

Constrained Density Estimation using Optimal Transport

Yinan Hu and Esteban G. Tabak

January 13, 2026

Abstract

A novel framework for density estimation under expectation constraints is proposed. The framework minimizes the Wasserstein distance between the estimated density and a prior, subject to the constraints that the expected value of a set of functions adopts or exceeds given values. The framework is generalized to include regularization inequalities to mitigate the artifacts in the target measure. An annealing-like algorithm is developed to address non-smooth constraints, with its effectiveness demonstrated through both synthetic and proof-of-concept real world examples in finance.

1 Introduction

The classical optimal transport (OT) problem seeks the map that moves mass from a source to a target measure while minimizing a prescribed cost function. The objective can be formalized in either Monge’s [12] or Kantronich’s formulation [10], a convex relaxation of the former that considers transport plans instead of deterministic maps. These foundational formulations have wide-ranging applications, including to economics [7] and machine learning [14].

In many practical scenarios, the source measure is known or readily inferable from empirical data but the target measure is not explicitly specified. Instead, it is only constrained by practical requirements or expert knowledge. For example, when applying Monge’s formulation to transportation problems, the placement of the mass in the target region may be constrained to lie entirely beyond a certain boundary or within a particular region, rather than by the specification of a precise location for each fraction of the total mass. Similarly, in economic applications, supply and demand may be subject to constraints such as maximal amounts available or minimal amounts required, rather than dictated through precise marginal distributions.

Many of these constraints can be naturally formulated in terms of the expected value of specific functions. For example, in nutritional planning, the focus may be on the total daily protein intake rather than the specific food items consumed. In finance, the price of options placed on an asset must agree with their expected value under the asset’s underlying risk-free measure. Also, when the target distribution is constrained through observations, these are often presented as expected values arising from repeated measurements.

These considerations lead to developing a generalized optimal transport model which incorporates expectation-based constraints, enhancing and expanding the applicability of optimal transport in fields such as finance, statistical inference and source allocation under uncertainty.

This paper introduces and develops such framework. Starting from a source or prior distribution, we seek the target density that minimizes the Wasserstein distance to the prior while satisfying a set of specified expectation constraints.

1.1 Related Work

The most common constraints on an estimated density investigated are linear combinations of moments of the desired estimators. Examples include [9], focusing on the role of variance inflation on the data upon the restored estimator, and [8], which proposes a framework involving smoothed bootstrapping constraints. The framed proposed here addresses general equality constraints and adopts the Wasserstein distance as metric to measure the distance between the prior and estimated distributions.

In finance, the pricing of options has long been a focus of research. A generalized path-integral approach is proposed in [4] to price common exotic options formulated using the classical Black-Scholes model [3], [6] uses Gaussian process regression techniques for the learning process to accelerate the pricing of options and [11] applies the flow-based generative network RealNVP, which constructs a normalizing flow of pricing distributions, to price path-dependent exotic options. Closer to our work is [2], which modifies the maximum likelihood estimator by introducing expectation constraints as penalty functions and applies it to the pricing of several exotic options. The method proposed in this paper modifies the objective function used in [2], providing an option pricing method with geometric underlying different from maximal likelihood.

Notation	Meaning
ρ	Prior measure
μ	Estimated target measure
p_ρ	Probability density function of ρ
p_μ	Probability density function of μ
x_1, x_2, \dots	Available samples from ρ
y_1, y_2, \dots	Estimated samples from μ
$\{f_k\}, k \in \{1, 2, \dots, K\}$	Functions with constrained expected values

Table 1: Notation used and their meaning

2 Problem formulation

This section formulates the problem of constrained optimal transport. For convenience, we summarize our notation in table 1. Given a prior measure ρ in the set \mathcal{P} of Borel measures on a Polish space X , we seek the target measure $\mu \in \mathcal{P}(X)$ with minimal Wasserstein distance to ρ satisfying the expectation constraints

$$\int_X f_k(y) d\mu = \bar{f}_k, \quad (1)$$

where $f_k \in L^1(\mu)$, $k = 1, 2, \dots, K$ are integrable functions and the $\{\bar{f}_k\}$ are given constants. When ρ and μ have corresponding probability density functions, we denote them by p_ρ and p_μ .

The corresponding constrained optimal transport can be formulated in two ways, inspired by Monge and Kantorovich's formulations of the classical problem respectively.

Definition 1 (Constrained optimal transport problem, Monge's formulation). *Given a cost function $c : X \times X \rightarrow \mathbb{R}$, a prior measure ρ and a set of functions $\{f_k\}$ and real values $\{\bar{f}_k\}$, we define the constrained optimal transport problem as:*

$$\begin{aligned} \min_{\substack{T:X \rightarrow X \\ \mu \in \mathcal{P}(X)}} J_c[T] &= \int_X c(x, T(x)) d\rho(x), \\ \text{s.t. } T\#\rho &= \mu, \\ \int_X f_k(y) d\mu &= \bar{f}_k, \quad k = 1, 2, \dots, K. \end{aligned} \quad (2)$$

Definition 2 (Transportation cost, Kantorovich's formulation). *Given a cost function $c(x, y)$, a prior distribution ρ and a target distribution μ , the Kan-*

Kantorovich transportation cost between them is defined as follows:

$$I_c(\rho, \mu) = \int_{X \times Y} c(x, y) d\pi(x, y), \quad (3)$$

where

$$\pi \in \Pi(\rho, \mu) = \{\pi \in \mathcal{P}(X \times Y), \pi_x = \mu, \pi_y = \rho\}. \quad (4)$$

Definition 3 (Constrained optimal transport, Kantorovich's formulation). *The Kantorovich formulation of constrained optimal transport is defined as follows:*

$$\begin{aligned} & \min_{\mu \in \mathcal{P}(X)} I_c(\rho, \mu) \\ & \text{s.t. } \int_X f_k(y) d\pi_y(y) = \bar{f}_k, \quad k = 1, 2, \dots, K. \end{aligned} \quad (5)$$

Notice that Monge and Kantorovich formulations of constrained optimal transport problems differ from their classical optimal transport counterparts, in that the target measure itself is unknown, not only the transportation map or plan.

3 Some simple examples

In order to illustrate the problem and the nature of its solutions, we consider some simple examples for which we can write exact or semi-exact solutions, using as cost the canonical squared distance: $c(x, y) = \|x - y\|_2^2$. We start with one-dimensional examples, where $X = \mathbb{R}$. In regular one-dimensional optimal transport problems with quadratic cost, the optimal map is always monotone (see for instance chapter 2 of [13].) The following lemma shows that such monotonicity carries over to our formulation.

Lemma 1 (The optimal map is monotone). *If the optimal transport map $T^* : X \rightarrow X$ for the optimization problem (2) exists, it increases monotonically.*

Proof. The optimal solution to the problem consists of the target measure μ and the map T pushing the prior ρ to μ . The map T necessarily solves the classical optimal transport problem between ρ and μ , since the constraints involve only μ , so any other map yielding the same μ will automatically satisfy them. Then T inherits all the properties of regular OT, including its monotonicity. \square

We compare the problem's solution for each choice of $\{f_k\}$ to the solution from an alternative formulation of density estimation with constraints,

based not on optimal transport but on the Kullback–Leibler divergence between distributions. This was the choice adopted in [2] as a metric for the discrepancy between prior and pricing distributions. They estimate the target density by solving the following optimization problem:

$$\begin{aligned} & \min_{\mu \in \mathcal{P}(X)} \text{KL}(\rho \| \mu), \\ & \text{s.t. } \int_X f_k(y) d\mu = \bar{f}_k, \quad k = 1, 2, \dots, K. \end{aligned} \tag{6}$$

The Kullback–Leibler divergence between distributions is a “vertical” measure of dissimilarity, comparing the value of both distributions at each point, while the optimal transport cost provides a “horizontal” measure, which quantifies the displacement in X -space required to make both distributions identical.

3.1 Indicator function

We first consider for f the indicator function for the complement of a closed interval:

$$f(s) = \mathbf{1}_{\mathbb{R} \setminus [a,b]}(s) = \begin{cases} 1 & s \in \mathbb{R} \setminus [a,b], \\ 0 & \text{o/w,} \end{cases} \quad \text{with constraint } \bar{f} = 0, \tag{7}$$

which requires the support of the target distribution to lie within the interval $[a, b]$.

The solutions of the corresponding OT and KL-based problems are quite different. For KL, we have

$$p_\mu^{KL}(y) = \begin{cases} \frac{p_\rho(y)}{\int_a^b p_\rho(y) dy} & y \in [a, b], \\ 0 & \text{o/w,} \end{cases} \tag{8}$$

where every point within the support of the target is stretched vertically by the same amount, while for OT,

$$\mu^{OT} = \rho|_{[a,b]} + c_a \delta_a + c_b \delta_b, \tag{9}$$

where $\rho|_{[a,b]}$ refers to the restriction of the measure ρ within the interval $[a, b]$, that is, for all $A \subset \mathbb{R}$,

$$\rho|_{[a,b]}(A) = \rho(A \cap [a, b]), \tag{10}$$

and

$$c_a = \int_{-\infty}^a p_\rho(z) dz, \quad c_b = \int_b^\infty p_\rho(z) dz, \tag{11}$$

a result proved in appendix 8.1.

3.2 The RELU function

Another example has as constraint the ‘RELU’ function:

$$f(x) = \begin{cases} x - \omega & x \geq \omega, \\ 0 & \text{otherwise,} \end{cases} \quad (12)$$

used for the pricing of options with strike price ω .

Here the solution of the KL-based problem is

$$p_\mu^{KL}(y) = \begin{cases} \frac{1}{Z} p_\rho(y) e^{\gamma(y-\omega)} & y > \omega \\ \frac{1}{Z} p_\rho(y) & \text{o/w} \end{cases}, \quad (13)$$

where

$$Z = \int_{\omega}^{\infty} p_\rho(y) e^{\gamma(y-\omega)} dy + \int_{-\infty}^{\omega} p_\rho(y) dy \quad (14)$$

and γ is the unique constant satisfying

$$\frac{1}{Z} \int_{\omega}^{\infty} p_\rho(y) e^{\gamma(y-\omega)} (y - \omega) dy = \bar{f}. \quad (15)$$

By contrast, the solution to the corresponding OT problem involves a horizontal, piecewise constant shifting of elements of the prior measure: for all $A \subseteq \mathbb{R}$,

$$\mu^{OT}(A) = \rho(T_*^{-1}(A)),$$

where $T_* : X \rightarrow Y$, $T_*(x) = x + \lambda \mathbf{1}_{[x_*, \infty)}(x)$. The parameters λ, x_* are determined by the choice among the following two candidate systems that yields the lower value of the objective function.

1)

$$\begin{cases} x_* = \omega - \frac{\lambda}{2}, \\ \int_{x_*}^{\infty} (x - \omega + \lambda) p_\rho(x) dx = \bar{f}, \end{cases}$$

2)

$$\begin{cases} x_* = \omega - \lambda, \\ \int_{\omega-\lambda}^{\infty} (x - \omega + \lambda) p_\rho(x) dx = \bar{f}; \end{cases}$$

as proved in Appendix 8.2.

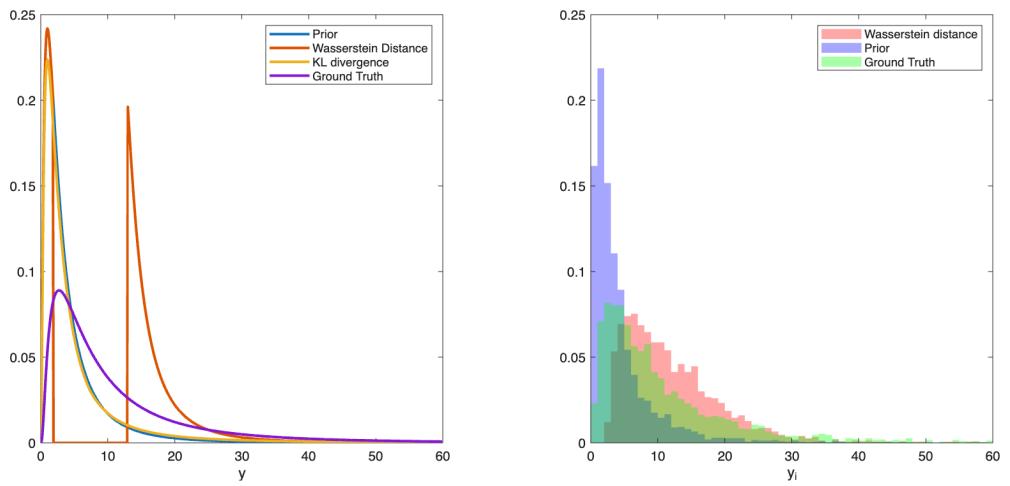


Figure 1: Left: the prior, the surrogate distribution, the target distribution restored from KL divergence and the exact target distribution restored from Wasserstein distance without smoothing; Right: the sample-based prior and the target distributions restored from Wasserstein distance with smoothing inequalities. The prior is Lognormal(1, 1). The [surrogate](#) measure is Lognormal(2, 2). Single RELU constraint (16) is adopted with $\omega = 7.3891$ and \bar{f}_k computed via (26).

3.3 Multiple RELU functions

Switching to cases of multiple expectation constraints, notice that these constraints may not be consistent, the simplest example being two constraints setting the expectation of the same function to two different values. Thus it is necessary to assume that there exists at least one distribution satisfying all of the constraints.

Assumption 1 (Assumption of consistency). *For problem (2), we assume that there exists a probability measure μ satisfying all the equality constraints:*

$$\int_Y f_k(y) d\mu = \bar{f}_k, \quad k = 1, 2, \dots, K.$$

We guarantee that this condition holds in our synthetic examples, by picking a surrogate distribution μ for the target and adopting for $\{\bar{f}_k\}$ the expected values of the $\{f_k\}$ on it.

For the RELU functions, we have

$$f_k(x) = \begin{cases} x - \omega_k & x \geq \omega_k, \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

where we assume for convenience that $\omega_1 < \omega_2 < \dots < \omega_K$. In finance, multiple RELU functions appear as constraints through the pricing of up-and-out options of an asset at different thresholding prices.

Similar to the single RELU case, the solution of the KL-based problem is

$$p_\mu^{KL}(y) = \begin{cases} \frac{1}{Z_K} p_\rho(y) & y < \omega_1, \\ \frac{1}{Z_K} p_\rho(y) e^{\sum_{i=1}^k \lambda_i (y - \omega_i)} & y \in [\omega_k, \omega_{k+1}], \\ \frac{1}{Z_K} p_\rho(y) e^{\sum_{i=1}^K \lambda_i (y - \omega_i)} & y > \omega_K, \end{cases} \quad (17)$$

where Z_K refers to the normalization coefficient:

$$Z_K = \int_{\omega_K}^{\infty} p_\rho(y) e^{\sum_{i=1}^K \lambda_i (y - \omega_i)} dy + \sum_{k=1}^{K-1} \int_{\omega_k}^{\omega_{k+1}} p_\rho(y) e^{\sum_{i=1}^k \lambda_i (y - \omega_i)} dy + \int_{-\infty}^{\omega_1} p_\rho(y) dy \quad (18)$$

and λ_k , $k = 1, 2, \dots, K$ meet the equality constraints:

$$\frac{1}{Z_K} \int_{\omega_k}^{\infty} p_\rho(y) e^{\sum_{i=1}^k \lambda_i (y - \omega_i)} (y - \omega_k) dy = \bar{f}_k, \quad k = 1, 2, \dots, K. \quad (19)$$

The solution for the OT-based problem is more complex; it can be best presented in a backward recursive way:

$$\mu^{OT}(A) = \rho(T_*^{-1}(A)), \quad (20)$$

where $T_*(x) = x + \lambda_{K*} \mathbf{1}_{[x_{K*}, \infty)}(x) + \sum_{k=1}^{K-1} \tau_{k*}(x) \mathbf{1}_{[x_{k*}, x_{k+1*}]}(x)$. Here λ_{K*}, x_{K*} are solutions of one of the following equations:

1-1)

$$\begin{cases} x_{K*} = \omega_K - \lambda_{K*}, \\ \int_{\omega_K - \lambda_{K*}}^{\infty} (z - \omega_K + \lambda_{K*}) p_\rho(z) dz = \bar{f}_K; \end{cases} \quad (21)$$

1-2)

$$\begin{cases} x_{K*} = \omega_K - \frac{\lambda_{K*}}{2}, \\ \int_{x_{K*}}^{\infty} (z - \omega_K + \lambda_{K*}) p_\rho(z) dz = \bar{f}_K. \end{cases} \quad (22)$$

And for $k = K-1, K-2, \dots, 1$,

$$\tau_k^*(x) = \begin{cases} \omega_{k+1} - x, & x \in [\omega_{k+1} - \lambda_{k*}, x_{k+1*}], \\ \lambda_{k*}, & x \in [\omega_k - \lambda_{k*}, \omega_{k+1} - \lambda_{k*}], \\ \omega_k - x, & x \in [x_{k*}, \omega_k - \lambda_{k*}], \end{cases} \quad (23)$$

Defining

$$\begin{aligned} \Delta\omega_k &= \omega_{k+1} - \omega_k; \\ H_k^+(x, \lambda) &= \int_x^{x_{k+1*}} (z - \omega_k + \lambda) p_\rho(z) dz; \\ H_k^-(x, \lambda) &= \int_x^{\omega_{k+1} - \lambda} (z - \omega_k + \lambda) p_\rho(z) dz + \Delta\omega_k \int_{\omega_{k+1} - \lambda}^{x_{k+1*}} p_\rho(z) dz; \\ \Delta\tilde{f}_k(\lambda) &= \bar{f}_k - \bar{f}_{k+1} + \Delta\omega_k \int_{x_{k+1*} - \lambda}^{\infty} p_\rho(z) dz, \end{aligned} \quad (24)$$

the parameters λ_{k*} and x_{k*} satisfy one set of the following system equations:

$$\begin{aligned} 2-1) \quad & x_{k*} = \omega_k - \frac{1}{2}\lambda_{k*}, \quad H_k^+(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-2) \quad & x_{k*} = \omega_k - \frac{1}{2}\lambda_{k*}, \quad H_k^-(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-3) \quad & x_{k*} = \omega_k - \lambda_{k*}, \quad H_k^+(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-4) \quad & x_{k*} = \omega_k - \lambda_{k*}, \quad H_k^-(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-5) \quad & \Delta\omega_k \int_{x_{k*}}^{\infty} p_\rho(z) dz = \Delta f_k, \quad \lambda_k \leq \omega_{k+1} - x_{k+1}, \end{aligned} \quad (25)$$

as proved in appendix (8.3).

In order to illustrate the different results for the KL divergence and the OT methods, we adopt as prior $p_\rho = \text{Lognormal}(0, 1)$ and as surrogate target

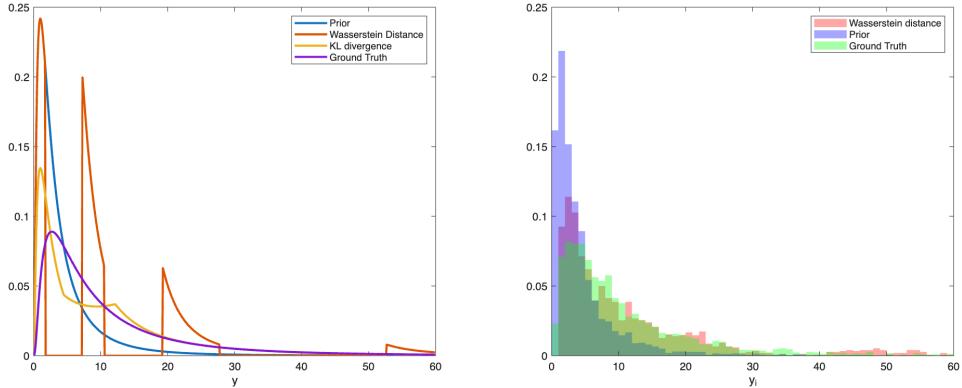


Figure 2: Left: Target distributions resulting from minimizing KL divergence as well as Wasserstein distance. The prior is Lognormal(1, 1). The surrogate measure is Lognormal(2, 2). We select $K = 3$ with \bar{f}_k ($k = 1, 2, \dots, K$) computed in (26). Right: the finite-sampled counterpart with inequality constraints. We choose $N = 2000$ samples from the prior distribution and compute the target samples.

$p_\mu^K = \text{Lognormal}(1, 1)$, under which we compute the prices of four options, $K = 5$, $\omega_k = -3 + k$, $k = 1, 2, \dots, K$ and

$$\bar{f}_k = \int_{\omega_k}^{\infty} (y - \omega_k) p_\mu^K(y) dy. \quad (26)$$

The resulting target distributions are displayed in the left of Figure 2.

We observe that the KL divergence approach tends to “stretch” the prior distribution vertically, resulting in sharp corners as artifacts, while the constrained optimal transport framework shifts samples horizontally, leading to accumulations and gaps as artifacts. In the next section, we will explore strategies to mitigate these artifacts through additional constraints.

4 Smoothing inequality constraints

The examples above make it clear that, when the functions $\{f_k\}$ are non-smooth, the optimal target distribution μ is often singular in two typical ways: assigning finite measure to small sets or having gaps, i.e. areas where μ vanishes surrounded by others where it does not. These are not artifacts: for instance, since the map with smallest cost transporting a prior distribution to a subset of \mathbb{R}^n needs not move any mass beyond the subset’s boundary, it will accumulate finite mass along it, and if only a fraction of the original mass

is required to move to said subset, then the most distant part of the prior will stay in place and a gap will form between it and the transported fraction. Yet these features may be undesirable: we may prefer the mass to be somewhat spread across the target subset, and regions with zero probability may be unrealistic models for the problem under consideration.

In order to mitigate such effects, we introduce the following two inequalities on the target probability density p_μ :

$$\int_X p_\mu^2(y) dy \leq M_\delta, \quad \int_D \frac{1}{p_\mu(y)} dy \leq M_0, \quad (27)$$

where the first inequality forbids measure accumulation in small sets and the second restricts areas of small density within a specified domain D .

5 A numerical algorithm

The previous sections established a theoretical framework for optimal transport with expectation constraints and provided several examples where explicit solutions under special choices of f_k, \bar{f}_k and p_ρ can be found. However, explicit analytical solutions for the general problem in (2) are typically not available, particularly in the presence of the smoothing constraints in (27), which are nonlinear in p_μ . Moreover, the prior ρ is often not provided in closed form but through a set of independent samples $\{x_i\}$.

In this section, we develop a numerical algorithm that uses the n samples $\{x_i\} \sim \rho$ as data and provides as output a corresponding set $\{y_i = T^*(x_i)\}$ of samples of the estimated distribution μ . We replace the prior measure ρ by its empirical counterpart $\hat{\rho}$:

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \quad (28)$$

and seek target samples $\{y_i\}_{i=1}^n$, such that the empirical target measure

$$\hat{\mu}^* = \frac{1}{n} \sum_{i=1}^n \delta_{y_i^*}, \quad (29)$$

satisfies the constraints and the empirical transport cost is minimized:

$$\min_{\{y_i\}} L = \frac{1}{n} \sum_{i=1}^n \|x_i - y_i\|^2 + \sum_{k=1}^K \lambda_k \left(\frac{1}{n} \sum_{i=1}^n f_k(y_i) - \bar{f}_k \right)^2, \quad (30)$$

where we have replaced the constraints by a quadratic penalty in the objective function with penalization parameters $\{\lambda_k\}$. This minimization problem can be solved easily through gradient descent if the functions $\{f_k\}$ are smooth and their support is broad enough (More on this below.) When they are not, as in most of our examples above, we mollify them into $\{f_k^\epsilon\}$, smooth functions with larger support that converge to the $\{f_k\}$ as $\epsilon \rightarrow 0$.

It is not solely to have well-defined derivatives of L that we mollify the $\{f_k\}$: the vanishing gradients of functions f_k within flat regions can stall the optimization. Consider as a simple example the Heaviside step function

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0, \end{cases}$$

which we can use to require a given fraction \bar{f} of the mass of μ to lie to the right of $x = 0$. The problem of using functions like this within (30) is not that $f'(0)$ is not well defined, but rather that $f'(x) = 0$ for all $x \neq 0$. Because of this, nothing in $\frac{\partial L}{\partial y_i}$ pushes any point y_i to the other side of $y = 0$, so all samples will remain at their original position $y_i = x_i$, which minimizes the transportation cost but does not satisfy the constraint. By contrast, the mollified

$$f^\epsilon(x) = \frac{1}{2} \left[\tanh\left(\frac{x}{\epsilon}\right) + 1 \right]$$

does not have this problem if ϵ is large enough for f^ϵ to be sensitive to the values of all $\{y_i\}$.

Motivated by this, we employ an annealing-like schedule. We start with a relatively large ϵ to ensure far from zero gradients across sample space. After the algorithm converges for a given ϵ , we reduce it (e.g., $\epsilon \leftarrow \epsilon / \beta$ for some $\beta > 1$) and repeat the optimization. Gradually reducing the mollification allows the solution to satisfy the constraints on the true $\{f_k\}$ without leaving sample-points behind.

Smoothing inequality constraints In order to enforce the inequality constraints in (27), we need to estimate the density function p_μ from the samples $\{y_i\}$ and evaluate the integrals via Monte Carlo. We choose to do this through kernels, writing

$$p_\mu(y) \leftarrow \frac{1}{n} \sum_{j=1}^n K(y, y_j),$$

$$\int p_\mu^2(y) dy = \int p_\mu(y) d\mu \leftarrow \frac{1}{n} \sum_{j=1}^n p_\mu(y_j),$$

$$\int_D \frac{1}{p_\mu(y)} dy = \int_Y \frac{\mathbf{1}_D}{p_\mu^2(y)} d\mu \leftarrow \frac{1}{n} \sum_{j=1}^n \frac{\varphi_\epsilon(y_j; D)}{\hat{p}_\mu^2(y_j)},$$

where $\varphi_\epsilon(\cdot; D)$ is a mollifier of the indicator function of D . For our examples, we have used Gaussian kernels K with bandwidth determined through the rule of thumb applied to the original $\{x_i\}$.

Then, using barrier functions for the inequality constraints, the full objective function becomes

$$L^{\epsilon,t} = \frac{1}{n} \sum_{i=1}^n \|x_i - y_i\|^2 + \sum_{k=1}^K \lambda_k \left(\frac{1}{n} \sum_{i=1}^n f_k^\epsilon(y_i) - \bar{f}_k \right)^2 + \frac{\lambda_\delta}{t} B_\delta + \frac{\lambda_0}{t} B_0^\epsilon, \quad (31)$$

where

$$B_\delta = -\log \left(-\frac{1}{n^2 h} \sum_{i,j} K\left(\frac{y_i - y_j}{h}\right) + M_\delta \right),$$

$$B_0^\epsilon = -\log \left(-nh^2 \sum_i \frac{\varphi_\epsilon(y_j; D)}{\left(\sum_j K\left(\frac{y_i - y_j}{h}\right)\right)^2} + M_0 \right).$$

Adaptive learning rate We use a backtrack line search to determine the step size $\eta^{(t)}$ at each iteration, ensuring sufficient decrease and robust convergence. For each update step, we begin with an initial guess of the learning rate slightly expanded from the size of the previous step: $\eta^{(t)} \leftarrow (1 + \delta)\eta^{(t-1)}$ and reduce it until the Armijo-Goldstein condition is satisfied:

$$L^\epsilon(y^{(t)} - \eta^{(t)} \nabla L^\epsilon(y^{(t)})) \leq L^\epsilon(y^{(t)}) + \alpha \eta^{(t)} \|\nabla L^\epsilon(y^{(t)})\|^2, \quad (32)$$

where $\alpha \in (0, 0.5]$ is a control parameter. This ensures that each step provides a meaningful reduction in the objective function.

We detail the algorithm for soft-constrained problems in Algorithm 1.

5.1 One-dimensional examples

The restored target measures in a sample-based setting using the Algorithm 1 are depicted on the right in Figures 1 and 2. We observe that the smoothing inequalities help mitigate gaps and accumulations by re-distributing the measures located at accumulations and spreading them out to fill in the gaps.

Algorithm 1 Solution to the soft-constrained optimization

```

1: Input: Parameters  $\{\lambda_k\}_1^K, n, \{f_k^\epsilon\}_1^K$ . Horizon  $T_{\max} \in \mathbb{Z}$ , tol.
2: Collect samples  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$  from the prior  $\rho$ .
3: Initialize:  $y^{(0)}$ . Set:  $\epsilon \leftarrow \epsilon_0$ .
4: for  $j = 1$  to  $J$  do ▷ Outer loop for annealing  $\epsilon$ 
5:   for  $t = 1$  to  $T_{\max}$  do ▷ Inner loop for gradient descent
6:     Compute gradient:  $g^{(t)} \leftarrow \nabla_y L^\epsilon(y^{(t)}; x)$ .
7:     Determine step size  $\eta^{(t)}$  using backtracking line search satisfying
       the Armijo condition.
8:     Update samples:  $y^{(t+1)} \leftarrow y^{(t)} - \eta^{(t)} g^{(t)}$ .
9:     if  $\|y^{(t+1)} - y^{(t)}\| < \text{tol}$  then
10:      break ▷ Converged for the current  $\epsilon$ 
11:    end if
12:   end for
13:   Update starting point for next round:  $y^{(0)} \leftarrow y^{(t+1)}$ .
14:   Anneal smoothing parameter:  $\epsilon \leftarrow \epsilon/\beta$ .
15: end for
16: Return: Final target samples  $y = (y_1, \dots, y_n)$ .

```

5.2 Two-dimensional examples

We apply next the constrained optimal transport framework with smoothing inequalities on 2D examples, i.e. with $X = \mathbb{R}^2$. For demonstration, we select for the equality constraints a single ($K = 1$) two-dimensional indicator function supported on the complement of a set D :

$$f(z_1, z_2) = \mathbf{1}_{\mathbb{R}^2 \setminus D} = \begin{cases} 1 & (z_1, z_2) \notin D \\ 0 & (z_1, z_2) \in D \end{cases}, \quad \bar{f} = 0. \quad (33)$$

Indicator function within a circle As a first example, $D = D_R = \{(z_1, z_2) \in \mathbb{R}^2 : z_1^2 + z_2^2 \leq R^2\}$ specifies a circle within which the support of the target measure is required to lie. The OT solution without smoothing inequalities is

$$\mu^{OT} = \rho|_{D_R} + \alpha_R \delta_{\partial D_R}, \quad \alpha_R = \iint_{z_1^2 + z_2^2 \geq R^2} f_\rho(z_1, z_2) dz_1 dz_2. \quad (34)$$

Figure 3 depicts the exact target measure (34) and the restored target measures with and without smoothing constraints.

Indicator function on a half plane: As another example, we consider the half plane represented by the set $D = D_3 = \{(z_1, z_2) : z_1 \geq R\}$. The OT solution without smoothing inequalities is

$$\mu^{OT} = \rho|_{D_3} + \alpha_3 \delta_{\partial D_3}, \quad \alpha_3 = \iint_{z_1 \leq R} f_\rho(z_1, z_2) dz_1 dz_2. \quad (35)$$

Figure 4 depicts the exact target measures and the estimated target measure with and without smoothing constraints. We observe that the unregularized estimated target measure is close to the exact measure both in terms of its continuous and its atom components, while the regularized target measure mollifies some of the target measures near the boundaries and pushes it further away.

6 Case study

In this section, we apply the density estimation framework proposed in Section 2 to a proof-of-concept problem in quantitative finance, recovering the risk-neutral pricing measure of an underlying asset given market prices of a few vanilla options. This recovered measure is then used to price exotic options, whose market prices are assumed to be unknown. We use synthetic data to validate the framework, comparing its performance against a surrogate measure and the Kullback-Leibler (KL) divergence minimization approach [2].

Background An option [5] is a contract to buy ('call') or sell ('put') a specific quantity of an underlying asset or instrument at a specified strike price on or before a specified date. Vanilla options are simple call/put options with no special or unusual features that can be exercised when the strike place performs better than the underlying asset price does. Mathematically, a vanilla call option can be expressed as the single RELU function in (12), where $x > 0$ is the asset price and $\omega > 0$ the strike price. Exotic options are more complex in terms of payment structure, expiration dates and strike prices.

The valuation of exotic options has been a research focus in finance for decades [1]. In order to estimate the value of exotic options on an underlying asset given only the pricing of its vanilla options, we proceed in two steps. We first use this article's constrained optimal transport framework in lieu of the KL divergence adopted in [2], to estimate the underlying risk-free measure, and then utilize this estimated pricing measure to price the exotic option.

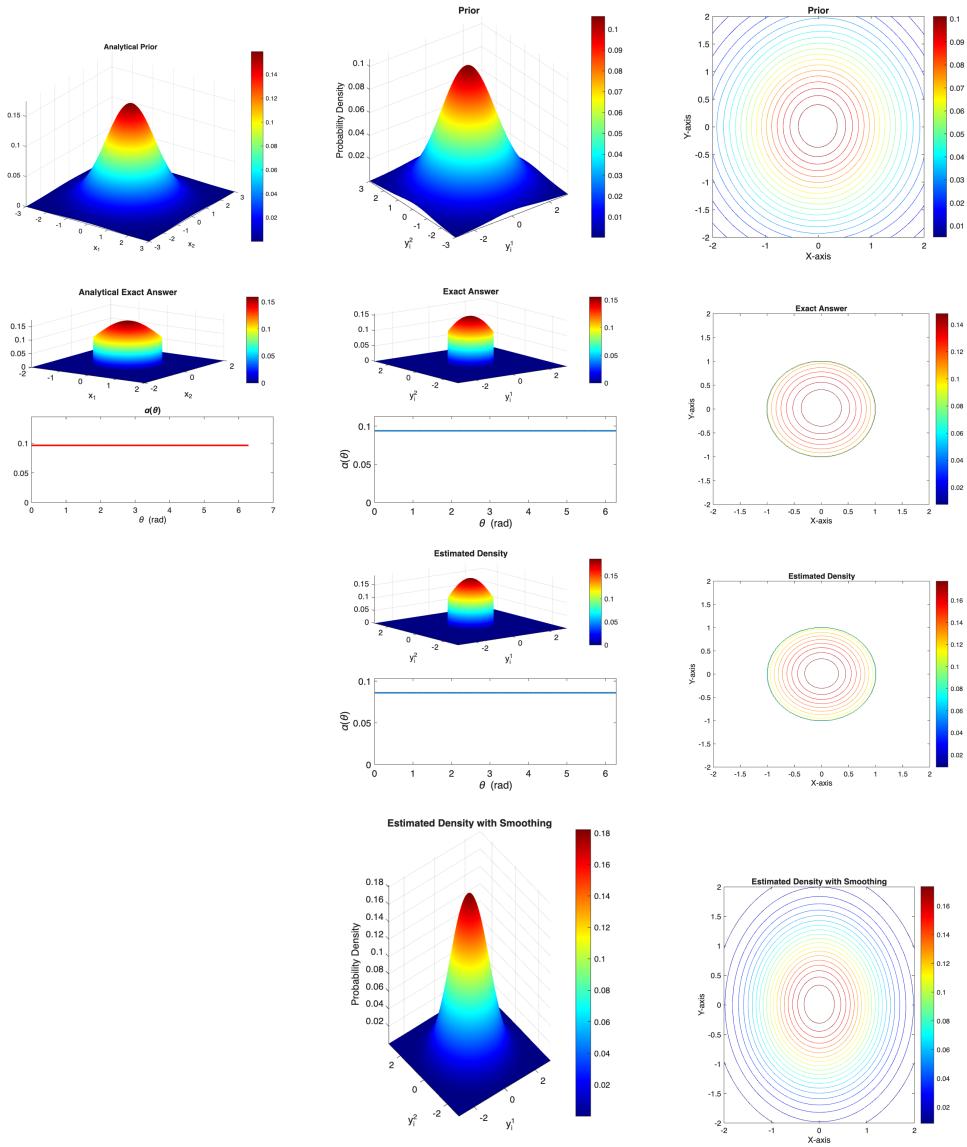


Figure 3: The surface plots and contour plots of the prior, the exact solution and the estimated target measures using the proposed framework without and with regularization. All measures are constructed using samples via kernel density estimation techniques. The exact solution and estimated measure without regularization both contain a delta measure supported on the circle $\{(z_1, z_2) | z_1^2 + z_2^2 = R^2\}$, denoted as $\alpha(\theta)\delta(R, \theta)$ and are zero outside the circle.

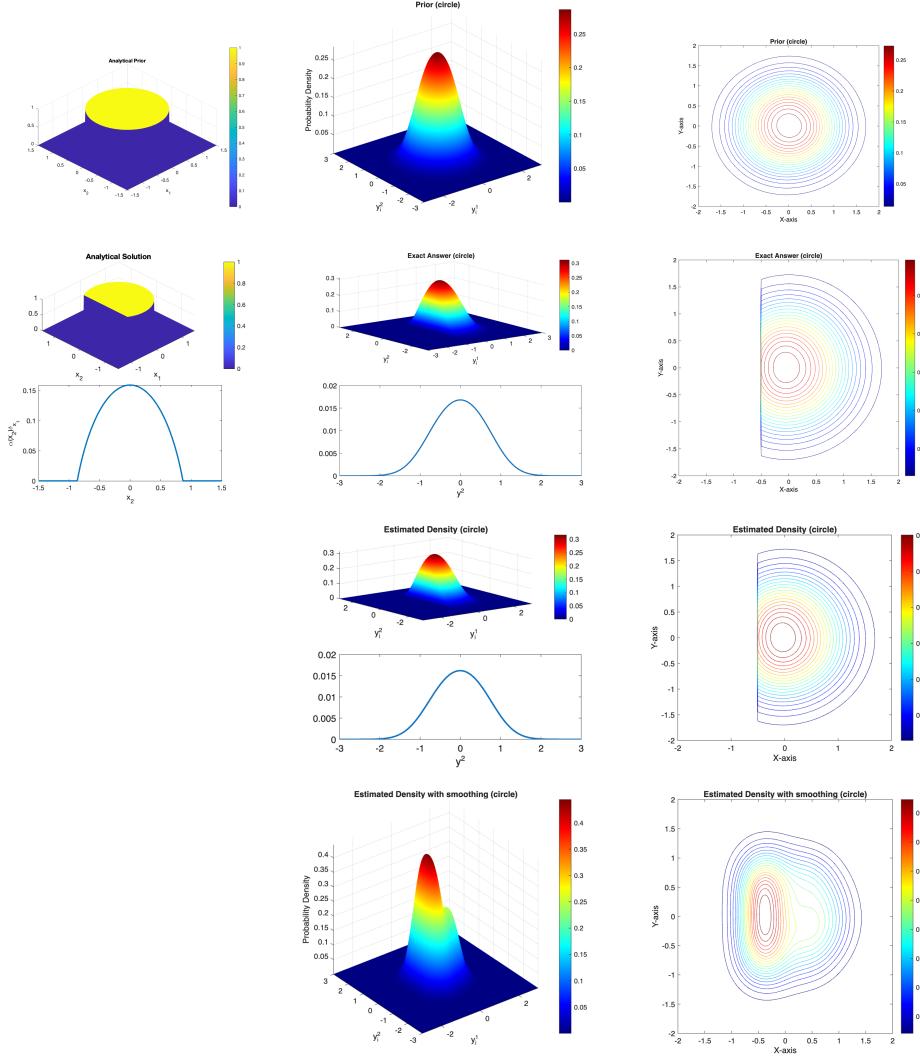


Figure 4: The surface plots and contour plots of the prior, the exact solution and the estimated target measures using the proposed framework without and with regularization. All measures are constructed using samples via kernel density estimation techniques. The exact solution and estimated measure without regularization both contain a delta measure supported along the line $\{(z_1, z_2) : z_1 = R\}$ and are zero in $\{(z_1, z_2) : z_1 < R\}$ with $R = -0.5$.

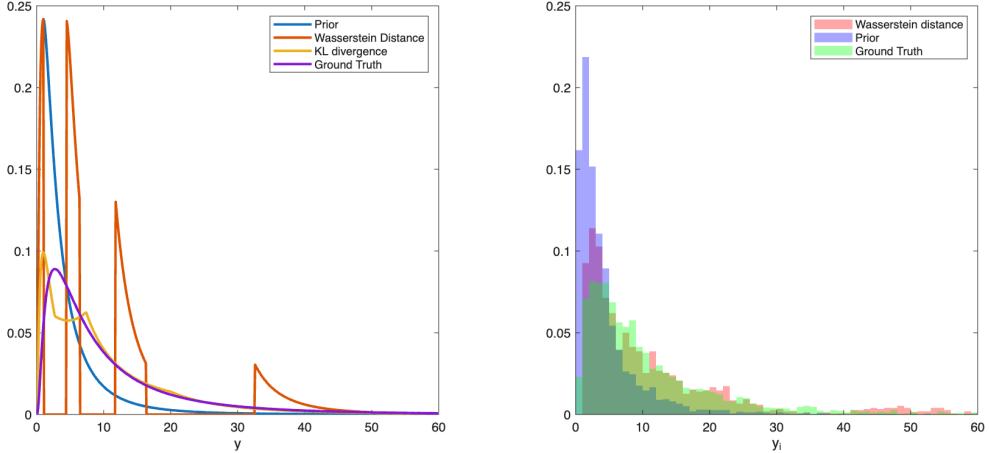


Figure 5: Left: the estimated pricing measure of the asset by the proposed framework and the surrogate measure. Right: the estimated histogram of the sampled target measure of the asset by the regularized framework, the surrogate measure and the The prior measure. The prior is p_ρ and \bar{f}_k are computed based on p_μ^K and threshold prices $\omega_1, \dots, \omega_K$ with $K = 3$.

We compare the results with their counterparts using the prior measure, the one estimated via KL divergence method and the ground-truth underlying pricing measure of the asset.

Synthetic examples We illustrate our model with several synthetic examples, where a surrogate underlying pricing measure, based on which the values of various exotic options are calculated, is assumed.

We assume that the surrogate underlying pricing measure of an asset is a log-normal distribution with parameters $\mu_K = 2, \sigma_K = 1$, or Lognormal(2,1). We are given the true pricing of $K = 3$ vanilla call options as constraints, with logarithm of threshold prices $\omega_k = -\frac{K-1}{2} + k + 1, k = 1, 2, \dots, K$. The equality constraint corresponding to the vanilla call options are given in (12) with \bar{f}_k computed based on p_μ^K : $\bar{f}_k = \int_{\omega_K}^{\infty} (y - \omega_K) p_\mu^K(y) dy$.

Choosing as prior measure of the price $p_\rho = \text{Lognormal}(1, 1)$, we depict the results generated by the KL divergence method and by our formulations in Figure 5. We can observe that our restored measure is closer to the true underlying price measure of the asset than the the one estimated by KL divergence method.

	Prior	Wasserstein KL Divergence		Smooth Wasserstein	surrogate
Up-and-out	2.7121 0.6735	10.355 5.7158	9.7641 5.3727	10.313 5.2194	10.053 5.7588
Cash-or-nothing	2.0000 2.0000	3.3446 3.3446	3.4580 3.1443	3.6720 3.1307	3.7319 3.3645
Asset-or-nothing	3.0988 2.2408	12.053 10.125	11.057 10.034	11.554 11.169	11.297 10.178

Table 2: Estimation of pricing of several exotic options using different methods. The prior $p_\rho = \text{Lognormal}(1, 1)$ and the surrogate pricing measure is Lognormal(2, 1). The three options g_1, g_2, g_3 are given in (36)(37)(38) respectively.

Pricing exotic options We now apply the estimated pricing measure to price several exotic options: down-and-out call option, cash-or-nothing option and asset-or-nothing option.

A down-and-out call option g_1 refers to an exotic option that provides a payoff only when the price lies above a certain strike price s_1 and payoff is the difference between the current price and the threshold H_0 as expressed as follows

$$g_1(x) = \max(x - s_1, 0) \mathbf{1}_{x \geq H_0}. \quad (36)$$

Here we select $H_0 = 20.0855, s_1 = 1.6487$ and $H_0 = 2.7183, s_1 = 2.1170$, respectively.

A cash-or-nothing option refers to returning a fixed cash when the pricing of the asset exceeds the strike price, or zero if it does not. With s_2 being the strike price, the price of an asset-or-nothing option g_2 can be expressed as

$$g_2(x) = C \mathbf{1}_{x \geq s_2}, \quad (37)$$

with C being the cash value. Here we select $C = 4$ and $s_2 = 2.7381, 1.6487$ respectively.

An asset-or-nothing option refers to returning the value of the asset when the pricing of the asset exceeds the strike price, or zero if it does not. With ω being the strike price, the price of an asset-or-nothing option g_3 can be expressed as

$$g_3(x) = x \mathbf{1}_{x \geq s_3}, \quad (38)$$

where we select $s_3 = 7.3891, 4.4817$ respectively.

We utilize our estimated pricing measures of the assets obtained by our method and the KL divergence method to compute pricing of the above-mentioned exotic options and compare them with the true pricing (i.e. pricing computed from the surrogate pricing of asset measures p_μ). Table 2 lists the computed prices, and Table 3 shows the relative errors of each estimate. The results in Table 3 show that all three estimation methods significantly outperform the prior, confirming that assimilating the information in the constraints is critical. Notably, a Wasserstein-based method (either the standard or smooth variant) achieves the lowest relative error in five of the six test cases. This suggests that the Wasserstein distance, as minimized by our framework, is an effective metric for regularizing the option pricing calibration problem.

Options	Prior	Wasserstein	KL Divergence	Smooth	Wasserstein
Up-and-out	-0.8831	-0.0075	-0.0670	-0.0937	
	-0.7302	0.0300	-0.0288	0.0258	
Cash-or-nothing	-0.4056	-0.0059	-0.0654	-0.0695	
	-0.4641	-0.1038	-0.0734	-0.0161	
Asset-or-nothing	-0.7799	-0.0053	-0.0141	0.0973	
	-0.7257	0.0669	-0.0212	0.0227	

Table 3: Relative error rates from the pricing of several exotic options using the estimated asset pricing measure with different frameworks: the prior measure, the proposed framework and the KL divergence framework.

7 Conclusions

In this paper, we develop a modified optimal transport framework that incorporates equality constraints to meet specific requirements. We provide both explicit analytical solutions and numerical implementations to illustrate the differences between our approach and the classical KL divergence method. Additionally, we design algorithms tailored for sample-based prior distributions.

A promising direction for future research is to consider scenarios where neither the source measure nor the target measure is explicitly specified, but only subject to certain constraints. Such situations arise frequently in operations research, where supplies and demands are only constrained rather than fully defined.

References

- [1] Marco Avellaneda. An introduction to option pricing and the mathematical theory of risk. *Probability Theory and Applications*, (6):351, 1999.
- [2] Marco Avellaneda, Craig Friedman, Richard Holmes, and Dominick Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Mathematical Finance*, 4(1):37–64, 1997.
- [3] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.
- [4] Giacomo Bormetti, Guido Montagna, N Moreni, and Oreste Nicrosini. Pricing exotic options in a path integral approach. *Quantitative Finance*, 6(1):55–66, 2006.
- [5] Mohamed Bouzoubaa and Adel Osseiran. *Exotic options and hybrids: A guide to structuring, pricing and trading*. John Wiley & Sons, 2010.
- [6] Jan De Spiegeleer, Dilip B Madan, Sofie Reyners, and Wim Schoutens. Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance*, 18(10):1635–1643, 2018.
- [7] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2016.
- [8] Peter Hall and Brett Presnell. Density estimation under constraints. *Journal of Computational and Graphical Statistics*, 8(2):259–277, 1999.
- [9] MC Jones. On correcting for variance inflation in kernel density estimation. *Computational Statistics & Data Analysis*, 11(1):3–15, 1991.
- [10] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.
- [11] Hyun-Gyo Kim, Se-Jin Kwon, Jeong-Hoon Kim, and Jeonggyu Huh. Pricing path-dependent exotic options with flow-based generative networks. *Applied Soft Computing*, 124:109049, 2022.
- [12] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

- [13] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser, 2015.
- [14] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

8 Appendix

8.1 The OT solution for the indicator function

We provide the solution for (2) when $K = 1$ and f is an indicator function (7) in the following proposition:

Proposition 1. *Let $X = Y = \mathbb{R}$ and $c(x, y) = \frac{1}{2}\|x - y\|_2^2$ be the pairwise cost for the objective function of the constrained optimization problem in (2), with $K = 1$. Let f_1 be the indicator function in (7) with $\bar{f}_1 = 0$. Assume that $\rho \in \mathcal{B}(\mathbb{R})$ is the prior measure. Then the optimal target measure $\mu^* \in \mathcal{B}(\mathbb{R})$ is given as follows:*

$$\mu^* = \rho|_{[a,b]} + c_a\delta(a) + c_b\delta(b), \quad (39)$$

where $\rho|_{[a,b]}$ refers to the restriction of the measure ρ within the interval $[a, b]$, that is, for all $A \subset \mathbb{R}$,

$$\rho|_{[a,b]}(A) = \rho(A \cap [a, b]), \quad (40)$$

$\delta(x_0)$ is the atomic measure centered at x_0 and

$$c_a = \int_{-\infty}^a p_\rho(z)dz, \quad c_b = \int_b^\infty p_\rho(z)dz. \quad (41)$$

Proof. The indicator function constraint with f_1 in (7) and $\bar{f} = 0$ requires that no mass is allowed in the target measure outside the interval $[a, b]$. Thus the least costly way to bring all mass to an interval is to leave the mass that is already within the interval in place, and bring all mass outside to the closest point in the interval, i.e. the interval's closest endpoint, which leads to the result (9). \square

8.2 Proof of proposition 2

Proposition 2. *Let (2) be the optimization problem with the function f_k expressed in (16) and $\bar{f}_k \geq 0$ (otherwise there is no optimal measure satisfying*

the equality constraint) and $k = 1$. Let $\rho \in \mathcal{B}(\mathbb{R})$ be the atomless prior measure with $\text{supp } \rho = X = \mathbb{R}$. Then we can derive the optimal target measure $\mu^* \in \mathcal{B}(\mathbb{R})$ as follows: for any $A \subset (-\infty, \infty)$,

$$\mu^*(A) = \rho(T_*^{-1}(A)), \quad (42)$$

where $T_* : X \rightarrow Y$, $T_*(x) = x + \tau(x)$, and $\tau : X \rightarrow \mathbb{R}$ is determined by parameters x_* and λ as follows:

$$\tau(x) = \begin{cases} \lambda & x \in [x_*, \infty), \\ 0 & \text{otherwise.} \end{cases} \quad (43)$$

The parameters λ, x_* are determined by solving the following two candidate systems and select the solution yielding the lower value of the objective function.

$$1) \quad \begin{cases} x_* = \omega - \lambda, \\ \int_{\omega-\lambda}^{\infty} (x - \omega + \lambda) p_\rho(x) dx = \bar{f}; \end{cases} \quad (44)$$

$$2) \quad \begin{cases} x_* = \omega - \frac{\lambda}{2}, \\ \int_{x_*}^{\infty} (x - \omega + \lambda) p_\rho(x) dx = \bar{f}, \\ \int_{\omega}^{\infty} (x - \omega) p_\rho(x) dx < \bar{f}. \end{cases} \quad (45)$$

Proof. From lemma 1 we have the following claim.

Claim 1. Suppose there exist $x_0, x_1 \in X$ such that $T(x_0) \geq \omega$, $T(x_1) \leq \omega$. Then we can find $x_* \in X$ such that

$$\begin{cases} x + \tau(x) \leq \omega & x \leq x_*, \\ x + \tau(x) > \omega & x > x_*. \end{cases} \quad (46)$$

Proof of claim 1. This is a direct result of lemma 1. \square

The previous claim helps build the following claim:

Claim 2. Let f_1 be the ‘RELU’ function as in (12). Then the equality constraint in (2) can be rewritten in terms of the source distribution as follows:

$$\int_{x_*}^{\infty} (x + \tau(x) - \omega) p_\rho(x) dx = \bar{f}. \quad (47)$$

Proof of claim 2. This is an application of the rule of lazy statistician. Using the notation $y = T(x) = x + \tau(x)$ as well as the theorem $\mathbb{E}_{X \sim f_X(x)}[g(X)] = \int g(x) f_X(x) dx$, where $g(x) = (x + \tau(x) - \omega)_+$.

By lemma 1 we know $(x + \tau(x) - \omega) \mathbf{1}_{x \geq x_*} = (y - \omega) \mathbf{1}_{y \geq \omega} =: g \circ T^{-1}(y)$. Thus $E_X[g(X)] = E_Y[g \circ T^{-1}(Y)]$, the right hand side of which corresponds to the RHS of the equation (47). \square

The previous claim substitutes the the original equality constraint in the optimization problem (2) with (47) equivalently.

Regarding the shifting on $[-\infty, x_*]$, we have

Claim 3. *For $x \leq x_*$ the optimal shift function*

$$\tau^*(x) = 0. \quad (48)$$

Proof of claim 3. By claim 1 for $x \leq x_*$, we have $x + \tau(x) \leq \omega$ so $(x + \tau(x) - \omega)_+ = 0$. By the expression of f in (16), when $x \leq x_*$ the term $x + \tau(x)$ contributes nothing to the expectation constraint (2) for $K = 1$ no matter how large $\tau(x)$ is, yet $|\tau(x)|^2$ is always nonnegative. \square

When $x > x_*$ (or $x + \tau(x) > \omega$), we denote $\iota(x) = \tau(x) - (\omega - x)$. Thus $\iota(x) \geq 0$. Then the optimization problem is converted into

$$\begin{aligned} & \min_{\substack{\iota(x), x_* \\ \iota(x) \geq 0}} \frac{1}{2} \int_{x_*}^{\infty} |\iota(x) + \omega - x|^2 p_{\rho}(x) dx, \\ & \text{s.t. } \int_{x_*}^{\infty} \iota(x) p_{\rho}(x) dx = \bar{f}. \end{aligned} \quad (49)$$

We associate a Lagrange multiplier $u : X \rightarrow \mathbb{R}$, $u \geq 0$ to the inequality $\iota \geq 0$. The (partial) corresponding Lagrangian \mathcal{L} is

$$\mathcal{L}(\iota, \lambda, u) = \int_{x_*}^{\infty} \left[\frac{1}{2} |\iota(x) + \omega - x|^2 - (u(x) + \lambda) \iota(x) \right] p_{\rho}(x) dx + \lambda \bar{f}. \quad (50)$$

So taking the first-order condition with respect to ι (variational derivative), we get

$$0 \equiv \frac{\delta \mathcal{L}}{\delta \iota} = (\iota(x) + \omega - x) - u(x) - \lambda. \quad (51)$$

Considering also the complementary slackness condition

$$u(x)\iota(x) = 0, \quad \forall x \geq x_*. \quad (52)$$

From (52) we know for every $x \geq x_*$, either $u(x) = 0$ or $\iota(x) = 0$ or both. We convert the optimization problem over all τ (or equivalently, ι) into the one over all $x_* \in \mathbb{R}$ and $\tau(x) \mathbf{1}_{x \geq x_*}$. Here we discuss by cases.

Case 1: $u(x) = 0$ Then $\iota(x) > 0$ strictly (from the monotone assumption, we already know $x + \tau(x) \geq \omega$ and now the complementary slackness condition implies under this condition $\iota(x) \neq 0$, but according to (51) the sum is zero, so ι must be positive). And we further know $\iota(x) = x - \omega + \lambda$, which, together with the strict positivity of ι , implies $x > \omega - \lambda$.

Case 2: $\iota(x) = 0$. Then $u(x) > 0$ strictly, and $u(x) = \omega - x - \lambda$, which implies $x < \omega - \lambda$.

In short, we can summarize the expression of ι_* when $x \geq x_*$ as

$$\iota_*(x) = \begin{cases} x - \omega + \lambda & x > \omega - \lambda \\ 0 & \text{o/w.} \end{cases} \quad (53)$$

Regarding the relationship between x_* and $\omega - \lambda$ there are two choices: either $x_* > \omega - \lambda$ strictly or $x_* \leq \omega - \lambda$.

a) Suppose the following holds true:

$$x_* > \omega - \lambda(x_*). \quad (54)$$

We know that for all $x > x^*$ condition $x > \omega - \lambda$ is satisfied, which leads to the expression of ι as in case 1. The complementary slackness conditions still apply. Notice λ is dependent on x_* via the following equality constraint:

$$H(x_*, \lambda) := \int_{x_*}^{\infty} (x - \omega + \lambda)p_{\rho}(x)dx = \bar{f}. \quad (55)$$

With a little abuse of notation we here use $\lambda(x_*)$ to stress such dependency. Accordingly,

$$\tau^*(x) = \begin{cases} \lambda(x_*) & x > x_*, \\ 0 & \text{o/w.} \end{cases} \quad (56)$$

Applying first-order condition on it gives:

$$x_* = \omega - \frac{1}{2}\lambda(x_*). \quad (57)$$

Equations (56), (57) and (55) provide a candidate solution for τ^* . We end up with the target distribution, which is split into two intervals $[x_* + \lambda(x_*), \infty)$ and $(-\infty, x_*]$.

In the meantime, to assure that x_* in (57) satisfies the condition (54), there shall exist such $\lambda > 0$ that

$$Q_0(\lambda) := \int_{\omega - \frac{\lambda}{2}}^{\infty} (x - \omega + \lambda)p_{\rho}(x)dx = \bar{f}. \quad (58)$$

Since ρ is atomless, p_ρ is continuous and so is Q_0 . Considering $Q_0(\lambda) \rightarrow \infty$ as $\lambda \rightarrow \infty$, the existence of such a $\lambda > 0$ is equivalent to $Q_0(0) := \int_{\omega}^{\infty} (x - \omega)p_\rho(x)dx < \bar{f}$.

b) If on the other hand

$$x_* \leq \omega - \lambda,$$

we know for all $x > x_*$, either $x_* < x < \omega - \lambda$ or $x > \omega - \lambda$. Thus by discussion of complementary slackness we know

$$\iota(x) = \begin{cases} x - \omega + \lambda, & x > \omega - \lambda, \\ 0, & x_* < x < \omega - \lambda, \end{cases} \quad (59)$$

Substituting the expression of $\iota(x)$ in (59) into the equality constraint, we get

$$H^-(\lambda) := \int_{\omega-\lambda}^{\infty} (x - \omega + \lambda)p_\rho(x)dx =: \bar{f}, \quad (60)$$

which only concerns λ . Since ρ is atomless and the left hand side of (60) is continuous and increasing in terms of λ , Thus the solution to (60) exists, which is denoted as λ_* . Now we aim to find out x_* . From the monotone assumption (46) we know an inequality constraint from the the expression of τ^* in (61):

$$\tau^*(x) = \begin{cases} \lambda_*, & x > \omega - \lambda_*, \\ \omega - x, & x_* < x < \omega - \lambda_*, \\ 0, & x < x_*, \end{cases} \quad (61)$$

where λ_* can be determined through (60). Using first-order condition we get

$$x_* = \omega - \lambda_*. \quad (62)$$

Summarizing the results of (62) and (61), we can now know the optimal shifting function τ^* as.

$$\tau^*(x) = \begin{cases} \lambda_* & x \geq \omega - \lambda_* = x_*, \\ 0 & x < \omega - \lambda_*. \end{cases} \quad (63)$$

It is clear to see that for both situations a) and b) we can always find proper λ and x_* . So the best strategy is to compute the values of the objective function solved by both ways and select the way leading to a lower value. \square

Remark: If $\text{supp } p_\rho = (a, b]$ instead of \mathbb{R} , the two candidates for x_* , λ still hold, provided that $x_* \in (a, b]$. In addition, there are two more candidate systems when $x_* = a$, $x_* = b$ and the corresponding λ are computed using (55).

Remark: The proof above carries over to the situation where ρ is a discrete measure by replacing integrals with finite sums. Specifically, if ρ is supported on finite points: $\{x_{(i)}\}_i$, that is,

$$\rho = \sum_i c_i \delta(x_{(i)}). \quad (64)$$

Then the candidate systems are

$$1) x_* = \omega - \lambda, \quad \sum_{i:x_{(i)} > x_*} c_i (x_{(i)} - \omega + \lambda) = \bar{f}; \quad (65)$$

$$2) x_* = \omega - \frac{\lambda}{2}, \quad \sum_{i:x_{(i)} > x_*} c_i (x_{(i)} - \omega + \lambda) = \bar{f}. \quad (66)$$

8.3 Proof of proposition 3

Proposition 3. Let (2) be the optimization with the function f_k being the class of ‘RELU’ functions in (16) and $\bar{f}_k \geq 0$ (otherwise there is no optimal measure satisfying the equality constraint). Then we can derive the optimal target measure μ^* as follows: for any $A \subset (-\infty, \infty)$,

$$\mu^*(A) = \rho(A - \tau_0), \quad (67)$$

where

$$\tau_0 = \begin{cases} 0 & x < x_{1*}, \\ \tau_{k*} & x \in [x_{k*}, x_{k+1*}], \quad k = 1, 2, \dots, K-1 \\ \tau_{K*} & x > x_{K*}. \end{cases} \quad (68)$$

where τ_{k*}, x_{k*} are to be specified in the following recursive way:

1) When $k = K$, we have

$$\tau_K = \begin{cases} \lambda_{K*} & x \in [x_{K*} + \lambda_{K*}, \infty], \\ 0 & x \in [x_{K*}, x_{K*} + \lambda_{K*}], \end{cases} \quad (69)$$

where λ_{K*}, x_{K*} are solutions of one of the following equations:

1-1)

$$\begin{cases} x_{K*} = \omega_K - \lambda_{K*}, \\ \int_{\omega_K - \lambda_{K*}}^{\infty} (z - \omega_K + \lambda_{K*}) p_\rho(z) dz = \bar{f}_K; \end{cases} \quad (70)$$

1-2)

$$\begin{cases} x_{K*} = \omega_K - \frac{\lambda_{K*}}{2}, \\ \int_{x_{K*}}^{\infty} (z - \omega_K + \lambda_{K*}) p_{\rho}(z) dz = \bar{f}_K. \end{cases} \quad (71)$$

2)

$$\tau_k^*(x) = \begin{cases} \omega_{k+1} - x, & x \in [\omega_{k+1} - \lambda_{k*}, x_{k+1*}], \\ \lambda_{k*}, & x \in [\omega_k - \lambda_{k*}, \omega_{k+1} - \lambda_{k*}], \\ \omega_k - x, & x \in [x_{k*}, \omega_k - \lambda_{k*}], \end{cases} \quad (72)$$

For each $k \in [K-1]$. For simplicity, we define

$$\begin{aligned} \Delta\omega_k &= \omega_{k+1} - \omega_k; \\ H_k^+(x, \lambda) &= \int_x^{x_{k+1*}} (z - \omega_k + \lambda) p_{\rho}(z) dz; \\ H_k^-(x, \lambda) &= \int_x^{\omega_{k+1} - \lambda} (z - \omega_k + \lambda) p_{\rho}(z) dz + \Delta\omega_k \int_{\omega_{k+1} - \lambda}^{x_{k+1*}} p_{\rho}(z) dz; \\ \Delta\tilde{f}_k(\lambda) &= \bar{f}_k - \bar{f}_{k+1} + \Delta\omega_k \int_{x_{k+1*} - \lambda}^{\infty} p_{\rho}(z) dz. \end{aligned} \quad (73)$$

Then λ_{k*} and x_{k*} satisfying one set of the following system equations:

$$\begin{aligned} 2-1) \quad x_{k*} &= \omega_k - \frac{1}{2}\lambda_{k*}, \quad H_k^+(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-2) \quad x_{k*} &= \omega_k - \frac{1}{2}\lambda_{k*}, \quad H_k^-(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-3) \quad x_{k*} &= \omega_k - \lambda_{k*}, \quad H_k^+(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-4) \quad x_{k*} &= \omega_k - \lambda_{k*}, \quad H_k^-(x_k, \lambda_k) = \Delta\tilde{f}_k; \\ 2-5) \quad \lambda_k &\leq \omega_{k+1} - x_{k+1}, \quad \Delta\omega_k \int_{x_{k*}}^{\infty} p_{\rho}(z) dz = \Delta f_k, . \end{aligned} \quad (74)$$

Proof. Similar to the proof of proposition 2, we denote $y = x + \tau(x)$. As an extension of claim 1 in the previous proof, we could find $x_{1*} \leq x_{2*} \leq \dots \leq x_{K*}$ such that

$$\begin{cases} x + \tau(x) \leq \omega_k & x < x_{k*}, \\ x + \tau(x) \geq \omega_k & x \geq x_{k*}, k = 1, 2, \dots, K. \end{cases} \quad (75)$$

Inspired by the K inequalities (75), we can decompose the real line into K intervals: $(\infty, x_{1*}], (x_{1*}, x_{2*}], \dots, (x_{K*}, \infty)$, where $x_{1*} \leq x_{2*} \leq \dots$ are to be

computed later. We can rewrite the K equality constraints by decompose the integrals on the LHS of the constraints into the ones on intervals:

$$\min_{\tau} \int_{-\infty}^{x_1} |\tau(x)|^2 f_\rho(x) dx + \sum_{k=1}^{K-1} \int_{x_{k*}}^{x_{k+1*}} |\tau(x)|^2 d\rho + \int_{x_{K*}}^{\infty} |\tau(x)|^2 d\rho \quad (76)$$

$$\text{s.t. } \int_{x_{K*}}^{\infty} \tau(x) d\rho = \bar{f}_K + (1 - c_{K*}) \omega_K - s_{K*}, \quad (77)$$

$$\int_{x_{k*}}^{x_{k+1*}} \tau(x) d\rho = \bar{f}_k - \bar{f}_{k+1} + (1 - c_{k*}) \omega_k - (1 - c_{k+1*}) \omega_{k+1*} \quad (78)$$

$$- (s_{k*} - s_{k+1*}), \quad k \in [K-1], \quad (79)$$

where we denote

$$c_{k*} = \int_{-\infty}^{x_{k*}} p_\rho(x) dx, \quad s_{k*} = \int_{x_{k*}}^{\infty} x p_\rho(x) dx. \quad (80)$$

We can solve the general case in a reverse dynamic programming way. That is, we begin with $k = K$, solving τ_K and x_{K*} , and then go backwards until $k = 1$.

Step 1: When $k = K$, we aim to solve the following problem:

$$\begin{aligned} & \min_{\substack{\iota_K(x), x_{K*} \\ \iota_K(x) \geq 0}} \int_{x_{K*}}^{\infty} |\iota_K(x) + \omega_K - x|^2 d\rho, \\ & \text{s.t. } \int_{x_{K*}}^{\infty} \iota_K(x) d\rho = \bar{f}_K. \end{aligned} \quad (81)$$

The problem (81) can be solved by the same way as discussed in the previous paragraph to find $\iota_K(x), x_{K*}$.

Step 2: When $k = 1, 2, \dots, K-1$, we aim to find the optimal τ^* when $x \in [x_{k*}, x_{k+1*}]$.

Having solved the case $k = K$, we substitute x_{K*} into (82) and obtain the optimal solution τ_{K-1*} parameterized by x_{K-1*}, λ_{K-1} , which could then be used in (82) to solve τ_{K-2*} , etc. Let $\iota_k(x) = \tau(x) - (\omega_k - x)$ for $k = 1, 2, \dots, K$ represent the shifting ι restricted within the interval $[x_{k*}, x_{k+1*}]$. From the monotone mp assumption we know $\iota_k(x) \geq 0$. Unlike the case of a single constraint, we also know that $\tau(x) - (\omega_{k+1} - x) \leq 0$, implying $\iota_k(x) \leq \omega_{k+1} - \omega_k$ within the interval $[x_{k*}, x_{k+1*}]$.

Inspired by the lemma, we know the optimal shift function τ^* is a constant in the region $[x_{k*}, x_{k+1*}]$, can be denoted as τ_k^* . Then the optimization problem is converted into

$$\begin{aligned} & \min_{\substack{\iota_k(x), x_{k*} \\ 0 \leq \iota_k(x) \\ \iota_k(x) \leq \omega_{k+1} - \omega_k}} \int_{x_{k*}}^{x_{k+1*}} |\iota_k(x) + \omega_k - x|^2 d\rho, \\ & \text{s.t. } \int_{x_{k*}}^{x_{k+1*}} \iota_k(x) d\rho = \bar{f}_k - \bar{f}_{k+1} - \int_{x_{k+1*}}^{\infty} (\omega_{k+1} - \omega_k) p_{\rho}(x) dx, \\ & \quad k = 1, 2, \dots, K-1. \end{aligned} \tag{82}$$

We introduce another Lagrange multiplier $v_k \geq 0$ to pair with the extra inequality constraint $\iota_k(x) \leq \omega_{k+1} - \omega_k =: \Delta\omega_k$. Also for brevity we denote

$$\Delta\tilde{f}_k = \bar{f}_k - \bar{f}_{k+1} - \int_{x_{k+1*}}^{\infty} (\omega_{k+1} - \omega_k) p_{\rho}(x) dx \tag{83}$$

We denote $u_k(x)$ as the Lagrange multiplier for the inequality constraint $\iota_k(x) \geq 0$. The KKT condition as follows:

$$\iota_k(x) + \omega_k - x - u_k(x) + v_k(x) - \lambda_k \equiv 0, \tag{84}$$

$$u_k(x)\iota_k(x) = 0, \tag{85}$$

$$v_k(x)(\iota_k(x) - \Delta\omega_k) = 0. \tag{86}$$

1) If $\iota_k(x) < \Delta\omega_k$ strictly, the last condition is active and thus $v_k(x) = 0$. In this situation the previous discussions on the case $K = 1$ an carry over here:

1a) If $\iota_k(x) > 0$ strictly, $u_k(x) = 0$ and we get $\iota_k(x) = x - \omega_k + \lambda_k$. The strict positivity of ι_k implies this could happen when $x > \omega_k - \lambda_k$. Considering $\iota_k < \Delta\omega_k$ strictly, we revise the region into $\omega_k - \lambda_k < x < \omega_{k+1} - \lambda_k$.

1b) If $\iota_k(x) = 0$, we know $u_k(x) > 0$ strictly. Together we have known $v_k(x) = 0$. Substituting choices of ι_k, u_k, v_k into the first-order equation, we know $x = \omega_k - \lambda_k - u_k(x)$, implying this happens when $x < \omega_k - \lambda_k$.

2) If $\iota_k(x) = \Delta\omega_k$ for some $x \in [x_{k*}, x_{k+1*}]$, we know $v_k(x) > 0$ strictly and $u_k(x) = 0$ (since $\iota_k(x) > 0$ strictly as well). As a result, we conclude $v_k(x) = \lambda_k + x - \omega_{k+1} > 0$ and thus $x > \omega_{k+1} - \lambda_k$. In this situation $\tau(x) = \omega_{k+1} - x$.

To summarize, we can derive $\iota_k(x)$ as follows:

$$\iota_k(x) = \begin{cases} 0, & x < \omega_k - \lambda_k, \\ x - \omega_k + \lambda_k, & \omega_k - \lambda_k < x < \omega_{k+1} - \lambda_k, \\ \Delta\omega_k, & x > \omega_{k+1} - \lambda_k. \end{cases} \tag{87}$$

The previous expression (87) specifies the optimal shift $\tau_k^*(x)$ within $[x_{k*}, x_{k+1*}]$ (or equivalently, $\iota_k(x)$) when x belongs to different sub-regions. However, the relationship among $x_{k*}, x_{k+1*}, \omega_k - \lambda_k, \omega_{k+1} - \lambda_k$ remains unknown (except we assume $x_{k*} \leq x_{k+1*}$ and we know $\omega_k \leq \omega_{k+1}$). There are in total six cases below: For different cases, we could rewrite the objective function $J_k = J|_{[x_{k*}, x_{k+1*}]}$ into at most three terms.

- 0) If $x_{k*} \leq x_{k+1*} \leq \omega_k - \lambda_k \leq \omega_{k+1} - \lambda_k$.
- 1) If $x_{k*} < \omega_k - \lambda_k < x_{k+1*} < \omega_{k+1} - \lambda_k$,
- 2) If $x_{k*} \leq \omega_k - \lambda_k \leq \omega_{k+1} - \lambda_k \leq x_{k+1*}$
- 3) If $\omega_k - \lambda_k < x_{k*} < \omega_{k+1} - \lambda_k < x_{k+1*}$,
- 4) If $\omega_k - \lambda_k < x_{k*} < x_{k+1*} < \omega_{k+1} - \lambda_k$,
- 5) $x_{k+1*} \geq x_{k*} \geq \omega_{k+1} - \lambda_k > \omega_k - \lambda_k$

Step 3: For $x \leq x_{1*}$ the optimal shift function

$$\tau^*(x) = 0, \quad (88)$$

as is similar from proposition. \square