Shape-Constrained Density Estimation Via Optimal Transport

Ryan Cumings-Menon*

November 13, 2018

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.[1] 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.

JEL Classification: C14

Keywords: Nonparametric density estimation, Kernel density estimation, Optimal transport, Log-concavity, $\rho-$concavity$/s-$concavity

*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. *Phone number:* (515) 556-3263.

[1] The maximum likelihood estimator is simply the average over Dirac delta functions centered at each datapoint, which is also known as Dirac catastrophe.

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 $\rho \to 0$.[2] 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 strictly increasing. Sufficient conditions for Myerson's 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

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

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 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's 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 regularity condition is satisfied, appears to follow the data closely. Our test also fails to reject that the population density satisfies Myerson's 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 choice of ρ. 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 primarily advantageous when using shape constraints that are not sufficiently restrictive to ensure parsimony by themselves. 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 provides an algorithm to calculate the estimator. Note that initializing this algorithm at a reasonable approximation of the density estimator provides a gain in computational efficiency, and an algorithm for finding an approximation is provided in the appendix. 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 $\mu_i = \mu(\mathbf{a}_i)$.

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

$$W_0(\mu_0(\cdot), \mu_1(\cdot)) := \min_{\psi \in \Psi(\mu_0(\cdot), \mu_1(\cdot))} \int_{\mathcal{A} \times \mathcal{A}} \|a_1 - a_2\|^2 \, d\psi(a_1, a_2), \tag{1}$$

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

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

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.[3] 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

$$W_\gamma(\mu_0, \mu_1) := \min_\psi \sum_{i,j} \psi_{ij} M_{ij} + \gamma \psi_{ij} \log(\psi_{ij}) \text{ subject to:} \tag{3}$$

$$\sum_j^m \psi_{ij} = \mu_{0i} \; \forall i \in \{1, 2, ..., m\} \tag{4}$$

$$\sum_i^m \psi_{ij} = \mu_{1j} \; \forall i \in \{1, 2, ..., m\}. \tag{5}$$

The corresponding Lagrangian is given by

$$\mathcal{L} = \left(\sum_{i,j} \gamma \psi_{ij} \log(\psi_{ij}) + \psi_{ij} M_{ij} \right) - \lambda_0^T \left(\sum_j^m \psi_{\cdot j} - \mu_0 \right) - \lambda_1^T \left(\sum_i^m \psi_{i\cdot} - \mu_1 \right), \tag{6}$$

[3]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).

and the first order conditions imply

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

In other words, there exists $v, w \in \mathbb{R}_+^m$ such that the optimal coupling has elements $\psi_{ij} = K_{ij}w_iv_j$, 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,

$$\psi = D_w K D_v, \tag{7}$$

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 $D_wKv = \mu_0$, and subsequently redefines v so that $D_vKw = \mu_1$, as summarized in Algorithm 1. Note that after combining these equalities we have $D_wK(\mu_1 \oslash (Kw)) = \mu_0$, which will be used in subsequent sections.

Algorithm 1 The iterative proportional fitting procedure.

function $\texttt{IPFP}(K, \mu_0, \mu_1)$
$w \leftarrow \mathbf{1}_m$
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,

$$W_\gamma(\mu_0, \mu_1) = \max_{(x,y)\in\mathbb{R}^{2m}} x^T\mu_0 + y^T\mu_1 - \gamma\sum_{i,j}\exp\left((x_i + y_j - M_{ij})/\gamma\right). \tag{8}$$

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

$$\mu_{0i}/\left(\sum_j \exp\left(y_j/\gamma\right)K_{ij}\right) = \exp\left(x_i/\gamma\right) \tag{9}$$

$$\mu_{1j}/\left(\sum_i \exp\left(x_i/\gamma\right)K_{ij}\right) = \exp\left(y_j/\gamma\right). \tag{10}$$

Note that after replacing $\exp\left(x/\gamma\right)$ with w and $\exp\left(y/\gamma\right)$ 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.

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,

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

$$sgn(\rho) f^\rho \in \mathcal{K}, \tag{11}$$

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,

$$\max_{(x,y)\in\mathbb{R}^{2m}} x_{-k}^T g^{1/\rho} + x_k \left(m - \sum_i g_i^{1/\rho} \right) + y^T \mu - \gamma \sum_{i,j} \exp\left((x_i + y_j - M_{ij})/\gamma\right), \quad (12)$$

and the final form of our optimization problem is,

$$\min_g W_\gamma(g^{1/\rho}, \mu) \text{ subject to:} \quad (13)$$

$$g_i = \alpha_i + \beta_i \mathrm{a}_i, \;\; sgn(\rho)\left(\alpha_i - \alpha_j + (\beta_i - \beta_j)^T \mathrm{a}_i\right) \leq 0 \;\; \forall\, i, j \in \{2, ..., m-1\}, \quad (14)$$

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 $\min_{i\neq j} \|\mathbf{a}_i - \mathbf{a}_j\| / \sqrt{\gamma} \to$*

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,[4]

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

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

$$\min_{\psi} \sum_{i,j} \psi_{ij} M_{ij} + \gamma \psi_{ij} \log(\psi_{ij}) \text{ subject to:} \tag{15}$$

$$\sum_{i}^{m} \psi_{ij} = \mu_j \; \forall j \in \{1, 2, ..., n\} \tag{16}$$

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

$$\psi = K D_v. \tag{17}$$

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

$$\sum_{i}^{m} \psi_{ij} = v_j \sum_{i}^{m} K_{ij} = \mu_j, \tag{18}$$

which implies

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

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,

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

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

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

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 $w := \exp(x/\gamma)$.[5] 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

[5] 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.

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.

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 Appendix B provides a method for finding an approximation in a computationally efficient manner. When running the main optimization algorithm described in the next section, 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 surprisingly similar to estimates using the method described by Koenker and Mizera (2010). However, it is likely that the condition $N\sqrt{\sigma^2 + \gamma/2} \to \infty$ would not hold in this case, which is a requirement of Theorem 2. If $N\sqrt{\sigma^2 + \gamma/2} \to 0$ and we redefine the location of the vertices in the mesh, $\mathbf{a}$, to be equal to the datapoints, then asymptotically μ_i and $f_{unc,i}$ would only be influenced by a single datapoint. The following theorem shows that we can at least ensure $\hat{f} \xrightarrow{p} \mu^\star$ in the case $\gamma = \sigma = 0$.

Theorem 3: *Suppose the locations of the vertices in the mesh* $\mathbf{a}$ *are defined to be the datapoints,* $\mu^\star(z) > 0$ *on its bounded domain* $\mathcal{A}$, *and* $\mu_i = 1/N$ *for* $i \in \{1, 2, ..., N\}$. *If there exists* $\epsilon > 0$ *such that* $\int_{\mathcal{A}} |x|^{2+\epsilon} \, d\mu^\star < \infty$, $\mu^\star$ *is* $\rho-$*concave, and* $\gamma = \sigma = 0$, *then* $\hat{f} \xrightarrow{p} \mu^\star$.

Proof: The Wasserstein distance between density functions, denoted by $W_0(\nu_1(z), \nu_2(z))$, is commonly approximated by discretizing, using $\nu_{j,i} = \nu_j(z_i)$ for $j \in \{1, 2\}$ and $z_i \in \{\mathbf{a}_i\}_i$, and solving a linear programming problem, as discussed in the previous section. As long as $\min\{r \mid \mathcal{A} \subset \cup_{i=1}^m B_r(\mathbf{a}_i)\}$ goes to zero asymptotically, $W_0(\nu_1, \nu_2)$ converges to its non-discretized counterpart, $W_0(\nu_1(z), \nu_2(z))$; see for example, (Cuturi and Peyré, 2016). Since $\mu^\star(z) > 0$ for all $z \in \mathcal{A}$ and $\{\mathbf{a}_i\}_i = \{z_i\}_i$, this condition holds asymptotically, so showing $W_0(\hat{f}(z), \mu^\star(z)) \xrightarrow{p} 0$, where $\hat{f}(z)$ is the linear interpolation of $\{(\mathbf{a}_i, \hat{f}_i)\}_{i=1}^N$, will imply the result.

Let $\mu(z) = \sum_i \delta_{z_i}(z)/N$ and $f(z)$ be an arbitrary proper density with $\int_{\mathcal{A}} |x|^{2+\epsilon} \, df$. After using the triangle inequality twice, we can bound $W_0(f(\cdot), \mu(\cdot))$ by $W_0(f(\cdot), \mu^\star(\cdot))$ $\pm W_0(\mu(\cdot), \mu^\star(\cdot))$ for all $N > 1$. Note that Wasserstein distances metrize weak convergence, in the sense that $W_0(\nu_0(z), \nu_N(z)) \xrightarrow{p} 0$ if and only if the distribution corresponding to $\nu_N(z)$ converges weakly to that of $\nu_0(z)$ and the first two moments of $\nu_N(z)$ converge to the first two moments of $\nu_0(z)$; see for example, Theorem 7.12 of (Villani, 2003). Thus, $W_0(\mu(\cdot), \mu^\star(\cdot)) \xrightarrow{p} 0$ by the result of Kiefer and Wolfowitz (1956), so our bounds on $W_0(f(\cdot), \mu(\cdot))$ imply that $W_0(f(\cdot), \mu(\cdot)) \xrightarrow{p} W_0(f(\cdot), \mu^\star(\cdot))$ for all such $f(\cdot)$.

Note that $W_0(f(\cdot), \mu^\star(\cdot))$ obtains its minimum value with respect to $f(\cdot)$ at densities with distribution functions that are arbitrarily close to the distribution function of $\mu^\star(\cdot)$. Since $\mu^\star$ is $\rho-$concave by definition, and $\hat{f}(\cdot)$ is also by our constraint, this distribution has well defined and continuous density function by Aleksandrov's theorem. Thus, $\hat{f}(\cdot) \xrightarrow{p} \mu^\star(\cdot)$ by the standard argument for consistency of an M-estimator; see for example, Theorem 5.7 of (van der Vaart, 2000).

□

For the five reasons discussed in the previous section, our primary focus is on cases in which μ, and f_{unc}, are also consistent estimators, so we will move onto our recommended approach for setting $\sqrt{\sigma^2 + \gamma/2}$ and m. Like kernel density estimators, the mean squared error of f_{unc} is minimized when $\sqrt{\sigma^2 + \gamma/2}$ is $O(N^{-1/(d+4)})$. Also, 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 constraints 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.

The following theorem establishes the limiting distribution of the estimator conditional on the set of constraints that bind asymptotically. Since one feature of the proposed estimator is that the resolution of the mesh can be set independently on the sample size, the set of active constraints would converge as long as m is fixed or if it grows sufficiently slowly in the sample size. 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 4: *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 $A g_{\Omega \cup \dot{\Omega}}$ is the second difference of $g_{\Omega \cup \dot{\Omega}}$, so the binding constraints can be denoted by $A g_{\Omega \cup \dot{\Omega}} \leq \mathbf{0}$.[6] 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,

$$(x_{\Omega \cup \dot{\Omega}} - x_k) \otimes \hat{g}_{\Omega \cup \dot{\Omega}}^{1/\rho - 1} / \rho = -A^T \lambda \tag{19}$$

$$(x_{-\Omega \cap -\dot{\Omega}} - x_k) \otimes \hat{g}_{-\Omega \cap -\dot{\Omega}}^{1/\rho - 1} / \rho = \mathbf{0} \tag{20}$$

$$A \hat{g}_\Omega = \mathbf{0} \tag{21}$$

[6] 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.

$$[\ \left(\hat{g}^{1/\rho}\right)^T \quad m - \sum_i \hat{g}_i^{1/\rho}\ ]^T = \exp(x/\gamma) \otimes \left(K\left(\mu \oslash \left(K \exp(x/\gamma)\right)\right)\right), \tag{22}$$

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,

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

(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

$$\begin{aligned}\nabla_\mu h_1(x, \mu) &= -D_w K D_{\mathbf{1} \oslash (Kw)} \\ &= -D_w K D_{v \oslash \mu} = -\psi D_{\mathbf{1} \oslash \mu},\end{aligned}$$

and

$$\begin{aligned}\nabla_w h_1(x, \mu) &= D_w K D_{\mu \oslash (Kw)^2} K - D_{K(\mu \oslash (Kw))} \\ &= D_w K D_v D_{\mathbf{1} \oslash \mu} D_v K - D_{\hat{f} \oslash w} \\ &= \left(\psi D_{\mathbf{1} \oslash \mu} \psi^T - D_{\hat{f}}\right) D_{\mathbf{1} \oslash w}.\end{aligned}$$

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,

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

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

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

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

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

so

$$G_{\dot{\Omega},\cdot} := \nabla_{\mu}\hat{f}_{\dot{\Omega}} = \nabla_{x_{\Omega\cup\dot{\Omega}}-x_k}\hat{f}_{\dot{\Omega}}P\nabla_{\mu}x = -\left(D_{(x_{\dot{\Omega}}-x_k)}\hat{f}^{-\rho}_{\dot{\Omega}} + \tilde{A}D_{(x_{\Omega\cup\dot{\Omega}}-x_k)\otimes(C\hat{f}^{\rho}_{\dot{\Omega}})^{1/\rho-2}}CD_{\hat{f}^{\rho-1}_{\dot{\Omega}}}\right)^{-1} \cdot \left(B + \tilde{A}D_{(C\hat{f}^{\rho}_{\dot{\Omega}})^{1/\rho-1}}\right)P\left(\psi D_{\mathbf{1}\oslash\mu}\psi^T - D_{\hat{f}}\right)^{+}\psi D_{\mathbf{1}\oslash\mu}\gamma/(1-\rho), \tag{23}$$

where $P_{|\Omega\cup\dot{\Omega}|\times m} := \nabla_x(x_{\Omega\cup\dot{\Omega}} - x_k)$ 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,

$$G_{-\Omega\cap-\dot{\Omega},\cdot} := \nabla_{\mu}\hat{f}_{-\Omega\cap-\dot{\Omega}} = K_{-\Omega\cap-\dot{\Omega},\cdot}. \tag{24}$$

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

$$G_{\Omega,\cdot} := \nabla_{\mu}\hat{f}_{\Omega} = D_{(C\hat{f}^{\rho}_{\dot{\Omega}})^{1/\rho-1}}CD_{\hat{f}^{\rho-1}_{\dot{\Omega}}}G_{\dot{\Omega},\cdot}. \tag{25}$$

□

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 5: *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 $N(\sigma^2+\gamma/2)^{d/2}\Big(W_\gamma(\hat{f},\mu) - W_\gamma(f_{unc},\mu)\Big) \geq c_\alpha$, where c_α satisfies $P(X^TBX \geq c_\alpha) = \alpha$ 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 $N(\sigma^2+\gamma/2)^{d/2}\left\|f_{unc}-\mu^\star\right\|^3$ 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.

□

4 The Optimization Algorithm

In this section we will derive a trust region algorithm to find the global minimum of (13) and (14). 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 $\rho \neq 0$. Appendix A contains these derivations for the log $-$concave case, which corresponds to the limit as $\rho \to 0$. 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 6: *The gradient of the $W_\gamma(\mu, g^{1/\rho})$ is*

$$r := \nabla_g W_\gamma(\mu, g^{1/\rho}) = D_{g^{1/\rho - 1}/\rho}(x_{-k} - x_k), \tag{26}$$

and the Hessian is

$$H := \nabla_g^2 W_\gamma(\mu, g^{1/\rho}) = ABA^T + C, \tag{27}$$

where $A := \left[\begin{array}{cc} D_{g^{1/\rho-1}/\rho} & -g^{1/\rho-1}/\rho \end{array}\right]$, $B := \gamma(D_{g^{1/\rho}} - \psi D_{\mathbf{1}\oslash\mu}\psi^T)^+$, *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,

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

$$= D_{g^{1/\rho-1}/\rho} \nabla_g (x_{-k}(g) - x_k(g)) + \left(\nabla_g g^{1/\rho-1}/\rho\right) D_{x_{-k}-x_k}. \tag{28}$$

Next 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

$$\nabla_g f(g) = \left[\begin{array}{c} D_{g^{1/\rho-1}/\rho} \\ -g^{1/\rho-1}/\rho \end{array}\right].$$

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 $f - D_w K(\mu \oslash (Kw)) = 0$, 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 $w := \exp(x/\gamma)$ 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 $\nabla_x \tilde{x}(x) = \left[\begin{array}{cc} I & -\mathbf{1} \end{array}\right]_{m-1\times 1}$. After combining all four derivatives, we have

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

$$= \gamma D_{g^{1/\rho-1}/\rho} \left[\begin{array}{cc} I & -\mathbf{1} \end{array}\right] D_{\mathbf{1}\oslash w}(D_{f\oslash w} - \psi D_{\mathbf{1}\oslash\mu}\psi^T D_{\mathbf{1}\oslash w})^+ \left[\begin{array}{c} D_{g^{1/\rho-1}/\rho} \\ -g^{1/\rho-1}/\rho \end{array}\right]$$

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

$$= ABA^T.$$

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 to be $\arg\min_i x_i$, 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, if p is an eigenvector of $D_{\mathbf{1}\oslash f}\psi D_{\mathbf{1}\oslash\mu}\psi^T D_{\mathbf{1}\oslash f}$, then 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: 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.

As described in the second to last paragraph, we initialize the algorithm described next at a very good approximation of $\hat{f}$, which is found using the method described in Appendix B. Note that this method does not require the specification of k. Also, when the mesh is enlarged beyond the convex hull of the data, so that μ_i and $\hat{f}_i$ are generally lower when $\mathbf{a}_i$ corresponds to a point near the boundary of the domain, w_i and v_i are also generally smaller at these boundary points.

□

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

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

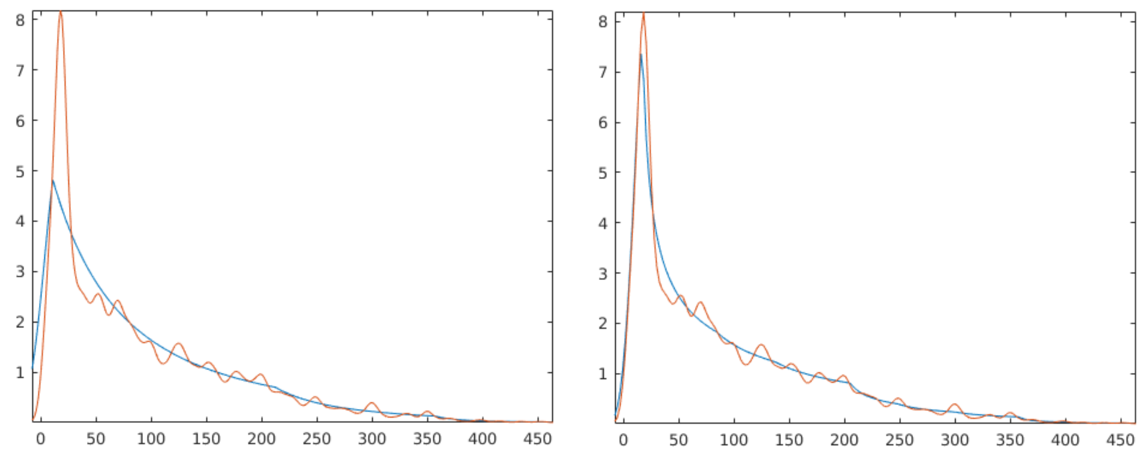
If the value of the objective function evaluated at $g_{-k}^{(i-1)} + \Delta$ results in an improvement over its value at $g_{-k}^{(i-1)}$, then $g_{-k}^{(i)}$ is defined to be $g_{-k}^{(i-1)} + \Delta$. If the improvement was significant, then the radius of the trust region, c, is increased, and otherwise it is decreased and the value of $g_{-k}^{(i)}$ is defined to be $g_{-k}^{(i-1)}$. This is described in Algorithm 2.

Algorithm 2 The parameter values for this method were set using the recommendations of Fan and Yuan (2001). `WassersteinDistance` uses Algorithm 1 to find w, and then computes $W_\gamma(\mu, f)$, the gradient, Hessian, and the index k (as defined in Theorem 6).

---

function `TrustRegion`($\mu, \rho, f^{(0)}, K$)
$(W_\gamma^{(0)}, H^{(0)}, r^{(0)}, k^{(0)}) \leftarrow$`WassersteinDistance(`$\mu, \rho, K, f^{(0)}$`)`
$g_{-k}^{(0)} \leftarrow f_{-k}^{(0)\rho}$
$\tilde{W}_\gamma^{(0)} \leftarrow W_\gamma^{(0)}$
$\eta \leftarrow 1$, $c \leftarrow 1$
for $i = 1, 2, \ldots$**:**
  $(W_\gamma^{(i)}, H^{(i)}, r^{(i)}, k^{(i)}) \leftarrow$`WassersteinDistance(`$\mu, \rho, K, f^{(i-1)}$`)`
  if $W_\gamma^{(i)} > W_\gamma^{(i-1)}$**:** $W_\gamma^{(i)} \leftarrow W_\gamma^{(i-1)}$, $H^{(i)} \leftarrow H^{(i-1)}$,$r^{(i)} \leftarrow r^{(i-1)}$, $k^{(i)} \leftarrow k^{(i-1)}$, $g \leftarrow f^{(i-2)\rho}$
  if $(W_\gamma^{(i)} - W_\gamma^{(i-1)})/\tilde{W}_\gamma^{(i-1)} < 0.25$**:** $c \leftarrow c/4 + \eta/8$
  else: $c \leftarrow 3c/2$
  $\Delta \leftarrow \arg\min_d \ d^T H d/2 + d^T r \ \text{ s.t. } \ d + g \in \mathcal{K}, \ \|d\| \leq c.$
  $\tilde{W}_\gamma^{(i)} \leftarrow \Delta^T H \Delta/2 + \Delta^T r$
  $\eta \leftarrow \|\Delta\|$
  $g \leftarrow g + \Delta$
  $g_k \leftarrow (m - \sum_i g_i^{1/\rho})^\rho$
  $f^{(i)} \leftarrow g^{1/\rho}$
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 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 algorithm's output. 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.

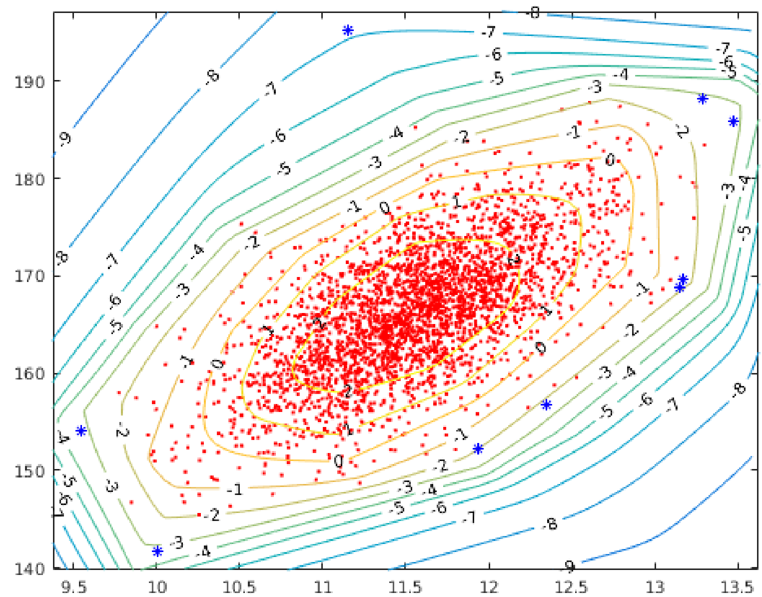


Figure 2: The data points (shown in red) consists of the height and finger length of 3,000 criminals from Macdonell (1902). The points on the convex hull of the data are illustrated with blue asterisks. The contour plot depicts the logarithm of the $-1/2-$concave density estimate to illustrate the tail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

To compute these estimates as well as the estimates given in the application section, each iteration of Algorithm 2 used MOSEK, a highly optimized quadratic program solver, to find Δ efficiently; however, this step is the limiting variable in terms of the time complexity of the algorithm. The computational efficiency of the algorithm can be improved by initializing $f^{(0)}$ at a good approximation of the final density estimate. Also, using the best available approximation of $\hat{f}$ to define $f^{(0)}$ ensures that w is as close as possible to $\mathbf{1}$, for the reasons given in Remark 1 of Theorem 2. This results in an increase in the numerical accuracy of the gradient and Hessian in Algorithm 2, and thus increased stability of the algorithm. The appendix provides a method for finding a particularly good approximation. While it does not have the same theoretical guarantees regarding convergence of a sequential quadratic programming algorithm, in many cases initializing Algorithm 2 at this input density results in convergence within two to four iterations.

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. It is worth noting that a similar sequential quadratic programming algorithm can also be used to find the global optimum in this more general setting when the derivatives corresponding to H and r exist.

5 Shape-Constraints More Generally

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,

$$\min_{g} W_\gamma(\mu, \alpha(g)) \tag{29}$$

$$\text{subject to: } \cap_i \beta_i(g) \leq 0, \tag{30}$$

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 7: *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 $A := \left[\begin{array}{cc} \nabla_g \alpha(g) & -\nabla_g \alpha(g)\mathbf{1} \end{array}\right]$, $B := \gamma(D_{\alpha(g)} - \psi D_{\mathbf{1}\oslash\mu}\psi^T)^+$, *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

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

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,

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

$$= ABA^T + C.$$

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 following Theorem 2 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 1: 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 regularity condition.[7]

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 $g(x)$ 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).

[7]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.

6 Myerson's 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

$$b_i + \frac{\hat{F}_b(b_i)}{(l-1)\hat{f}_b(b_i)}, \tag{31}$$

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 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 (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.

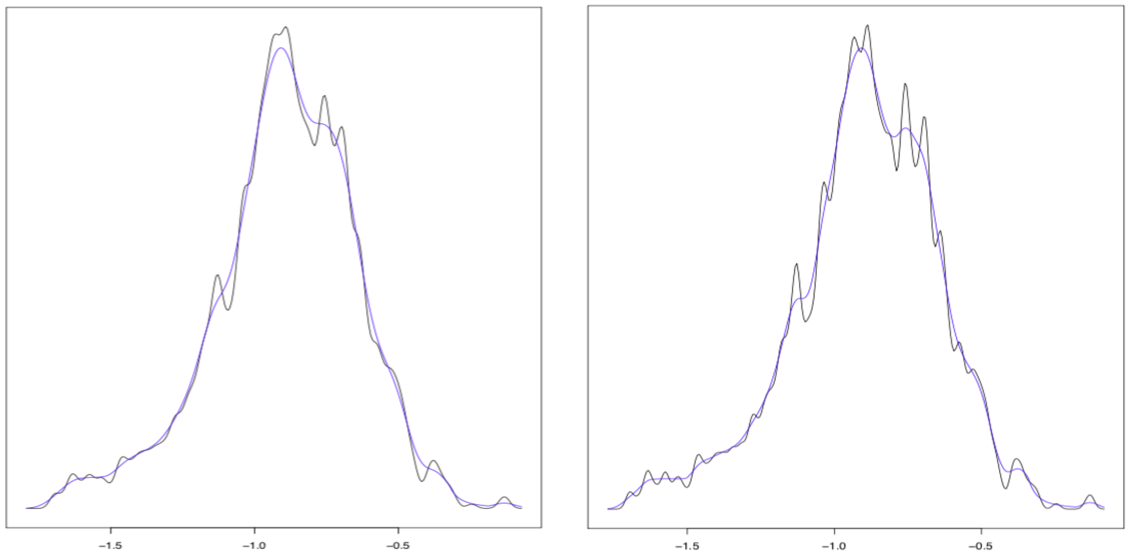
After we estimated the valuations for each firm, we estimated f_{unc} by setting $\sqrt{\gamma/2+\sigma^2}$ using Scott's (1992) rule of thumb; however, the resulting virtual valuations function was not monotonic. This could be an innocuous idiosyncrasy of the data or it could be evidence that Caltrans's 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 output of the algorithm 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.[8]

Since the density already appears parsimonious, there is little reason to decrease $\sqrt{\gamma/2+\sigma^2}$; 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 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.

[8]Myerson's 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.



(a) (b)

Figure 3: The input density and the output density, subject to Myerson's 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.

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 Algorithm 1. 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.[9] Convexity of the optimization problem in this case follows from Theorem 7.

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

Appendix A: 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 $W_\gamma(\mu, g^{1/\rho})$ is defined as

$$\max_{(x,y)\in\mathbb{R}^{2m}} x_{-k}^T \exp(g_{-k}) + x_k \left(m - \sum_i \exp(g_{-k,i})\right) + y^T\mu$$
$$-\gamma \sum_{i,j} \exp\left((x_i + y_j - M_{ij})/\gamma\right). \tag{32}$$

The index k can be chosen in the same way as described above to ensure the objective function is convex.

The gradient in this case is equal to

$$r_i := \frac{\partial W_\gamma(\mu, g^{1/\rho})}{\partial g_{-k,i}} = (x_i - x_k)\exp(g_{-k,i}), \tag{33}$$

and the Hessian is

$$H := \nabla^2 W_\gamma(\mu, g^{1/\rho}) = ABA^T + C \tag{34}$$

where

$$A := \left[\begin{array}{cc} D_{\exp(g)} & -\exp(g) \end{array}\right],$$

$$B := \gamma(D_{\exp(g)} - \psi D_{\mathbf{1}\oslash\mu}\psi^T)^{-1}, \text{ and}$$

$$C := D_{\exp(g)} \; D_{x_{-k}-x_k}.$$

Appendix B: An Approximation based on Alternating 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 explores a class of divergence measures defined by

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

where $\varphi : \mathbb{R}^m \to \mathbb{R}$ is a convex function. In other words, if $\hat{\varphi}_y(x)$ is the first order Taylor series expansion of $\varphi(\cdot)$ at $y \in \mathbb{R}^m$, 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 (1967) 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[10]

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

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

$$P_l(y) := \arg\min_{A_l x = b_l} D_{\varphi(\cdot)}(x \mid y).$$

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 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 (1967) projections of the coupling ψ onto the constraints $\psi \mathbf{1}_m = \mu_1$ and $\psi^T \mathbf{1}_m = \mu_0$. To approximate $f^\star$, we need to solve

$$\min_\psi \sum_{i,j} \psi_{ij} M_{ij} + \gamma \psi_{ij} \log(\psi_{ij}) \quad \text{subject to:} \tag{35}$$

$$\psi^T \mathbf{1}_m = \mu, \tag{36}$$

$$\psi \mathbf{1}_m = \alpha_i + \beta_i \mathrm{a}_i, \text{ and } (\alpha_i + \beta_i \mathrm{a}_i)^\rho \leq (\alpha_j + \beta_j \mathrm{a}_i)^\rho \ \ \forall\, i, j \in \{2, ..., m-1\}. \tag{37}$$

We can only guarantee that MABP converges to the global minimum if the constraints are affine. The constraint in (37) is not convex, so theory cannot provide us with a guarantee that MABP converges. 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 our 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

$$D_{W_\gamma(\cdot)}(\psi_{ij} \mid \bar{\psi}_{ij}) = \gamma \sum_{i,j} \psi_{i,j} \log\left(\frac{\psi_{ij}}{e \bar{\psi}_{ij}}\right) + \bar{\psi}_{ij}.$$

[10] 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.

As previously mentioned, the Bregman projection onto the constraint in (36) is $v \leftarrow \mu \oslash (K^T w)$. The constraints given by (37) can be combined to define the projection,

$$P_2(\bar{\psi}) := \underset{\psi,\alpha,\beta}{\arg\min} \sum_{i,j} \psi_{i,j} \log\left(\frac{\psi_{ij}}{e\bar{\psi}_{ij}}\right) \text{ subject to}$$

$$\psi \mathbf{1}_m = \alpha_i + \beta_i \mathrm{a}_i, \text{ and } (\alpha_i + \beta_i \mathrm{a}_i)^\rho \leq (\alpha_j + \beta_j \mathrm{a}_i)^\rho \;\; \forall\, i,j \in \{2,...,m-1\}. \tag{38}$$

Rather than attempting to solve (38) numerically, we can use the change of variable $f = g^{1/\rho}$, as in Section 3. The following problem has Kuhn-Tucker conditions that are equivalent to (38) but provide a reduction in dimensionality.

$$\underset{g,\alpha,\beta}{\arg\min} \sum_i g_i^{1/\rho} \log\left(\frac{g_i^{1/\rho}}{e\bar{v}_i}\right) \text{ subject to:}$$

$$g_i = \alpha_i + \beta_i \mathrm{a}_i \text{ and } \alpha_i + \beta_i \mathrm{a}_i \leq \alpha_j + \beta_j \mathrm{a}_i \;\; \forall\, i,j \in \{2,...,m-1\},$$

where $\bar{v}_i := \sum_j \bar{\psi}_{ij}$.[11]

To ensure this optimization problem is convex we need to have $(\rho-1)\log(g_i^{1/\rho}/\bar{v}_i) \leq 1$ for every $i \in \{1,2,...,m\}$. As discussed in Section 2, v and w are only unique up to a multiplicative constant, so this can easily be achieved by dividing v by $c \in \mathbb{R}$ and multiplying w by c in iterations in which this inequality may fail to hold. The pseudocode for this method is given in Algorithm 3. In this implementation we renormalize v and w whenever $\max_i\, (\rho-1)\log(g_i^{1/\rho}/\bar{v}_i) > 3/4$, and define c as $2^{sgn(\rho)}$. Generally five to thirty iterations are sufficient to provide a good initialization for Algorithm 2.

[11] Note that after v is defined using $v \leftarrow \mu \oslash (Kw)$, the equality $\psi = D_w K D_v$ from the second section implies that solving for the optimal density is equivalent to solving for the optimal value of w such that the density $D_w K v$ satisfies the shape constraint. The variable $\bar{v}$ is Kv, the component of f that is already fixed by the $v-$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 when the algorithm converges due to the assignment $v \leftarrow \mu \oslash (Kw)$, but renormalization at the end of the algorithm is necessary in the absence of convergence.

function `MABP`(μ, K, ρ)
$w \leftarrow \mathbf{1}_m$
$f \leftarrow \mu$
$c \leftarrow 2^{sgn(\rho)}$
for $i = 1, 2, \ldots$**:**
  $v \leftarrow \mu \oslash (Kw)$
  $\bar{v} \leftarrow Kv$
  if $\max_i\ (\rho - 1)\log(f_i/\bar{v}_i) > 3/4$**:** $v \leftarrow v/c,\ w \leftarrow cw$
  $g \leftarrow \arg\min_g \sum_i g_i^{1/\rho} \log\left(\frac{g_i^{1/\rho}}{e\bar{v}_i}\right)$ s.t. $g \in \mathcal{K}$
  $f \leftarrow g^{1/\rho}$
  $w \leftarrow f \oslash (Kv)$
$f \leftarrow mf/(f^T\mathbf{1})$
return f

References

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

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.

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.

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

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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

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

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

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

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

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.

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

Seregin, A., & Wellner, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Annals of statistics, 38*(6), 3751.

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. (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.

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

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Villani, C. (2003). *Topics in optimal transportation* (No. 58). American Mathematical Society.

Wellner, J. A. & Laha, N. (2017). Bi-s-concave distributions. *arXiv preprint.*