( m \times 1 \) vectors respectively. In other words, if \( x_i \in X \) and \( y_i \in Y \) are two points on the discretized domain of \( \mu_0 \) and \( \mu_1 \), then \( \pi_{ij} = \pi(x_i, y_j) \), \( \mu_0_j = \mu_0(x_j) \), and \( \mu_1_i = \mu_1(y_i) \). For notational convenience, we will also define the \( m \times n \) matrix \( M \) so that \( M_{ij} = \|x_i - y_j\|^2 \). After discretizing, (2) can be written as

\[
W_\gamma(\mu_0, \mu_1) := \min_{\pi} \frac{1}{mn} \sum_{i,j} \pi_{ij} M_{ij} + \gamma \pi_{ij} \log(\pi_{ij}) \text{ subject to:}
\]

\[
\frac{1}{m} \sum_i \pi_{ij} = \mu_0_j \forall j \in \{1, 2, ..., n\}
\]

\[
\frac{1}{n} \sum_j \pi_{ij} = \mu_1_i \forall i \in \{1, 2, ..., m\}.
\]

The corresponding Lagrangian is given by
\[ \mathcal{L} = \sum_{i,j} \frac{1}{mn} (\gamma \pi_{ij} \log(\pi_{ij}) + \pi_{ij} M_{ij}) + \lambda_0 \left( \frac{1}{m} \sum_i \pi_{ij} - \mu_{0j} \right) + \lambda_1 \left( \frac{1}{n} \sum_j \pi_{ij} - \mu_{1i} \right), \]

and the first order conditions imply that there exists \((v, w) \in \mathbb{R}^{m+n}_+\) such that the optimal coupling is equal to \(K_{ij} w_i v_j\), where \(K_{ij} := \exp(-M_{ij}/\gamma)\). This can also be written as

\[ \pi_\star = D_w K D_v, \]

In other words, adding the entropy term to the objective function reduces the dimensionality of the optimal coupling from \(mn\) to \(m + n\). The \(m + n\) elements of \(v\) and \(w\) must be set so that the \(m + n\) equality constraints given in (4) and (5) hold. Naively counting the degrees of freedom in this way does not lead us astray in this case; Sinkhorn (1967) shows that the vectors \(v\) and \(w\), such that (4) and (5) hold, are unique up to a multiplicative scalar. The iterative proportional fitting procedure is an efficient method of computing these vectors; see Krupp (1979). This method iteratively redefines \(v_j\) so that \(\frac{1}{m} \sum_i \pi_{ij} = \mu_{0j}\) for all \(j\), and subsequently \(w_i\) is redefined so that \(\frac{1}{n} \sum_j \pi_{ij} = \mu_{1i}\) for all \(i\), as summarized in Algorithm 1.

**Algorithm 1** This algorithm describes the iterative proportional fitting procedure.

The methods in the fourth section of this paper will also require \(w\) (and some functions of \(w, \mu_0,\) and \(\mu_1\)) as output. After we introduce these functions in the next section, we can add these additional values to the return statement.

**function** WassersteinDistance\((\mu_0, \mu_1, K)\)

\[
\begin{align*}
w &\leftarrow 1_m \\
\text{until convergence:} & \\
\quad v &\leftarrow \mu_0 \otimes (K^T w) \\
\quad w &\leftarrow \mu_1 \otimes (K v) \\
\text{return } & W_\gamma(\mu_0, \mu_1)
\end{align*}
\]

The first order condition for the dual of (3)-(5) can also be used to show that the assignments of \(v\) and \(w\) given in Algorithm 1 must both hold at the optimal coupling. Cuturi and Doucet (2014) show that the dual of (3)-(5) has the particularly simple form

\[ W_\gamma(\mu_0, \mu_1) = \max_{(x, y) \in \mathbb{R}^{m+n}} x^T \mu_0 + y^T \mu_1 - \gamma \sum_{i,j} \exp((x_i + y_j - M_{ij})/\gamma - 1). \]

The first order conditions of (8) imply

\[ \mu_{0i} / \left( \sum_j \exp(y_j/\gamma - 1) \exp(-M_{ij}/\gamma) \right) = \exp(x_i/\gamma) \text{ and} \]
\[ \mu_{1j} / (\sum_i \exp(x_i/\gamma) \exp(-M_{ij}/\gamma)) = \exp(y_j/\gamma - 1), \]

and, after replacing \( \exp(y_j/\gamma - 1) \) with \( v_j \) and \( \exp(x_i/\gamma) \) with \( w_i \), we can show that these formulas are equivalent to the updates given in Algorithm 1.

In the next section we will consider minimizing regularized Wasserstein distance over \( \mu_1 \) subject to a shape constraint, and we will define the minimizer as the shape-constrained density estimator. Before we move onto formulating this minimization problem, a few comments regarding the effect of \( \gamma \) on the minimizer will be useful.

Higher values of \( \gamma \) correspond to \( \pi^*_\gamma \) being more dispersed. As \( \gamma \) is decreased toward zero, \( W_\gamma(\mu_0, \mu_1) \) converges to \( W_0(\mu_0, \mu_1) \) and \( \pi^*_\gamma \) converges to the optimal unregularized coupling, \( \pi^*_0 \), at a rate of \( c_1 \exp(-c_2/\gamma) \) for fixed \( c_1, c_2 > 0 \); see Benamou et al. (2015) and Cuturi (2013). Decreasing \( \gamma \) also tends to increase the number of iterations that are required for Algorithm 1 to converge. After the proof of convergence is presented in the next section, Remark 2 provides a rule-of-thumb for setting \( \gamma \).

### 3 Shape-Constrained Density Estimation

The input of the density estimator proposed in this paper is a kernel density estimator, \( \mu \), based on \( N \) i.i.d. datapoints drawn from a continuous population density, \( \mu^* \). For simplicity, let’s begin by examining the one dimensional case. The shape-constrained density estimator in this case is defined as the solution to

\[ \min_f W_\gamma(\mu, f) \] subject to \( \left( \frac{1}{2} f^\rho_{j+1} + \frac{1}{2} f^\rho_{j-1} \right)^{1/\rho} \leq f_j \quad \forall j \in \{2, ..., m - 1\}. \tag{9} \]

The set of feasible densities is generally not convex. To ameliorate this problem, we will use a similar formulation as Koenker and Mizera (2010), which instead solves for \( g := f^\rho \). Also, we will exclude the \( k^{th} \) element of \( g \) in the objective function and constraints. Let \( g_{-k} \) be equal to \( g \) after omitting this \( k^{th} \) element. To use Algorithm 1, the elements of \( f \) that do not correspond to this \( k^{th} \) element, or \( f_{-k} \), are set equal to \( g^{1/\rho}_{-k} \) and \( f_k \) is set equal to \( m - \sum_i g^{1/\rho}_{-k,i} \). In a slight abuse of notation, I will also denote the objective function as \( W_\gamma(\mu, g^{1/\rho}_{-k}) \). In other words, \( W_\gamma(\mu, g^{1/\rho}_{-k}) \) is defined as

\[ \text{2A counterexample could be constructed for the case in which } \rho < 0, \text{ and } f^{(1)} \text{ and } f^{(2)} \text{ are both defined by evaluating Gaussian density functions with different means. Since Gaussian densities are log-concave, } f^{(1)} \text{ and } f^{(2)} \text{ are } \rho-\text{concave for any } \rho < 0 \text{ (as long as they are defined using a mesh that is sufficiently dense). However, } (f^{(1)} + f^{(2)})/2 \text{ is bimodal, so it is not } \rho-\text{concave for any value of } \rho. \]
\[
\max_{(x,y) \in \mathbb{R}^{m+n}} x^T \frac{1}{\rho} g_{-k} + x_k \left( m - \sum_i g_{-k,i}^{1/\rho} \right) + y^T \mu - \gamma \sum_{i,j} \exp \left( (x_i + y_j - M_{ij})/\gamma - 1 \right).
\]

(10)

This objective function is convex in \( g_{-k} \) as long as a suitable index \( k \) is chosen, a feature that is discussed in the next section in more detail. The final form of our optimization problem is

\[
\min_{g_{-k}} W_\gamma (\mu, g_{-k}^{1/\rho}) \quad \text{subject to:}
\]

\[
sgn(\rho) g \in \mathcal{K}_{-k},
\]

(12)

where \( \mathcal{K}_{-k} \) is the cone of locally concave functions after discretizing and omitting all constraints that contain \( g_k \). In other words, it can be viewed as a discretization of the set of functions \( \{ u(x) \mid u : X \to \mathbb{R}, \nabla^2 u(x) \leq 0 \forall x \in X \} \). For the one dimensional case, this optimization problem is equivalent to

\[
\min_{g_{-k}} W_\gamma (\mu, g_{-k}^{1/\rho}) \quad \text{subject to:}
\]

\[
\left( \frac{1}{2} g_{i+1} + \frac{1}{2} g_{i-1} - g_i \right) sgn(\rho) \leq 0 \forall i \in \{2, 3, ..., k-2, k+2, ..., m-1\}.
\]

The following theorem shows that the density corresponding to the minimizer of (11) and (12), say \( f^* \), converges to \( \mu^* \) when \( \mu^* \) is \( \rho \)-concave. We use the nondiscretized counterpart of the optimization problem for simplicity and discuss the discretized case in the remarks that follow the proof.

**Theorem 1:** If \( \mu^* : X \subseteq \mathbb{R}^d \to \mathbb{R}_+ \) is \( \rho' \)-concave for \( \rho' < \rho \) and continuous, \( X \) is bounded, \( \mu \) is a kernel density estimator using a Gaussian kernel with a covariance matrix equal to \( \sigma^2 V > 0 \), \( \sigma \) and \( \gamma \) are chosen so that \( \sigma \) and \( \sqrt{\sigma^2 + \gamma/2} \) converge to zero at an asymptotic rate of \( O(N^{-1/(d+4)}) \), then \( f^* \) converges to \( \mu^* \) and the asymptotic mean integrated squared error of \( f^* \) is \( O_p(N^{-4/(d+4)}) \).

**Proof:** Let \( W^* \) equal \( \min_{f} W_\gamma (\mu, f) \). \( W^* \) can be found by solving

\[
\min_{\pi} \int_{\mathbb{R}^m \times \mathbb{R}^m} \pi(x, y) \| x - y \|^2 + \gamma \pi(x, y) \log \pi(x, y) \delta x \delta y \quad \text{such that:}
\]

\[
\int_{\mathbb{R}^m} \pi(x, y) \delta y = \mu(x) \quad \forall x \in \mathbb{R}^m.
\]

(13)

The corresponding Lagrangian is

\[
\mathcal{L} = \int_{\mathbb{R}^m \times \mathbb{R}^m} \pi(x, y) \| x - y \|^2 + \gamma \pi(x, y) \log \pi(x, y) + \lambda(x) \left( \int_{\mathbb{R}^m} \pi(x, y) \delta y - \mu(x) \right) \delta x \delta y,
\]

(14)
and the first order condition implies
\[
\pi(x, y) = \exp(-\|x - y\|^2 / \gamma) \exp(-\lambda(x) / \gamma - 1).
\]
After combining this equality with the constraint, we have
\[
\mu(x) = \exp(-\lambda(x) / \gamma - 1) \kappa,
\]
where \(\kappa := \int_{\mathbb{R}^m} \exp(-\|x - y\|^2 / \gamma) \delta y\). This equality implies
\[
\tilde{f}(y) := \int_{\mathbb{R}^m} \pi(x, y) \delta x = \int_{\mathbb{R}^m} \mu(x) / \kappa \exp(-\|x - y\|^2 / \gamma) \delta x.
\]
Since \(\tilde{f}(y)\) is a marginal of \(\pi(x, y)\), it has a total mass of one, so \(\kappa\) is set so that
\[
\tilde{f}(y) = \mu * \phi_{\gamma/2L},
\]
where \(*\) denotes a convolution and \(\phi_{\gamma/2L}(\cdot)\) is a Gaussian density with a mean equal to zero and a covariance matrix equal to \(\gamma/2L\). Since \(\mu(x)\) is a kernel density estimator, it can be expressed as
\[
\mu(x) = (\sum_i \delta(z_i)) * \phi_{\sigma^2 V}(x)/N,
\]
where \(\{z_i\}_{i=1}^N\) are the datapoints and \(\delta(z_i)\) is a Dirac delta function centered at \(z_i\). After substituting this definition into the equation \(\tilde{f}(y) = \mu * \phi_{\gamma/2L}\), we have
\[
\tilde{f}(y) = (\sum_i \delta(z_i)) * \phi_{\sigma^2 V}(y) * \phi_{\gamma/2L}(y)/N.
\]
The convolution of two Gaussian densities has the simple form \(\phi_{\sigma^2 V}(y) * \phi_{\gamma/2L}(y) = \phi_{\sigma^2 V + \gamma/2L}(y)\), so
\[
\tilde{f}(y) = (\sum_i \delta(z_i)) * \phi_{\sigma^2 V + \gamma/2L}(y)/N.
\]
Parzen (1962) shows that \(\tilde{f}(\cdot) \to \mu^*(\cdot)\) and the asymptotic mean integrated squared error between \(\tilde{f}(\cdot)\) and \(\mu^*(\cdot)\) converges in probability to zero at a rate of \(O_p(N^{-4/(d+4)})\) as long as \(\sqrt{\sigma^2 + \gamma/2} \to 0\) and \(N/\sqrt{\sigma^2 + \gamma/2} \to \infty\), but other choices of these parameters do not result in the same asymptotic rate of convergence (Parzen, 1962). Also, the theorem does not require the assumption that the kernel used to generate \(\mu\) is Gaussian. Any kernel can be used as long as the convolution of that kernel with a Gaussian density satisfies the required conditions given by Parzen (1962). On a more theoretical note, even though we assume that \(\sigma > 0\), the limit of \(\tilde{f}\) as \(\sigma \to 0\) does exist; thus the first part of the proof also shows that kernel density estimators can be viewed as the minimizers \(W_\gamma(\sum_i \delta(z_i)/N, f)\) with respect to \(f\).
Remark 2: Algorithm 1 can be slow to converge or fail to converge if $\gamma$ is chosen to be too small, but it is worth noting that the requirement that $\gamma \to 0$ is not problematic when the conditions given in the theorem hold because Algorithm 1 converges in one iteration when its inputs are the densities $\mu$ and $K(\mu \otimes (K^T1_m))$. The density $K(\mu \otimes (K^T1_m))$ can be viewed as the discretized counterpart of the density $\tilde{f}$ in the proof of Theorem 2, in the sense that $K(\mu \otimes (K^T1_m)) = \arg\min_f W_\gamma(\mu, f)$. For this reason $f^* = K(\mu \otimes (K^T1_m))$ if $K(\mu \otimes (K^T1_m))$ satisfies the shape constraints. As long as $\mu^*$ is $\rho'$-concave for $\rho' < \rho$, this will be the case asymptotically since $\mu \to \mu^*$ as $N \to \infty$ and since $K(\mu^* \otimes (K^T1_m)) \to \mu^*$ as $\gamma \to 0$ as long as $X \subset Y$.

We can use the interpretation of $\tilde{f}$ from the proof above to find a rule of thumb for setting $\gamma$. Specifically, $\sqrt{\sigma^2 + \gamma/2}$ can be set using a rule-of-thumb bandwidth estimator while ensuring $2\sigma^2/\gamma$ is a constant; for examples of rule-of-thumb bandwidth estimators see Silverman, 1986; Scott, 1992. In practice, setting these parameters so that $\sigma^2/\gamma = 1/4$ works well. Combining this equality with Scott’s rule of thumb implies $\sigma^2V = N^{-2/(d+4)}S/3$ and $\gamma = N^{-2/(d+4)}\text{tr}(S/d)4/3$, where $S$ and $\text{tr}(S)$ are the covariance matrix of the data and its trace respectively, and $d$ is the number of dimensions.

Other than cases in which $\mu$ is fairly far from the set of densities that are $\rho$-concave, this rule for setting $\sigma^2$ and $\gamma$ tends to lead to values that are higher than required for convergence of Algorithm 1. If one wishes to ensure that the shape constraints are more directly responsible for the parsimony of $f^*$, rather than the smoothing caused by setting $\sqrt{\sigma^2 + \gamma/2}$ to a value that is larger than required, $\gamma$ can be decreased to a point in which the shape constraints bind over larger subsets of the domain. Given an approximation of $f^*$, a root finding algorithm can be used to approximate the value of $\gamma$ that lead to convergence within a fixed number of iterations (a few hundred to a thousand iterations would be reasonable choices), and the output of this root finding algorithm can be thought of as an approximate lower bound on $\gamma$. The fifth section derives a function that is capable of approximating $f^*$.

4 Main Algorithm

In this section we will derive a trust region algorithm to find the global minimum of (11) and (12). We start this process by deriving the gradient and Hessian of $W_\gamma(\mu, g_{-k}^{1/\rho})$ for cases in which $\rho \neq 0$. Appendix B contains these derivations for the log-concave case, which corresponds to the limit as $\rho \to 0$. When the densities are defined in multiple dimensions, Appendix A describes a modification of Algorithm 1 that can provide a substantial gain in computational efficiency.

---

3 Algorithm 1 would begin by initializing $w$ at $1_m$ and then apply $v \leftarrow \mu \otimes (K^T1_m)$. The $w$-update would then set $w$ equal to $f \otimes K(\mu \otimes (K^T1_m))$, but since this is also equal to $1_m$, the algorithm has already converged.
The gradient of \( W_\gamma(\mu, g^{1/\rho}_\perp) \) can be found by applying the envelope theorem to (10). This implies
\[
r_i := \frac{\partial W_\gamma(\mu, g^{1/\rho}_\perp)}{\partial g_i} = (x_i - x_k)g^{-1/\rho - 1}_{\perp, i} / \rho.
\] (15)

To derive the Hessian of \( W_\gamma(\mu, g^{1/\rho}_\perp) \), we need to find the gradient of \( x_i \) and \( x_k \) with respect to \( g - k \). Implicit differentiation of the first order conditions of (10) imply
\[
\frac{\partial x_i}{\partial g - k} = \begin{cases} 
\gamma g^{-1/\rho - 1}_{\perp, i} / \left( \exp(x_i / \gamma) \sum_j \exp(-M_{ij} / \gamma) \exp(y_j / \gamma - 1) \rho \right) & \text{if } i = j \\
0 & \text{if } i \neq j
\end{cases}
\]
and
\[
\frac{\partial x_k}{\partial g - k} = -\gamma g^{-1/\rho - 1}_{\perp, k} / \left( \exp(x_k / \gamma) \sum_j \exp(-M_{kj} / \gamma) \exp(y_j / \gamma - 1) \rho \right)
\]

We can combine these derivatives with the first order of conditions of (8) to show that the Hessian is
\[
H := \nabla^2 W_\gamma(\mu, g^{1/\rho}_\perp) = D_b + a (f^{1-\rho}_\perp) \cdot (f^{1-\rho}_\perp)^T,
\] (16)
where
\[
a := \gamma / (f_k \rho^2) \text{ and }
\]
\[
b_i := ((x_i - x_k)(1 - \rho) + \gamma) f^{-1-2\rho}_{\perp, i} / \rho^2.
\]

Note that choosing choosing \( k \) to be \( \arg \min_i w_i \) is a sufficient condition for this matrix to be positive definite; however, it is generally possible to choose \( k \) to correspond to one of the elements on the boundary of the domain of \( \mu^* \) to ensure that the density estimate satisfies the shape constraints everywhere on the interior of its domain while maintaining the convexity of the objective function.

Having already derived the gradient and Hessian of \( W_\gamma(\mu, g^{1/\rho}_\perp) \), 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
\[
\Delta \leftarrow \arg \min_d \left\{ d^T H d / 2 + d^T r \mid (d + g^{(i-1)}_{\perp}) sgn(\rho) \in K_{\perp}, \|d\| \leq c \right\}.
\]

If the value of the objective function evaluated at \( g^{(i-1)}_{\perp} + \Delta \) results in an improvement over its value at \( g^{(i-1)}_{\perp} \), then \( g^{(i)}_{\perp} \) is defined to be \( g^{(i-1)}_{\perp} + \Delta \). If the improvement was significant, then the radius of the trust region, \( c \), is increased; otherwise it is decreased. This is described in Algorithm 2.
Algorithm 2 The parameter values for this method were set using the recommendations of Fan and Yuan (2001). DefineGamma(\(\mu, \rho, f^{(0)}\)) finds the value of \(\gamma\) such that Algorithm 1 converges within a given range of iterations and then defines \(K\) using this value of \(\gamma\). WassersteinDistance is similar to Algorithm 1 after adding the gradient, Hessian, and the index \(k\) (as defined in section three) to its output.

```
function TrustRegion(\(\mu, \rho, f^{(0)}, K = []\))
    if \(K = []\):
        \(K \leftarrow \text{DefineGamma}(\mu, \rho, f^{(0)})\)
    \((W^{(0)}, H^{(0)}, r^{(0)}, k^{(0)}) \leftarrow \text{WassersteinDistance}(\mu, \rho, f^{(0)}, K)\)
    \(g^{(0)}_{-k} \leftarrow f^{(0)}_k\)
    \(\tilde{W}^{(0)} \leftarrow W^{(0)}\)
    \(\eta \leftarrow 1, c \leftarrow 1\)
    for \(i = 1, 2, \ldots:\)
        \((W^{(i)}, H^{(i)}, r^{(i)}, k^{(i)}) \leftarrow \text{WassersteinDistance}(\mu, \rho, K, f^{(i-1)})\)
        \(W^{(i)} > W^{(i-1)}; W^{(i)} \leftarrow W^{(i-1)}, H^{(i)} \leftarrow H^{(i-1)}, r^{(i)} \leftarrow r^{(i-1)}, k^{(i)} \leftarrow k^{(i-1)}, g \leftarrow f^{(i-2)}_k\)
        if \((W^{(i)} - W^{(i-1)})/\tilde{W}^{(i-1)} < 1/4; c \leftarrow c/4 + \eta/8\)
        else: \(c \leftarrow 3c/2\)
        \(\Delta \leftarrow \arg\min_d d^T H d/2 + d^T r\) s.t. \(d + g_{-k} \in \mathcal{K}_{-k}, \|d\| \leq c\).
        \(W^{(i)} \leftarrow \Delta^T H \Delta/2 + \Delta^T r\)
        \(\eta \leftarrow \|\Delta\|\)
        \(g_{-k} \leftarrow g_{-k} + \Delta\)
        \(g_k \leftarrow (m - \sum_{i} g^{1/\rho}_{-k,i})^{\rho}\)
        \(f^{(i)} \leftarrow g^{1/\rho}\)
    return \(f\)
```

Figures 1 and 2 illustrate two examples of the output of Algorithm 2. 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 density estimate 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).
Figure 1: The red density in each plot is a kernel density estimate of the rotational velocity of stars from Hoffleit and Warren (1991). The blue density in (a) is the $-1/2$–concave density estimate, while in (b) this density is the $-2$–concave density estimate.

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 estimate 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 show that the logarithm of the MLE 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.

MOSEK, a highly optimized quadratic program solver, generally finds $\Delta$ in a timely manner; however, this can still be a time consuming step when $f$ and $\mu$ are defined in two dimensions. The algorithm is more computationally efficient if $f^{(0)}$ is a good approximation of $f^\star$. The next section presents an algorithm that finds an approximation of $f^\star$ for this purpose.

5 Bregman Projections

Algorithm 1 can be derived using the method of alternating Bregman projections (MABP), which is also the basis for the algorithm proposed in this section (Bregman, 1967). Bregman explored a class of divergence measures defined by

$$D_{\varphi}(x \mid y) := \varphi(x) - \varphi(y) - (x - y)^T \nabla \varphi(y),$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a convex function. In other words, if $\hat{\varphi}_y(x)$ is the first order Taylor series expansion of $\varphi(\cdot)$ at $y \in \mathbb{R}^n$, then $D_{\varphi}(x \mid y) = \varphi(x) - \hat{\varphi}_y(x)$. Many distance measures and divergences can be viewed as Bregman divergences.
For example, squared Euclidian distance can be derived using \( \varphi(x) := \|x\|^2 \), and \( \varphi(x) := \sum_i x_i \log(x_i) \) results in Kullback-Leibler divergence.

Bregman described a way to to minimize \( \varphi(\cdot) \) subject to multiple sets of affine constraints using this divergence measure. As an example, let’s consider

\[
\min_x \varphi(x) \text{ subject to: } A_1 x = b_1, \ A_2 x = b_2.
\]

For \( l \in \{1, 2\} \), let the Bregman projection of \( y \) onto the constraint \( A_l x = b_l \) be denoted by

\[
P_l(y) := \arg \min_{A_l x = b_l} D_{\varphi(l)}(x \mid y).
\]

MABP begins by initializing \( x^{(0)} \) at \( \arg \min_x \varphi(x) \) and the \( i \)th iteration takes \( x^{(i-1)} \) as input and defines \( x^{(i)} \) as

\[
x^{(i)} \leftarrow P_2(P_1(x^{(i-1)})).
\]

This is often a very efficient way to solve an optimization problem when there is an analytic solution for one or both of the projections. For example, the two updates found in Algorithm 1 can be viewed as Bregman projections of the coupling \( \pi \) onto either \( \pi 1_n/n = \mu_1 \) or \( \pi^T 1_m/m = \mu_0 \). For simplicity, we will derive the algorithm proposed in this section in the one dimensional case. To approximate \( f^* \), we need to solve

\[
\min_\pi \frac{1}{mn} \sum_{i,j} \pi_{ij} M_{ij} + \gamma \pi_{ij} \log(\pi_{ij}) \quad \text{subject to:} \\
\pi^T 1_m/m = \mu, \tag{17}
\]

\[
\pi 1_n/n = f, \ \text{and} \ \left( \frac{1}{2} f_{j+1}^\rho + \frac{1}{2} f_{j-1}^\rho \right)^{1/\rho} \leq f_j \ \forall \ j \in \{2, \ldots, m-1\}. \tag{18}
\]

We can only guarantee that MABP converges to the global minimum if the constraints are affine. The constraint in (19) is not convex, so we cannot guarantee that MABP will converge. Regardless, MABP is often employed with reasonable success in nonconvex cases; for examples, see Bauschke, Borwein, and Combettes (2003) and the references therein. MABP also performs well in this setting, and since the output of the algorithm will only be used to initialize Algorithm 2, inaccuracies in the output will not impact our final density estimates.

The Bregman divergence corresponding to the objective function in (3) is

\[\text{The equality constraints could be replaced with inequalities. However, the constraints must have a nonempty intersection, be closed, and be affine. Bauschke and Lewis (2000) prove that a similar algorithm, which replaces the requirement that the constraints are affine with a convexity assumption, also converges to the global minimum.}\]
\[ \frac{1}{mn} \sum_{ij} x_{ij} \log(x_{ij}/y_{ij}) - x_{ij} + y_{ij}. \]

As previously mentioned, the Bregman projection onto the constraint in (18) is \( v \leftarrow \mu \otimes (K^T w) \). The constraints given by (19) can be combined to define the projection,

\[ P_2(\pi) := \arg \min_{\pi} \frac{1}{mn} \sum_{ij} \pi_{ij} \log \left( \frac{\pi_{ij}}{e^{\pi_{ij}}} \right) \text{ subject to } \]

\[ 2 \left( \frac{1}{n} \sum_j \pi_{ij} \right)^\rho - \left( \frac{1}{n} \sum_j \pi_{i+1,j} \right)^\rho - \left( \frac{1}{n} \sum_j \pi_{i-1,j} \right)^\rho \geq 0. \quad (20) \]

The first order conditions of the corresponding Lagrangian imply

\[ \pi_{ij} = \bar{\pi}_{ij} \exp \left( mn\rho \left( \frac{1}{n} \sum_j \bar{\pi}_{ij} \right)^{\rho-1} (\lambda_{i-1} + \lambda_{i+1} - 2\lambda_i) \right). \quad (21) \]

After summing over \( j \) on both sides and dividing by \( n \) we have

\[ f_i = \bar{v}_i \exp \left( mn\rho f_i^{\rho-1} \theta_i \right), \quad (22) \]

where \( \bar{v}_i := \frac{1}{n} \sum_j \bar{\pi}_{ij} \) and \( \theta_i := \lambda_{i-1} + \lambda_{i+1} - 2\lambda_i \).

Rather than attempting to solve (20) numerically, we can use the change of variable \( f = g^{1/\rho} \) from the third section. The following problem has equivalent KKT conditions, but the dimension is reduced from \( mn \) to \( m \).

\[ \arg \min_g \frac{1}{n} \sum_i g_i^{1/\rho} \log \left( g_i^{1/\rho} \right) \text{ subject to: } \]

\[ \text{sgn}(\rho)(2g_i - g_{i+1} - g_{i-1}) \leq 0 \quad \forall i \in \{2, 3, ..., m - 1\} \]

However, to ensure this optimization problem is convex we need to add the constraint \( g_i \leq \bar{v}_i^\rho \exp(-\rho/(1-\rho)) \) for every \( i \in \{1, 2, ..., m\} \), which may be a source of inaccuracy for this algorithm when \( \rho \) is set to a small value. To give some idea how this constraint affects the density estimate, let’s explore the cases in which \( \rho \) is equal to \(-5\) and \( 0 \). The constraint can also be thought of as specifying a lower bound on \( f_i \) of \( \bar{v}_i \exp(-1/(1-\rho)) \), so when \( \rho \) takes the low value of \(-5\), \( f_i \) is constrained to be greater than approximately 0.85\( \bar{v}_i \). However, when \( \rho = 0 \), \( f_i \) is constrained to be greater than \( \bar{v}_i/e \approx 0.37\bar{v}_i \). The pseudocode for this method is given in Algorithm 3. Generally five to thirty iterations are sufficient to provide a reasonable value for \( f(0) \).

\[ \text{Note that after } v \text{ is defined using } v \leftarrow \mu \otimes (K^T w), \text{ the equality } \pi^*_\gamma = D_wK D_v \text{ from the second section implies that solving for the optimal density is equivalent to solving for the optimal value of } w \text{ such that the density } D_wK v \text{ satisfies the shape constraint. The variable } \bar{v} \text{ is } K v, \text{ the component of } f \text{ that is already fixed by the } v-\text{update.} \]
Algorithm 3 Produces an approximate shape-constrained density estimate using MABP. Note that no constraint regarding the mass of $f$ was made. The mass of $f$ will be correct in the limit due to the assignment $v \leftarrow \mu \odot (Kw)$.

function $\text{MABP}(\mu, K, \rho)$

$w \leftarrow 1_m$

for $i = 1, 2, ...$

$v \leftarrow \mu \odot (Kw)$

$\bar{v} \leftarrow Kw$

$g \leftarrow \arg \min \frac{1}{n} \sum_i g_i^{1/\rho} \log \left( \frac{g_i^{1/\rho}}{e^M} \right)$ s.t. $2g_i - g_{i+1} - g_{i-1} \geq 0$, $g_i \leq h^\rho \exp(-\rho/(1-\rho))$

$f \leftarrow g^{1/\rho}$

$w \leftarrow f \odot (Kv)$

return $f$

6 Alternative Shape Constraints

In this section we will show that minimizing the regularized Wasserstein distance subject to a shape constraint is a convex optimization problem for a wide variety of shape constraints. To do this we define $\omega(x, y; g)$ as

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

where $f_k = \alpha(g_k)$ and $\alpha : \mathbb{R}^{m-1} \rightarrow \mathbb{R}^{m-1}_+$ has a continuous second derivative. This allows us to express the Wasserstein distance as

$$W_\gamma(\mu, \alpha(g_k)) = \max_{(x, y) \in \mathbb{R}^{m+n}} \omega(x, y; g). \quad (23)$$

Now consider a shape-constrained density estimator that is defined as the minimum of

$$\min_{g_k} W_\gamma(\mu, \alpha(g_k)) \quad (24)$$

subject to: $\beta(g_k) \leq 0$, \quad (25)

where $\beta : \mathbb{R}^{m-1} \rightarrow \mathbb{R}^{m-1}$ is a continuous and convex function. The envelope theorem and implicit differentiation implies that the Hessian of $W_\gamma(\mu, \alpha(g_k))$ is

$$\nabla^2_{g_k} W_\gamma(\mu, \alpha(g_k)) = \nabla^2_{g_k} \omega(x, y; g) + \left( \nabla_{g_k} \nabla x \omega(x, y; g_k) \right) \left( -\nabla^2_{g_k} \omega(x, y; g_k) \right)^{-1} \left( \nabla x \nabla_{g_k} \omega(x, y; g_k) \right).$$

For the ease of notation, let the matrices $A, B,$ and $C$ correspond to the matrices above so that $\nabla^2_{g_k} W_\gamma(\mu, \alpha(g_k)) = A + BCB^T$. The first order conditions of (23) imply
\[ C = (-\nabla_x^2 \omega(x, y; g))^{-1} = \gamma D_{1m} \sigma_f. \]

In other words, \( C \) is a positive definite matrix as long as each element of \( f \) is strictly positive, so \( BCB^T \) is also positive definite. It can also be shown that each nonzero element of \( A \) is multiplied by \( x_i - x_j \) for \( i, j \in \{1, 2, ..., n\} \), so each element of \( A \) is zero whenever each element of \( x \) is equal. As described in Remark 2 of Theorem 1, this occurs when \( f \) is the minimizer of \( W^2_\gamma(\mu, f) \), say \( \tilde{f} \). By initializing our optimization procedure at

\[
\arg\min_g g^T \left( \nabla^2 g_{-k} W_\gamma(\mu, \alpha(g_{-k})) \middle| g_{-k} = \alpha^{-1}(\tilde{f}) \right) g
\]

subject to: \( \beta(g_{-k}) \leq 0 \),

the Hessian is positive definite along the entire path of convergence as long as \( \tilde{f} \) is sufficiently close to the set of densities that satisfy the shape constraint. In the case of shape constraints that have the effect of increasing the parsimony of \( f \) in some way, this can often be ensured by increasing \( \sigma \). This approach allows the framework presented in this paper to be applied to other shape constraints than those that can be described using \( \rho^{-}\text{concavity}. \)

Constraining density estimates to have a monotonic virtual valuations function would be particularly useful in Economics. Ewerhart (2013) shows that this is equivalent to constraining \( g(x) = 1/(1 - F(x)) \) to be convex, where \( F(x) \) is the distribution function. We can solve for \( F(x) \) and take the first derivative to show that the corresponding density function is \( f(x) = g'(x)/g(x)^2 \), which can be approximated on a mesh to solve (24) and (25).

Wellner (2017) shows that constraining a density to have an increasing hazard function is equivalent to constraining \( g(x) = -\log(1 - F(x)) \) to be convex. After using a similar strategy as above, this implies, \( f(x) = \exp(-g(x))g'(x) \). Both of these examples lead to the matrix \( A \) having positive diagonal terms; however, guaranteeing that \( A \) is positive definite would require \( f \) to satisfy a set of inequalities that do not appear to have an obvious interpretation. Regardless, we can ensure positive definiteness by increasing \( \sigma \). In some cases this may increase the dispersion of \( f \) by more than we would prefer. In these cases, it would be best to compare the resulting density estimate with an estimate subject to a stronger constraint and check which density fits the data more closely. Ewerhart (2013) shows that a sufficient condition for the density \( f(x) \) to satisfy Myerson’s (1981) regularity condition is for \( f(x) \) to be \( \rho^{-}\text{concave} \) for \( \rho > -1/2 \), and a log-concavity constraint can be used to ensure that the hazard function is monotonic.

7 An Application

The California Department of Transportation (Caltrans) uses a first price auction 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. If \( f(x) \) is the density of 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 allocate the item being auctioned to the person with the highest bid are optimal when the virtual valuations of the bidders, which is given by \( x - (1 - F(x)) / f(x) \), is monotonically increasing.

To examine whether this condition is plausible, we need to estimate the valuations (or 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

\[
b_i + \frac{\hat{F}_b(b_i)}{(n-1)f_b(b_i)},
\]

where \( n \) 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.

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 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 (after normalizing by the Caltrans’s engineers’s 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 (2013) 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 estimate the valuations for each firm, we used a kernel density estimate with a bandwidth set using Silverman’s rule of thumb; however, the resulting virtual valuations function was not increasing in two regions of the domain of the density. This could be an innocuous idiosyncrasy of the data or it could be evidence that
Caltrans’s choice of auction format is suboptimal. To investigate which possibility is more plausible, we used a kernel density estimate with a bandwidth equal to one third of the bandwidth implied by Silverman’s rule of thumb. Afterward we found the density that minimizes the regularized Wasserstein distance from this input density subject to the constraint that the virtual valuations function is monotonic. Despite the fact that $\gamma$ and $\sigma^2$ were set so that $\sqrt{\sigma^2 + \gamma/2}$ was approximately seven percent less than the recommendation given in Remark 2 of Theorem 1, the Hessian of the objective function was convex along the path of convergence.

The input density and the output of the algorithm are shown in Figure 3. The fact that the output density has multiple modes demonstrates that this constraint is a significant relaxation of $\rho$–concavity. It is also apparent from the figure that the density of valuations is fairly close to the set of densities that have monotonic virtual valuations.

We also tested the null hypothesis that the valuations follow our shape-constrained density estimate against the alternative hypothesis that this is not the correct density using an Anderson-Darling (1954) test. Since our density estimate is non-standard, 10,000 samples were drawn from the shape-constrained density estimate to approximate the critical value. Using a significance level of 0.1, the approximate critical value is 2.20. Since our test statistic is 0.33, we fail to reject our hypothesis that the density estimate that is constrained to satisfy Myerson’s regularity condition is the correct density.
Figure 3: The density estimate of the valuations of firms is shown in red, and the density shown in blue minimizes the regularized Wasserstein distance from the red density subject to a constraint that the virtual valuations function of the density is monotonic. The absolute value of the valuations can be viewed as the cost per dollar of the Caltrans’s engineer’s estimates.

8 Conclusion

One of the differences between other shape-constrained density estimators proposed in the literature and the one proposed here is that a density is used as input rather than the data itself. The estimators proposed by Cule, Samworth, and Stewart (2010) and Koenker and Mizera (2010) provide density estimates that are zero outside of the convex hull of the data. One byproduct of using a kernel density estimator as input is that the resulting shape-constrained density estimate has non-zero tails outside of the convex hull of the data. Another implication of using a kernel density estimator as input is that $g$ is not necessarily affine on each facet of a triangulation of the data points, as in Cule, Samworth, and Stewart (2010).

The framework presented here can also be extended to allow $\gamma$ to take different values at each column of the matrix $K$, which would be appealing in two situations. When one would like $f^*$ to be as close as possible to $\mu$, $\gamma$ can be decreased below what would have otherwise been possible in regions where the shape-constrained density
estimator is closer to $\mu$ without interfering with the convergence of Algorithm 1. Secondly, this would allow for the development of methods that set $\gamma$ and $\sigma^2$ using an adaptive approach that is similar to the one described by Sheather and Jones (1991) for kernel density estimators.

Acknowledgements

The author would like to thank Roger Koenker for his guidance throughout this project as well as Alfred Galichon, Gabriel Peyré, and the attendees of the Econometrics seminar at the University of Illinois at Urbana Champaign for their insightful comments.
Appendix A: Estimation in Multiple Dimensions

Solomon et al. (2015) showed that Algorithm 1 can be approximated by using convolutions to blur \( v \) and \( w \) rather than by multiplying \( v \) and \( w \) by the matrix \( K \). This can result in a gain in computational efficiency when these densities are defined in two dimensions. We will denote the application of this Gaussian filter by the function \( K(\cdot) \). The pseudocode of \( \text{WassersteinDistance}(\cdot) \) when \( \mu_0, \mu_1 \in \mathbb{R}^{m \times n}_+ \) is given in Algorithm 4.

**Algorithm 4** The function \( \text{WassersteinDistance}(\cdot) \) when \( \mu_0, \mu_1 \in \mathbb{R}^{m \times n}_+ \). The function \( K(\cdot) \) denotes a Gaussian filter.

```
function \text{WassersteinDistance}(\mu_0, \mu_1, K)
    w \leftarrow 1_{m \times n}
    \text{until convergence:}
        v \leftarrow \mu_0 \odot K(w)
        w \leftarrow \mu_1 \odot K(v)
    \text{return } W_\gamma(\mu_0, \mu_1)
```

As in Koenker and Mizera (2010), concavity of \( g \cdot \text{sgn}(\rho) \) is ensured in the two dimensional case by constructing the discrete local Hessian of \( g \) at each point on the interior of its domain. The elements of this Hessian are equal to

\[
E_{1,1} = g_{i,j+1} + g_{i,j-1} - 2g_{i,j},
\]

\[
E_{2,2} = g_{i+1,j+1} + g_{i-1,j-1} - 2g_{i,j},
\]

and \( E_{1,2} = E_{2,1} = (g_{i-1,j+1} + g_{i+1,j-1} - g_{i-1,j-1} - g_{i+1,j+1})/4 \).

The shape constraints in this case correspond to constraining \( E \cdot \text{sgn}(\rho) \) to be negative semidefinite at all indices \( i \) and \( j \) on the interior of the matrix \( g \). The determinant of \( E \) can be formulated as a rotated conic, which can be implemented in MOSEK.

Appendix B: Log-Concavity Constraints

A different approach is necessary for \( \rho \to 0 \), which corresponds to log-concavity. In this case \( g = \log(f) \) is constrained to be concave, and \( W_\gamma(\mu, g^{1/\rho}_{-k}) \) is defined as

\[
\max_{(x,y) \in \mathbb{R}^{m+n}} x^T_k \exp(g_{-k}) + x_k (m - \sum_i \exp(g_{-k,i})) + y^T \mu - \gamma \sum_{i,j} \exp((x_i + y_j - M_{ij})/\gamma - 1).
\]  

(27)

The index \( k \) can be chosen in the same way as described above to ensure this function is convex.
The gradient in this case is equal to

$$r_i := \frac{\partial W_\gamma(\mu, g_{1/\rho}^k)}{\partial g_{-k,i}} = (x_i - x_k) \exp(g_{-k,i}),$$  \hspace{1cm} (28)

and the Hessian is

$$H := \nabla^2 W_\gamma(\mu, g_{1/\rho}^k) = D_b + a f_{-k} \cdot f_{-k}^T,$$  \hspace{1cm} (29)

where

$$a := \gamma / f_k$$

and

$$b_i := (x_i - x_k + \gamma) f_{-k,i}.$$
References

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association, 49*(268), 765-769.

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.*

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.

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.

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.

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).

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.

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: Techniques and Applications* (pp. 786-794). ICOTA, Hong Kong.

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

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

Kantorovich, L. V. (1942). On the translocation of masses. In Dokl. Akad. Nauk SSSR (Vol. 37, pp. 199-201).

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

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.

Kruithof, J. (1937). Telefoonverkeersrekening. De Ingenieur, 52, E15-E25.

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

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

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

Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33(3), 1065-1076.

Scott, D. W. (1992). Multivariate Density Estimation. Wiley.

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. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. The Annals of Statistics, 795-810.

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.

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.

Student. (1908). The probable error of a mean. Biometrika, 1-25.

Wellner, J. A. & Laha, N. (2017). Bi-s-Concave Distributions. arXiv preprint.