DATA DENOISING WITH SELF CONSISTENCY, VARIANCE MAXIMIZATION, AND THE KANTOROVICH DOMINANCE

BY JOSHUA ZOEN-GIT HIEW[1,b], TONGSEOK LIM[*2,c] [ID],
BRENDAN PASS[1,a], AND MARCELO CRUZ DE SOUZA[3,d]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton AB Canada,* [a]*pass@ualberta.ca;*
[b]*joshuazo@ualberta.ca*

[2]*Mitch Daniels School of Business, Purdue University, West Lafayette, IN 47907, USA ,* [c]*lim336@purdue.edu*

[3]*Financial System Monitoring Department, Central Bank of Brazil, Fortaleza, CE, Brazil ,* [d]*marcelo.souza@bcb.gov.br*

We introduce a new framework for data denoising, partially inspired by martingale optimal transport. For a given noisy distribution (the data), our approach involves finding the closest distribution to it among all distributions which 1) have a particular prescribed structure (expressed by requiring they lie in a particular domain), and 2) are self-consistent with the data. We show that this amounts to maximizing the variance among measures in the domain which are dominated in convex order by the data. For particular choices of the domain, this problem and a relaxed version of it, in which the self-consistency condition is removed, are intimately related to various classical approaches to denoising. We prove that our general problem has certain desirable features: solutions exist under mild assumptions, have certain robustness properties, and, for very simple domains, coincide with solutions to the relaxed problem.

We also introduce a novel relationship between distributions, termed Kantorovich dominance, which retains certain aspects of the convex order while being a weaker, more robust, and easier-to-verify condition. Building on this, we propose and analyze a new denoising problem by substituting the convex order in the previously described framework with Kantorovich dominance. We demonstrate that this revised problem shares some characteristics with the full convex order problem but offers enhanced stability, greater computational efficiency, and, in specific domains, more meaningful solutions. Finally, we present simple numerical examples illustrating solutions for both the full convex order problem and the Kantorovich dominance problem.

1. Introduction. Denoising or dimensional reduction is a central problem in statistics and machine learning [6], [17], consisting of inferring an unknown probability distribution μ (the true data or signal) from an observed distribution ν (the noisy data).

Given a structured domain $\mathcal{D}$ of Borel probability measures to which the signal is assumed to belong, there are at least two distinct approaches to denoising. The first involves finding the $\mu \in \mathcal{D}$ which is *closest to the data ν* in an appropriate sense. When the distance between μ and ν is measured by the Wasserstein metric ((3) below), this corresponds to the *relaxed problem* in our nomenclature here ((8) below) and, for different choices of $\mathcal{D}$, encompasses principal components, k-means clustering, and the version of principal curves in [17].

A second approach involves identifying a $\mu \in \mathcal{D}$ which *lies in the middle of ν* in a certain sense. A precise formulation of this notion is known as self-consistency in the statistics literature and was first formulated by Hastie and Stuetzle in their seminal work [12] introducing principal curves. While self-consistency (expressed in our work here as the existence

---

of a martingale coupling $\pi \in \mathcal{M}(\mu, \nu)$ defined in (6) below) can naturally be interpreted as, conditional on the signal being X, the average of the noise around X vanishing, it also arises (roughly speaking) as the first order variation of the distance function from the data [12]. Despite this connection, solutions obtained from the self-consistency approach do not even locally generally minimize the distance to the data among $\mu \in \mathcal{D}$ [10], and, to the best of our knowledge, there is not an existing general paradigm for data denoising capturing key features from both of these two approaches.

Our first contribution here is to propose such a framework, by exploiting ideas from the theory of martingale optimal transport (MOT) [2], [13], [8]. Heuristically, this problem amounts to finding the closest μ to ν *among those $\mu \in \mathcal{D}$ which can be coupled to ν in a self-consistent way*; more precisely, it is to minimize

$$(1) \qquad \min_{\mu \in \mathcal{D}} \min_{\pi \in \mathcal{M}(\mu, \nu)} \int |x - y|^2 d\pi(x, y).$$

We note that this can be interpreted as a backwards martingale optimal transport problem somewhat reminiscent of the one in [18]. Since by Strassen's theorem, the measures μ for which a self-consistent (or martingale, in the language of MOT) coupling to ν exist are precisely those which are less than ν in convex order, denoted by $\mu \preceq_C \nu$, and the distance between measures is measured by minimizing the expected squared distance among such couplings, it is straightforward to show that this problem is in fact equivalent to *maximizing* the variance among measures $\mu \in \mathcal{D}$ which are dominated by ν in convex order. We show that this novel denoising problem has certain desirable properties; solutions always exist for reasonably nice domains $\mathcal{D}$, and are robust in the sense that if the measure ν is close to some $\rho \in \mathcal{D}$ (i.e., the noise is small) with $\rho \preceq_C \nu$, our solution μ must be quantifiably close to ρ as well. Furthermore, for particularly simple domains, we show that problem (1) in fact coincides with the following *relaxed problem*

$$(2) \qquad \min_{\mu \in \mathcal{D}} \min_{\pi \in \Pi(\mu, \nu)} \int |x - y|^2 d\pi(x, y)$$

where $\Pi(\mu, \nu)$ is the set of couplings between μ and ν (without the martingale or self-consistency condition imposed). This is not surprising, since the self-consistency condition arises as a sort of optimality condition in (8); this equivalence is in fact implicit in standard analysis of k-means clustering problems (although we have not seen (1) explicitly formulated in this setting, and we establish the equivalence more generally here by identifying a general condition on the domain $\mathcal{D}$ for which it holds).

Since they serve as a key motivation for our work here, let us digress briefly to describe in more detail how various notions of principal curves appearing in the literature relate to our framework. Hastie and Stuetzle defined a principal curve of ν to be a smooth curve $s \in \mathbb{R} \mapsto f(s)$ such that for each s, the barycenter of the points y which project to $f(s)$ is $f(s)$. In our language, letting P be the projection map $P(y) = \operatorname{argmin}_{f(s)} \|y - f(s)\|$, $\mu = P_\# \nu$ and $\pi = (P, \operatorname{Id})_\# \nu \in \mathcal{M}(\mu, \nu)$, where Id denotes the identity map and $P_\# \nu$ the pushforward of the data distribution ν by P.[1] Several variants of this definition have been defined since. Notably, Tibshirani relaxed the projection requirement by looking for (in our nomenclature) a probability measure μ supported on a smooth curve $s \mapsto f(s)$, together with a martingale coupling $\pi \in \mathcal{M}(\mu, \nu)$ (so μ need not be the projection of ν onto f, but must still be in convex order with it) [23]. Neither the definition in [12] nor the one in [23] had a variational aspect analogous to (1), although the idea of minimizing the distance to the data

---

[1]Let $F : \mathcal{X} \to \mathcal{Y}$ be a measurable map and ρ be a distribution on $\mathcal{X}$. The pushforward of ρ by the map F, denoted by $F_\# \rho$, is the distribution on $\mathcal{Y}$ satisfying $F_\# \rho(B) = \rho(F^{-1}(B))$ for every measurable set B in $\mathcal{Y}$.

was clearly present in the formulation of the self-consistency condition as discussed above. Heuristically, when $\mathcal{D}$ is taken to be the set of all curves (neglecting for now issues about the regularity of the curves), our problem (1) is to find the principal curve in the sense of Tibshirani which is closest to the data.

Existence of principal curves as defined by Hastie-Stuetzle for general distributions ν is not known. Partially to address this, another notion was introduced by Kegl et al [17]. In our language, their principal curves are solutions to the relaxed problem (2) when $\mathcal{D}$ is the set of all continuous curves of length at most L. With this definition, they easily established existence for any ν. On the other hand, the self-consistency condition is lost.

For this same domain $\mathcal{D}$, we can define a principal curve as a solution to (1); a straight-forward argument (given in a more general setting in the first part of Theorem 2.4) then implies existence of a solution. To the best of our knowledge, this is the first notion of a self-consistent principal curve for which solutions generally exist and have the desirable feature of minimizing noise, or being as close as possible to the data.

Returning to the discussion of general domains $\mathcal{D}$, despite its advantages outlined above, our formulation (1) does come with certain drawbacks. First, checking whether a self-consistent coupling between $\mu \in \mathcal{D}$ and ν exists, or, equivalently, checking whether μ and ν are in convex order, is not straightforward, and so (1) is computationally challenging. Secondly, we demonstrate that (1) can be *unstable* with respect to variations in the data ν, and third, for certain problems of interest, the domain $\mathcal{D}$ may contain very few measures μ such that $\mu \preceq_{\mathrm{C}} \nu$, making the problem (1) trivial and its solution uninformative (see the example in Section 4.4 and in particular Remark 10).

To address these issues, we introduce a weakening of the convex order relation, and a corresponding variational problem (see (22) below). The new dominance relation between two probabiltiy measures, which we call the Kantorovich dominance relation (KDR in short), amounts to imposing the existence of a coupling between μ and ν which, though not necessarily self-consistent, enjoys some features of self-consistency; the resulting denoising problem (22) therefore falls *in between* the original problem (1) with the full self-consistency condition and the fully relaxed problem (2). We argue that this dominance relation is natural, by demonstrating that, like self-consistency, it also arises as an optimality condition for the relaxed problem in a certain sense; indeed, for a large class of domains, which we name cones, we show that the new problem is equivalent to the relaxed problem. In addition, we show that, like (1), (22) enjoys quantifiable robustness properties as the noise becomes small, but, in contrast to (1), (22) is stable with respect to perturbations in the data distribution ν.

We illustrate the properties of our new order dominance relation by discussing its relationship to several established data analysis techniques, including PCA [14], [11], k-means clustering [20], [16], principal curves [19], [15], and Gaussian denoising [6]; (versions of) each of these arise for appropriate choices of the domain $\mathcal{D}$.

The remainder of this article is organized as follows. In Section 2, we present the full, self-consistency problem formulation, establish the existence of an optimal solution under the convex ordering constraint, and provide examples illustrating a variety of feasible domains. In Section 3, we introduce the Kantorovich dominance and investigate its key properties. In Section 4, we examine the variance maximization problem under the Kantorovich dominance, along with equivalent optimization formulations and applications. Section 5 is reserved for numerical examples.

2. General problem formulation and basic properties. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of all probability measures on $\mathbb{R}^d$ with finite second moments, and let $\mathcal{P}_{2,0}(\mathbb{R}^d) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \mid \int x\, d\mu(x) = 0\}$ represent the set of centered probability measures in $\mathcal{P}_2(\mathbb{R}^d)$. In this paper, all probability measures are assumed to be in $\mathcal{P}_2(\mathbb{R}^d)$, unless stated otherwise.

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein distance $\mathcal{W}_2(\mu, \nu)$ between μ and ν is defined as

$$(3) \qquad \mathcal{W}_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, d\pi(x, y),$$

where $\mathcal{W}_2^2(\mu, \nu) = \left(\mathcal{W}_2(\mu, \nu) \right)^2$, and $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν, i.e.,

$$\Pi(\mu, \nu) = \{\pi = \mathcal{L}(X, Y) \mid \mathcal{L}(X) = \mu, \, \mathcal{L}(Y) = \nu\},$$

where $\mathcal{L}(X)$ denotes the law (distribution) of the random variable X [22], [24]. We denote $\mathcal{L}(X) = \mu$ and $X \sim \mu$ interchangeably. $|\cdot|$ denotes the Euclidean norm.

In a typical application, we will assume $Y \sim \nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$ is the given data distribution, satisfying the following model assumption:

$$(4) \qquad Y = X + R,$$

where $X \sim \rho \in \mathcal{P}_{2,0}(\mathbb{R}^d)$, $R \sim \epsilon \in \mathcal{P}_{2,0}(\mathbb{R}^d)$. Our goal is to recover ρ from the observed data ν perturbed by ϵ. With prior knowledge or constraint about ρ, we assume it belongs to a known domain $\mathcal{D} \subset \mathcal{P}_{2,0}(\mathbb{R}^d)$. Given $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$, we thus arrive at the following problem (which is simply (1) rewritten in probabilistic notation):

$$(5) \qquad \min_{\mu \in \mathcal{D}} \min_{\pi \in \mathcal{M}(\mu, \nu)} \mathbb{E}_\pi[|X - Y|^2],$$

where $\mathcal{M}(\mu, \nu)$ represents the set of martingale couplings between μ and ν:

$$(6) \qquad \mathcal{M}(\mu, \nu) = \{\pi = \mathcal{L}(X, Y) \mid \mathcal{L}(X) = \mu, \, \mathcal{L}(Y) = \nu, \, \mathbb{E}_\pi[Y|X] = X\}.$$

The martingale condition is often referred to as *self-consistency* in the statistics literature. In the generic model (4), it says that conditional on X, the average noise vanishes, $\mathbb{E}[R|X] = 0$.

We say that μ and ν in $\mathcal{P}_2(\mathbb{R}^d)$ are in convex order, denoted by $\mu \preceq_C \nu$, if $\int f \, d\mu \leq \int f \, d\nu$ for any convex function f. Strassen's theorem states that $\mathcal{M}(\mu, \nu) \neq \emptyset$ if and only if $\mu \preceq_C \nu$.

If $\pi \in \mathcal{M}(\mu, \nu)$ is a martingale coupling of $X \sim \mu$ and $Y \sim \nu$, we have $\mathbb{E}_\pi[XY] = \mathbb{E}_\mu[X \mathbb{E}_\pi[Y|X]] = \mathbb{E}_\mu[|X|^2]$. Therefore, if $\mu, \nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$, it follows that

$$\mathbb{E}_\pi[|X - Y|^2] = \mathbb{E}_\nu[|Y|^2] - \mathbb{E}_\mu[|X|^2] = \text{Var}(\nu) - \text{Var}(\mu),$$

where $\text{Var}(\mu)$ denotes the variance of μ, defined as $\text{Var}(\mu) = \int |x - \int x \, d\mu(x)|^2 d\mu(x)$.

Since ν is given and fixed, this shows that problem (5) can equivalently be formulated as:

$$(7) \qquad \max_{\mu \in \mathcal{D}, \, \mu \preceq_C \nu} \text{Var}(\mu).$$

REMARK 1. Since $\mathcal{D}$ typically consists of measures supported on low dimensional spaces (for instance, curves), this formulation appears to achieve a spectacular dimensional reduction, as it involves optimizing over $\mu \in \mathcal{D}$, rather than (μ, π) with $\mu \in \mathcal{D}$ and $\pi \in \mathcal{M}(\mu, \nu)$. The catch, of course, is that the constraint $\mu \preceq_C \nu$ involves ν in a sophisticated way and is not straightforward to check.

REMARK 2. We have $\mu \preceq_C \nu$ if and only if $\int f \, d\mu \leq \int f \, d\nu$ for every Lipschitz convex function f. To see this, assume the inequality holds for all Lipschitz convex f, and let g be any convex lower semicontinuous function. Then there exists a sequence of Lipschitz convex functions (indeed, one may take piecewise affine convex functions) $(g_n)_{n \geq 1}$ which increases to g pointwise. For each n, $\int g_n \, d\mu \leq \int g_n \, d\nu \leq \int g \, d\nu$. Letting $n \to \infty$ and applying the monotone convergence theorem yields $\int g \, d\mu \leq \int g \, d\nu$, which is the convex order condition.

Note that, with the definition of Wasserstein distance in hand, we can rewrite the relaxed problem (2) from the introduction as:

$$\min_{\mu \in \mathcal{D}} \mathcal{W}_2^2(\mu, \nu). \tag{8}$$

Since self-consistency can be seen as a first order optimality condition for the functional in (8) [12], it is not surprising that for certain simple domains $\mathcal{D}$, problems (5) and (8) are in fact equivalent. This is well known for problems such as k-means clustering (see Example 3 below); we offer here a general condition on $\mathcal{D}$ under which equivalence holds.

The π-conditional barycentric recentering operation described below is a well-known procedure in the k-means clustering algorithm. (For this connection, one may assume μ is supported on k points in $\mathbb{R}^d$.)

DEFINITION 2.1.   Given $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and a coupling $\pi = \mu \otimes \pi_x \in \Pi(\mu, \nu)$,[2] define the π-conditional barycentric map $c_\pi(x) := \int y \, d\pi_x(y)$ for μ-a.e. x. We call $c_{\pi\#}(\mu)$ the π-conditional barycentrically recentered first marginal of π, and $\mathsf{C}_\#\pi$ as the π-conditional barycentrically recentered coupling of π, where $\mathsf{C}(x, y) = \mathsf{C}_\pi(x, y) := (c_\pi(x), y)$.

Note that $\mathsf{C}_\#\pi \in \mathcal{M}(c_{\pi\#}(\mu), \nu)$ is a martingale measure, and thus $c_{\pi\#}(\mu) \preceq_{\mathrm{C}} \nu$.

PROPOSITION 2.2.   *Assume that there exists an optimal $\mu \in \mathcal{D}$ for the relaxed problem (8) and an optimal $\pi \in \Pi(\mu, \nu)$ for (3) such that $c_{\pi\#}(\mu) \in \mathcal{D}$. Then $c_{\pi\#}(\mu)$ is optimal in both (8) and (5), and consequently, the two problems are equivalent (by yielding the same value).*

PROOF.   Since $c_\pi(x)$ is the barycenter of π_x, we must have

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\mathsf{C}_\#\pi(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |c_\pi(x) - y|^2 d\pi(x, y)$$

$$= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |c_\pi(x) - y|^2 d\pi_x(y) \right) d\mu(x)$$

$$\leq \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |x - y|^2 d\pi_x(y) \right) d\mu(x) = \mathcal{W}_2^2(\mu, \nu),$$

where the last equality is from the assumption that π is optimal for (3). Since $\mathcal{W}_2^2(\mu, \nu)$ is equal to the minimal value in (8) (since $\mu \in \mathcal{D}$) and less than or equal to the minimal value in (5) (since $c_{\pi\#}(\mu) \in \mathcal{D}$), this implies optimality of $c_{\pi\#}(\mu)$ in both problems.   □

We now turn our attention to the existence and robustness of solutions. Given $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, we define the set of probability measures less than or equal to ν in convex order as:

$$\mathcal{M}_\nu := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \mid \mu \preceq_{\mathrm{C}} \nu\}. \tag{9}$$

LEMMA 2.3.   i) *$\mathcal{M}_\nu$ is compact in the $\mathcal{W}_2$-metric for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.*
ii) *Let $\mathcal{W}_2(\nu_n, \nu) \to 0$ as $n \to \infty$, and $\mu_n \in \mathcal{M}_{\nu_n}$. Then the sequence $\{\mu_n\}_n$ is precompact, meaning there exists a subsequence $\{\mu_{n_k}\}_k$ that converges in $\mathcal{W}_2$ to some $\mu \in \mathcal{M}_\nu$.*

---

[2] A disintegration of $\pi \in \Pi(\mu, \nu)$ w.r.t. μ, denoted by $\pi = \mu \otimes \pi_x$, means that for any Borel sets $A, B \subset \mathbb{R}^d$, it holds $\pi(A \times B) = \int_A \pi_x(B) d\mu(x)$. Using conditional probability notation, $\pi_x(B) = \mathcal{P}(Y \in B \mid X = x)$.

PROOF. i) The set $\mathcal{M}_\nu$ is clearly closed: If $\{\mu_n\}_n \subset \mathcal{M}_\nu$ and $\lim_{n\to\infty} \mathcal{W}_2(\mu_n, \mu) = 0$ so that $\int f \, d\mu_n \leq \int f \, d\nu$ for any L-Lipschitz convex function f, then by duality, we have

$$\left| \int f \, d\mu_n - \int f \, d\mu \right| \leq L\mathcal{W}_1(\mu_n, \mu) \leq L\mathcal{W}_2(\mu_n, \mu),$$

proving $\int f \, d\mu \leq \int f \, d\nu$. Hence, $\mu \in \mathcal{M}_\nu$ (see Remark 2).

To show $\mathcal{M}_\nu$ is compact, let $\{\mu_n\}_n \subset \mathcal{M}_\nu$. By monotone convergence, for $\epsilon > 0$, there exists $a > 0$ with $\int (2|x|^2 - a^2)^+ \, d\nu < \epsilon$. Then we have

$$\int (2|x|^2 - a^2)^+ \, d\nu \geq \int (2|x|^2 - a^2)^+ \, d\mu_n \geq \int_{|x| \geq a} (2|x|^2 - a^2)^+ \, d\mu_n \geq \int_{|x| \geq a} |x|^2 \, d\mu_n.$$

By the second inequality, $\{\mu_n\}_n$ is tight. Hence there is a subsequence $\{\mu_k\}_k$ and a probability measure μ such that $\mu_k \to \mu$ weakly. The lemma will follow if $\mathcal{W}_2(\mu_k, \mu) \to 0$. By [24, Theorem 7.12], it suffices to show that for any $\epsilon > 0$, there is $R > 0$ such that $\int_{|x| \geq R} |x|^2 \, d\mu_k(x) < \epsilon$ for all k. This is shown by the third inequality, completing the proof.

ii) As before, for $\epsilon > 0$, there exists $a > 0$ such that $\int (|x| - a)^+ \, d\nu < \epsilon$. Then,

$$\mu_n(\mathbb{R}^d \setminus B_{a+1}(0)) \leq \int (|x| - a)^+ \, d\mu_n \leq \int (|x| - a)^+ \, d\nu_n < \epsilon,$$

for all large n, since $\mathcal{W}_2(\nu_n, \nu) \to 0$ implies $\int (|x| - a)^+ \, d\nu_n \to \int (|x| - a)^+ \, d\nu$. This yields a subsequence $\{\mu_k\}_k$ and a probability measure μ such that $\mu_k \to \mu$ weakly. As before, there exists $a > 0$ such that $\int (2|x|^2 - a^2)^+ \, d\nu < \epsilon$, implying

$$\int_{|x| \geq a} |x|^2 \, d\mu_k \leq \int (2|x|^2 - a^2)^+ \, d\nu_k < \epsilon,$$

for all large k. By [24, Theorem 7.12], we conclude that $\mathcal{W}_2(\mu_k, \mu) \to 0$. $\qquad\square$

REMARK 3. The preceding lemma holds with an almost identical proof hold if we replace $\mathcal{W}_2$ with the p-Wasserstein distance, defined by $\mathcal{W}_p^p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \, d\pi(x, y)$ for $p \in [1, \infty)$, asserting that $\mathcal{M}_{\nu, p} := \{ \mu \in \mathcal{P}_p(\mathbb{R}^d) \mid \mu \preceq_C \nu \}$, is $\mathcal{W}_p$-compact.

We now show that our optimization problem (7) admits a solution and can recover the true distribution ρ as the noise diminishes.

THEOREM 2.4. i) *If $\mathcal{D} \subset \mathcal{P}_2(\mathbb{R}^d)$ is closed under the $\mathcal{W}_2$-metric and $\mathcal{D} \cap \mathcal{M}_\nu$ is non-empty, then problem (7) attains a solution.*

ii) *Let μ^* be a solution to (7). Then for any $\rho \in \mathcal{D}$ with $\rho \preceq_C \nu$, we have:*

$$\mathcal{W}_2(\mu^*, \rho) \leq \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\rho)} + \mathcal{W}_2(\nu, \rho).$$

Consequently, $\mathcal{W}_2(\mu^*, \rho) \to 0$ as $\mathcal{W}_2(\nu, \rho) \to 0$, i.e., as the noise diminishes.

PROOF. i) By Lemma 2.3 the set $\mathcal{M}_\nu$ is $\mathcal{W}_2$-compact. Hence, $\mathcal{D} \cap \mathcal{M}_\nu$ is $\mathcal{W}_2$-compact. Since the functional $\mu \mapsto \mathrm{Var}(\mu)$ is continuous in the $\mathcal{W}_2$-metric, part i) follows.

ii) Recall that $\mathcal{W}_2(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \sqrt{\mathbb{E}_\pi[|X - Y|^2]}$. We proceed as follows:

$$\begin{aligned} \mathcal{W}_2(\mu^*, \rho) &\leq \mathcal{W}_2(\mu^*, \nu) + \mathcal{W}_2(\nu, \rho) \\ &\leq \sqrt{\mathbb{E}_\pi[|X - Y|^2]} + \mathcal{W}_2(\nu, \rho), \quad \text{for any } \pi \in \mathcal{M}(\mu^*, \nu) \\ &= \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\mu^*)} + \mathcal{W}_2(\nu, \rho) \\ &\leq \sqrt{\mathrm{Var}(\nu) - \mathrm{Var}(\rho)} + \mathcal{W}_2(\nu, \rho), \end{aligned}$$

where the last inequality follows from the optimality of μ^* in problem (7). $\qquad\square$

2.1. *Examples of domains.*   The following examples clarify the connection between our framework, k-means clustering, and principal curves.

EXAMPLE 1 ($\mathcal{D}$ as the set of measures on Lipschitz curves).   Let $\Omega \subset \mathbb{R}^d$ be a compact and convex set, and fix parameters $L, T > 0$. Consider the set of Lipschitz curves and the probability measures supported on them, defined by

$$\text{(10)} \qquad \mathcal{C}_L = \big\{\alpha : [0, T] \to \Omega \,\big|\, |\alpha(t) - \alpha(s)| \leq L|t - s|\big\},$$

$$\text{(11)} \qquad \mathcal{D}_L = \big\{\mu \in \mathcal{P}_2(\mathbb{R}^d) \,\big|\, \text{spt}(\mu) \subset \text{Im}(\alpha) \text{ for some } \alpha \in \mathcal{C}_L\big\},$$

where $\text{Im}(\alpha) = \{\alpha(t) \,|\, t \in [0, T]\}$ and $\text{spt}(\mu)$ denotes the support of μ. In practice, Ω can be the convex hull of the support of ν, or a closed ball that contains ν's support.

To show $\mathcal{D}_L$ is closed in $\mathcal{W}_2$-metric, let $(\mu_n)_n$ be a sequence in $\mathcal{D}_L$ converging to μ, and let $(\alpha_n)_n \subset \mathcal{C}_L$ with $\text{spt}(\mu_n) \subset \text{Im}(\alpha_n)$. We need to show that $\text{spt}(\mu) \subset \text{Im}(\alpha)$ for some $\alpha \in \mathcal{C}_L$. By the Arzelà–Ascoli theorem, there exists a subsequence of $(\alpha_n)_n$ (which we still denote by $(\alpha_n)_n$) that converges uniformly on $[0, T]$ to some $\alpha \in \mathcal{C}_L$. This implies $\text{spt}(\mu) \subset \text{Im}(\alpha)$ as follows: for any $\epsilon > 0$, there exists N such that for all $n \geq N$, $\text{Im}(\alpha_n) \subset \mathcal{N}_\epsilon(\text{Im}(\alpha))$, where

$$\mathcal{N}_\epsilon(\text{Im}(\alpha)) := \big\{x \in \mathbb{R}^d \,|\, |x - y| \leq \epsilon \text{ for some } y \in \text{Im}(\alpha)\big\}.$$

Since $\text{spt}(\mu_n) \subset \text{Im}(\alpha_n)$ and $\mu_n \to \mu$, we have $\text{spt}(\mu) \subset \mathcal{N}_\epsilon(\text{Im}(\alpha))$. Taking the limit as $\epsilon \to 0$ yields $\text{spt}(\mu) \subset \text{Im}(\alpha)$. We note that the relaxed problem (8) for this domain is equivalent to the version of principal curves proposed in [17].

EXAMPLE 2 ($\mathcal{D}$ as the set of measures on monotone increasing curves).   A set $\Gamma \subset \mathbb{R}^2$ is said to be monotone if, for any $(x_1, x_2), (y_1, y_2) \in \Gamma$, the following condition holds:

$$\text{(12)} \qquad (y_1 - x_1)(y_2 - x_2) \geq 0.$$

Let MON denote the collection of all monotone sets in $\mathbb{R}^2$. We define the search space as

$$\text{(13)} \qquad \mathcal{D}_{\text{MON}} = \big\{\mu \in \mathcal{P}(\mathbb{R}^2) \,|\, \text{spt}(\mu) \subset \Gamma, \Gamma \in \text{MON}\big\}.$$

To show that $\mathcal{D}_{\text{MON}}$ is closed under the $\mathcal{W}_2$-metric, consider a sequence $\mu_n \to \mu$ in $\mathcal{W}_2$ with $\mu_n \in \mathcal{D}_{\text{MON}}$. $\mathcal{W}_2$-convergence implies weak convergence, so by the Portmanteau theorem, for any $(x_1, x_2), (y_1, y_2) \in \text{supp}(\mu)$, there exist $(x_1^n, x_2^n), (y_1^n, y_2^n) \in \text{supp}(\mu_n)$ such that $(x_1^n, x_2^n) \to (x_1, x_2)$ and $(y_1^n, y_2^n) \to (y_1, y_2)$. By continuity of the product function, we have $(y_1 - x_1)(y_2 - x_2) \geq 0$ for any $(x_1, x_2), (y_1, y_2) \in \text{spt}(\mu)$, confirming that $\mu \in \mathcal{D}_{\text{MON}}$. Monotonicity, reflecting an increasing dependence on the signal variables X_1, X_2, is a natural modelling assumption in many situations (such as when the noisy data variables Y_1, Y_2 are highly correlated). It is also closely related to the Lipschitz condition in the preceding example, as monotone sets $\Gamma \subset \mathbb{R}^2$ are well known to be 1-Lipschitz graphs $(\bar{x}_1, \bar{x}_2) = (\bar{x}_1, F(\bar{x}_1))$ of the anti-diagonal $\bar{x}_2 = \frac{1}{\sqrt{2}}[x_1 - x_2]$ over the diagonal $\bar{x}_1 = \frac{1}{\sqrt{2}}[x_1 + x_2]$ [21].

REMARK 4.   The last example is closely related to the backwards martingale optimal transport problem studied in [18]. For a given $\nu \in \mathcal{P}_2(\mathbb{R}^2)$, they attempt to minimize $\mathbb{E}_\pi[(Y_1 - X_1)(Y_2 - X_2)]$ among measures $\mu \in \mathcal{P}_2(\mathbb{R}^2)$ and martingale couplings $\pi \in \mathcal{M}(\mu, \nu)$. This differs from our framework in that their cost function $c(x, y) = (y_1 - x_1)(y_2 - x_2) = (\bar{y}_1 - \bar{x}_1)^2 - (\bar{y}_2 - \bar{x}_2)^2$ is increasing along the diagonal direction but decreasing along the anti-diagonal direction and they do not restrict to measures μ in any particular domain $\mathcal{D}$. Despite the latter difference, they show that their optimal μ in fact belongs to $\mathcal{D}_{\text{MON}}$. Thus, solving either the problem in the last example or the problem in [18] yields a measure $\mu \in \mathcal{D}_{\text{MON}}$ with $\mu \preceq_C \nu$. These two measures need not be the same, as shown in Supplementary Material A.

The following example relates our framework to data clustering problems.

EXAMPLE 3 ($\mathcal{D}$ as the set of measures supported on sets of bounded cardinality). Fix $m \in \mathbb{N}$. Consider the following search space

(14)
$$\mathcal{D}^m = \left\{ \mu = \sum_{i=1}^m u_i \delta_{x_i} \,\bigg|\, x_i \in \mathbb{R}^d, \, u_i \geq 0 \,\forall i, \, \sum_{i=1}^m u_i = 1 \right\}.$$

We also consider the set of discrete measures of fixed weight as follows. Let $u = (u_1, ..., u_m)$ be a vector of fixed weights, where $u_i \geq 0$ and $\sum_{i=1}^m u_i = 1$. We set

(15)
$$\mathcal{D}_u^m = \left\{ \mu = \sum_{i=1}^m u_i \delta_{x_i} \,\bigg|\, x_i \in \mathbb{R}^d \right\}.$$

Proposition 2.2 implies that (5) and (8) are equivalent for either of these domains. Note that the relaxed problem (8) with $\mathcal{D}^m$ is exactly the classical k-means clustering problem, while with $\mathcal{D}_u^m$ it is a fixed weight clustering problem.

2.2. *Stability.* The following example illustrates that solution stability may not hold if the input data ν varies. Note that this differs from the scenario of diminishing noise.

EXAMPLE 4 (Instability of denoising with self-consistency). We demonstrate that the stability of problem (5) does not generally hold, inspired by [4].
Define the one-step probability kernel κ_θ from $\mathbb{R}^2$ to $\mathcal{P}(\mathbb{R}^2)$ by

(16)
$$\kappa_\theta(x_1, x_2) = \tfrac{1}{2}\delta_{(x_1+\cos(\theta), x_2+\sin(\theta))} + \tfrac{1}{2}\delta_{(x_1-\cos(\theta), x_2-\sin(\theta))}.$$

Define $\mu = \tfrac{1}{3}\delta_{(-1,0)} + \tfrac{1}{3}\delta_{(0,0)} + \tfrac{1}{3}\delta_{(1,0)}$, $\theta_n = \pi(1 - \frac{1}{2(n+1)})$, and ν_n as the second marginal of the martingale transport $\mu \otimes \kappa_{\theta_n} \in \mathcal{M}(\mu, \nu_n)$, depicted as red dots in Figure 1) for each $n \in \mathbb{N}$. Then as $n \to \infty$, we have $\nu_n \to \nu_\infty := \kappa_\pi \mu = \tfrac{1}{6}\delta_{(-2,0)} + \tfrac{1}{6}\delta_{(-1,0)} + \tfrac{1}{3}\delta_{(0,0)} + \tfrac{1}{6}\delta_{(1,0)} + \tfrac{1}{6}\delta_{(2,0)}$.
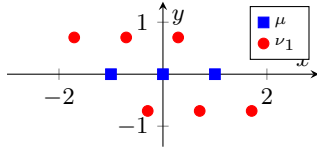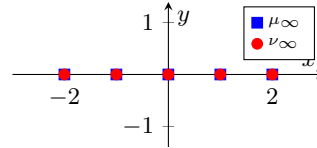

FIG 1. *Support of μ and ν_1*

FIG 2. *Support of $\mu_\infty = \nu_\infty$*

Set $\mathcal{M}_{\nu,\text{MON}} = \{\mu \in \mathcal{D}_{\text{MON}} \mid \mu \preceq_C \nu\}$. For each $n \in \mathbb{N} \cup \{\infty\}$, let μ_n be the solution to (7) with respect to ν_n, which aims to maximize the variance by optimally spreading out the mass. This results in $\mu_n = \mu$ for all $n \in \mathbb{N}$ due to the monotonicity constraint. However, in the limit $n \to \infty$, ν_∞ itself is monotone, implying $\mu_\infty = \nu_\infty$. This example shows that μ is not stable as a solution to (7) with respect to ν in the sense that $\nu_n \to \nu_\infty \not\Rightarrow \mu_n \to \mu_\infty$, demonstrating the instability of problem (5).

In contrast to (5), the relaxed problem (8) enjoys the following stability property.

PROPOSITION 2.5 (Stability of the relaxed problem). *If $\nu_n \to \nu$ in $\mathcal{W}_2$ metric, and $\mu_n \in \arg\min_{\mu \in \mathcal{D}} \mathcal{W}_2^2(\mu, \nu_n)$ converges in $\mathcal{W}_2$ to μ^*, then $\mu^* \in \arg\min_{\mu \in \mathcal{D}} \mathcal{W}_2^2(\mu, \nu)$.*

This proposition implies that if $\mathcal{D}$ is compact and (8) has a unique minimizer μ^* for ν, then any sequence of minimizers μ_n for ν_n converges to μ^*.

PROOF. Since μ_n minimizes $\mathcal{W}_2^2(\cdot, \nu_n)$ over $\mathcal{D}$, we have

$$\mathcal{W}_2^2(\mu_n, \nu_n) \leq \mathcal{W}_2^2(\xi, \nu_n) \quad \forall \xi \in \mathcal{D}.$$

By the continuity of $\mathcal{W}_2$ with respect to its own topology, taking the limit yields

$$\mathcal{W}_2^2(\mu^*, \nu) = \lim_{n \to \infty} \mathcal{W}_2^2(\mu_n, \nu_n) \leq \lim_{n \to \infty} \mathcal{W}_2^2(\xi, \nu_n) = \mathcal{W}_2^2(\xi, \nu) \quad \forall \xi \in \mathcal{D}.$$

This confirms that μ^* is a minimizer of $\mathcal{W}_2^2(\cdot, \nu)$ over $\mathcal{D}$. □

2.3. *Preliminary numerical simulations.* As a proof of concept, we generate 110 and 1000 observations of $Y = X + R$, where X is monotone and R is Gaussian noise. We set the domain $\mathcal{D}$ to be either the set of measures supported on a monotone set or the set of measures supported on a curve with finite length. We solve problem (5) by adapting a generalized Lloyd algorithm. The self-consistency constraint is softly enforced via a penalty term in the objective function. For comparison, we also generated principal curves based on different definitions from [12], [17], and [23].[3] The results are shown in Figure 3. The figures illustrate that the curve fitting method based on problem (5) performs reasonably well in capturing the underlying structure of the data compared to other principal curve methods.
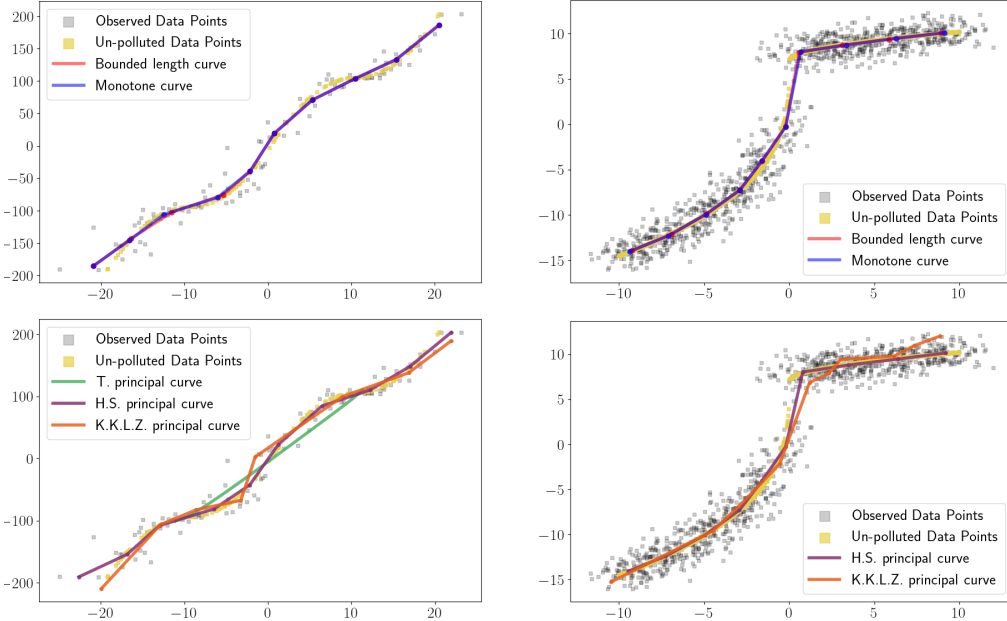

FIG 3. *Principal curve fitting methods applied to 110 data points (Left) and 1000 monotone data points (Right). "Monotone curve" and "Bounded length curve" correspond to our solutions of problem (5) over the domain $\mathcal{D}$, defined as either the set of measures supported on a monotone set or those supported on a curve with finite length. These solutions are computed using a generalized Lloyd algorithm, where the self-consistency constraint is softly enforced via a penalty term in the objective function. For comparison, we also present the results of fitting principal curves using the methods by Hastie-Stuetzle [12], Kégl-Krzyzak-Linder-Zeger [17], and Tibshirani [23].*

---

[3]For the method proposed by [23], we only generate the curve using 110 data points, as we encounter overflow issues with larger sample sizes.

3. The Kantorovich dominance relation. Despite its conceptual appeal, the learning problem with self consistency (5) (or equivalently (7)) has certain drawbacks: the convex order constraint is difficult to check, the problem may be unstable (recall Example 4), and the intersection of the certain natural domains $\mathcal{D}$ with the set $\{\mu \,|\, \mu \preceq_{\mathrm{C}} \nu\}$ may be nearly empty, and thus not contain a reasonable solution (see, for example, the application in Section 4.4).

These challenges prompt us to introduce a weaker relation, the Kantorovich dominance relation, which still captures some aspects of the martingale property, while avoiding some of the computational and theoretical difficulties associated with the convex order relation.

Let $\langle \cdot, \cdot \rangle$ denote the inner product. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we say that μ is less than ν in the Kantorovich dominance relation (KDR), and write $\mu \preceq_{\mathrm{K}} \nu$, if there is $\pi \in \Pi(\mu, \nu)$ such that

$$(17) \qquad \int \langle x - b_\nu, y - x \rangle d\pi(x, y) = 0, \ \text{ where } \ b_\nu = \int y \, d\nu(y) \text{ is the barycenter of } \nu.$$

Let $T_k(x) = x + k$ be a translation map. (17) says $\mu \preceq_{\mathrm{K}} \nu$ iff the translated measures $\mu' = T_{-b_\nu \#}(\mu)$ and $\nu' = T_{-b_\nu \#}(\nu)$ (thus ν' is centered) admits a coupling $\pi' \in \Pi(\mu', \nu')$ such that

$$\int \langle x, y - x \rangle \, d\pi'(x, y) = 0.$$

By definition, the KDR relation is translation invariant: if $\mu \preceq_{\mathrm{K}} \nu$, then for any common shift vector a, the translated measures $\mu_a := (T_a)_\# \mu$ and $\nu_a := (T_a)_\# \nu$ also satisfy $\mu_a \preceq_{\mathrm{K}} \nu_a$. Accordingly, throughout this paper we assume for simplicity that ν is centered (i.e., $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$), unless stated otherwise. Since ν represents the given data distribution, this centering step is straightforward in practice.

REMARK 5. If $\mu \preceq_{\mathrm{C}} \nu$, then there exists a martingale coupling $\pi \in \mathcal{M}(\mu, \nu)$ such that

$$\int \langle x, y - x \rangle d\pi(x, y) = \int \langle x, \Big(\int (y - x) d\pi_x(y) \Big) \rangle d\mu(x) = 0.$$

Therefore, the KDR is a relaxation of the convex order relation, capturing a weaker form of the martingale property. Precisely, $v(x) := \int (y - x) d\pi_x(y)$ for some coupling π, the KDR requires only that v is orthogonal to Id in $L^2(\mu)$.

A key part of the motivation for the Kantorovich dominance is that it arises as a sort of optimality condition for the relaxed problem (8) for a certain class of domains (see Theorem 4.9 below), analogously to the way that self consistency arises as an optimality condition for particularly simple domains (Proposition 2.2).

REMARK 6. If ν is centered, for $\pi = \mu \otimes \nu$ (the product measure), we have

$$\int \langle x, y - x \rangle d\pi(x, y) = \langle \int x \, d\mu(x), \int y \, d\nu(y) \rangle - \int |x|^2 d\mu(x) = - \int |x|^2 d\mu(x) \le 0.$$

Since the image of the weakly continuous mapping $\pi \mapsto \int \langle x, y - x \rangle d\pi(x, y)$ over the connected set $\Pi(\mu, \nu)$ is itself connected, we conclude that (17) is equivalent to

$$(18) \qquad D(\mu, \nu) := \sup_{\pi \in \Pi(\mu, \nu)} \int \langle x, y - x \rangle d\pi(x, y) \ge 0.$$

(18) motivates the term *Kantorovich dominance*, as it shows that the "Kantorovich cost" $\sup_{\pi \in \Pi(\mu, \nu)} \int \langle x, y \rangle d\pi$ dominates (i.e., is greater than or equal to) the second moment of μ.

Since $\mathcal{W}_2^2(\mu, \nu) = \int |x|^2 d\mu + \int |y|^2 d\nu - 2 \sup_{\pi \in \Pi(\mu, \nu)} \int \langle x, y \rangle d\pi$, (18) is equivalent to

$$(19) \qquad \mathcal{W}_2^2(\mu, \nu) \le \int |y|^2 d\nu - \int |x|^2 d\mu.$$

It follows immediately from (19) that the set of probability measures dominated by ν

$$\mathcal{M}^K_{\nu,\Omega} := \{\mu \in \mathcal{P}_2(\Omega) \mid \mu \preceq_K \nu\}$$

is closed with respect to $\mathcal{W}_2$-metric for any compact $\Omega \subset \mathbb{R}^d$. However, in general, the set

$$\mathcal{M}^K_\nu := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \mid \mu \preceq_K \nu\}$$

is not closed with respect to the $\mathcal{W}_2$ metric (see later discussion).

REMARK 7. Let $B_\mu(r) = \{\rho \mid \mathcal{W}_2(\mu,\rho) \leq r\}$. [25] showed that $\mu \preceq_C \nu$ if and only if

$$(20) \qquad \mathcal{W}_2^2(\nu,\rho) - \mathcal{W}_2^2(\mu,\rho) \leq \int |y|^2 d\nu - \int |x|^2 d\mu$$

holds for all $\rho \in B_\mu(\infty) = \mathcal{P}_2(\mathbb{R}^d)$. In contrast, $\mu \preceq_K \nu$ if (20) holds for $\rho \in B_\mu(0)$, i.e., for $\rho = \mu$. Exploring intermediate radii $r \in (0,\infty)$ remains an open direction for further study.

We observe that the tail probability of μ is controlled by the second moment of ν if $\mu \preceq_K \nu$.

LEMMA 3.1. *If $\mu \preceq_K \nu$, then we have*

$$(21) \qquad \mu(\{|x|^2 \geq \alpha\}) \leq \frac{1}{\alpha} \int_Y |y|^2 d\nu \quad \textit{for every } \alpha > 0.$$

(21) easily follows from (19) and Markov's inequality:

$$\int |y|^2 d\nu \geq \mathcal{W}_2^2(\mu,\nu) + \int |x|^2 d\mu \geq \int |x|^2 d\mu \geq \alpha\mu(\{|x|^2 \geq \alpha\}).$$

If we further assume that Ω is compact and $\nu(\Omega) = 1$ (e.g., $\Omega = \mathrm{con}(\mathrm{spt}(\nu))$ with compact $\mathrm{spt}(\nu)$, where $\mathrm{con}(\Omega)$ denotes the convex hull of Ω), then we obtain a stronger compactness.

COROLLARY 3.2. *$\mathcal{M}^K_\nu$ is a precompact subset of $\mathcal{P}(\mathbb{R}^d)$ with respect to the weak topology. If Ω is compact, then $\mathcal{M}^K_{\nu,\Omega}$ is compact in both the weak and $\mathcal{W}_2$ topologies.*

PROOF. By Lemma 3.1 and Prokhorov's theorem, $\mathcal{M}^K_\nu$ is precompact in the weak topology. Since $\mathcal{W}_2$ convergence is equivalent to weak convergence when Ω is compact (see Remark 2.8 of [1]), and $\mathcal{M}^K_{\nu,\Omega}$ is closed, the result follows. $\square$

However, $\mathcal{M}^K_\nu$ is not necessarily compact in the $\mathcal{W}_2$ topology, even if $\mathrm{spt}(\nu)$ is compact. This motivates the consideration of the domain $\mathcal{M}^K_{\nu,\Omega}$ with a compact Ω instead of $\mathcal{M}^K_\nu$.

EXAMPLE 5. Let $\sigma_c \in \mathcal{P}(\mathbb{R}^d)$ denote the uniform probability measure over a sphere with radius $c \geq 0$ centered at the origin. For $0 \leq r \leq 1 \leq R$, let $\mu = (1-\lambda)\sigma_r + \lambda\sigma_R$ and $\nu = \sigma_1$. The optimal coupling π for $\mathcal{W}_2^2(\mu,\nu)$ sends $(1-\lambda)\sigma_r$ radially to $(1-\lambda)\sigma_1$ and $\lambda\sigma_R$ to $\lambda\sigma_1$.

If $r \in (0,1)$ is fixed, and let $R = \frac{1+\sqrt{1+4(r-r^2)(\frac{1}{\lambda}-1)}}{2}$, it is straightforward to verify that inequality (19) still holds, and hence $\mu \preceq_K \nu$. This shows that the support of μ can be arbitrarily large (by letting $\lambda \to 0$), and that the Kantorovich order relation is not equivalent to the convex order relation, even in one dimension.

This example also demonstrates that $\mathcal{M}^K_\nu$ is not $\mathcal{W}_2$ compact in $\mathcal{P}_2(\mathbb{R}^d)$, even if ν is compactly supported. If we let $r = 1/2$, R as above, and $\mu_\lambda = (1-\lambda)\sigma_r + \lambda\sigma_R$, then as $\lambda \to 0$, μ_λ converges weakly to the uniform probability measure μ over a sphere with radius $1/2$ and second moment $1/4$. However, as $\lambda \to 0$, we have $\int |x|^2 d\mu_\lambda \to 1/2 \neq 1/4$.
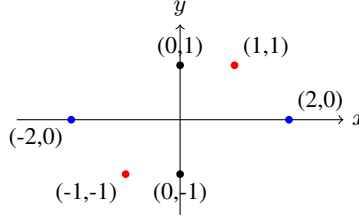
FIG 4. *An example showing that the KDR is not transitive and hence not a partial order.*

Although the KDR reflects certain aspects of the convex order, it is not itself a partial order, as it generally fails to satisfy transitivity. We illustrate this with examples in one and two dimensions.

EXAMPLE 6. Let $\mu = \frac{1}{3}\delta_{-4} + \frac{2}{3}\delta_2$, $\nu = \frac{1}{3}\delta_{-4} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_4$ and $\theta = \frac{2}{3}\delta_{-3} + \frac{1}{3}\delta_6$. Then $\mathcal{W}_2^2(\mu,\nu) = \frac{8}{3}$, $\mathcal{W}_2^2(\nu,\theta) = \frac{14}{3}$ and $\mathcal{W}_2^2(\mu,\theta) = 14$. On the other hand, $\int |x|^2 d\mu = 8$, $\int |y|^2 d\nu = \frac{32}{3}$ and $\int |z|^2 d\theta = 18$. By (19), $\mu \preceq_{\mathrm{K}} \nu$ and $\nu \preceq_{\mathrm{K}} \theta$, but we do not have $\mu \preceq_{\mathrm{K}} \theta$.

EXAMPLE 7. Let $\mu = \frac{1}{2}\delta_{(0,-1)} + \frac{1}{2}\delta_{(0,1)}$, $\nu = \frac{1}{2}\delta_{(-1,-1)} + \frac{1}{2}\delta_{(1,1)}$, and $\theta = \frac{1}{2}\delta_{(-2,0)} + \frac{1}{2}\delta_{(2,0)}$ (see Figure 4). Clearly, $\mathcal{W}_2^2(\mu,\nu) = 1$ and $\mathcal{W}_2^2(\nu,\theta) = 2$. Meanwhile, $\int |x|^2 d\mu = 1$, $\int |y|^2 d\nu = 2$, and $\int |z|^2 d\theta = 4$. Thus, $\mu \preceq_{\mathrm{K}} \nu$ and $\nu \preceq_{\mathrm{K}} \theta$ by (19). However, since $\mathcal{W}_2^2(\mu,\theta) = 5 > 3 = \int |z|^2 d\theta - \int |x|^2 d\mu$, we do not have $\mu \preceq_{\mathrm{K}} \theta$.

4. Maximizing variance under the Kantorovich dominance relation. Throughout the sequel, we will assume that Ω is a compact, convex set containing 0, unless stated otherwise. An example is $\Omega = \mathrm{con}(\mathrm{spt}(\nu))$. Let $\nu \in \mathcal{P}_{2,0}(\Omega)$. Motivated by our earlier discussion, given a domain $\mathcal{D} \subset \mathcal{P}_2(\mathbb{R}^d)$, we now study the following variance maximization problem:

$$(22) \qquad \max_{\mu \in \mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K} \mathrm{Var}(\mu).$$

THEOREM 4.1. i) *If $\mathcal{D} \subset \mathcal{P}_2(\mathbb{R}^d)$ is closed in the $\mathcal{W}_2$-metric and $\mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ non-empty, then problem* (22) *admits a solution.*

ii) *Let μ^* be any solution to* (22). *Then for any $\rho \in \mathcal{D} \cap \mathcal{M}_\nu^K \cap \mathcal{P}_{2,0}(\mathbb{R}^d)$, we have:*

$$\mathcal{W}_2(\mu^*,\rho) \leq \sqrt{\int |y|^2 d\nu - \int |z|^2 d\rho} + \mathcal{W}_2(\nu,\rho).$$

Consequently, problem (22) *exhibits a denoising property, meaning that as the noise in the measure ν decreases, the optimal solution μ^* tends to recover the original measure.*

PROOF. i) Since $\mathcal{M}_{\nu,\Omega}^K$ is compact in the $\mathcal{W}_2$-metric by Corollary 3.2, and $\mathcal{D}$ is $\mathcal{W}_2$-closed, it follows that $\mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ is also $\mathcal{W}_2$-compact. The continuity of the variance functional with respect to the $\mathcal{W}_2$-metric guarantees that problem (22) admits a solution.

ii) The proof is similar to the proof of Theorem 2.4.

$$\mathcal{W}_2(\mu^*,\rho) \leq \mathcal{W}_2(\mu^*,\nu) + \mathcal{W}_2(\nu,\rho)$$

$$\leq \sqrt{\int |y|^2 d\nu - \int |x|^2 d\mu^*} + \mathcal{W}_2(\nu,\rho) \leq \sqrt{\int |y|^2 d\nu - \int |z|^2 d\rho} + \mathcal{W}_2(\nu,\rho),$$

where the second inequality is by the KDR and the last inequality follows from the optimality of μ^*, which gives $\mathbb{E}_\rho[|X|^2] = \mathrm{Var}(\rho) \leq \mathrm{Var}(\mu^*) = \mathbb{E}_{\mu^*}[|X|^2] - |\mathbb{E}_{\mu^*}[X]|^2 \leq \mathbb{E}_{\mu^*}[|X|^2]$. □

4.1. *Stability.* Under a certain condition on $\mathcal{D}$, we now show problem (22) is stable, in contrast to (5), as shown in Example 4. The hypothesis we will impose on $\mathcal{D}$ in the following definitions is quite weak; in particular, it is satisfied by all domains we have considered.

DEFINITION 4.2. Let $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$. We say that $\mathcal{D}$ is *approachable from the interior with respect to ν* if, for any $\mu \in \mathcal{D}$ with $\mu \neq \delta_0$ and $\mu \preceq_K \nu$, there exists a sequence $\{\mu_n\} \subset \mathcal{D}$ such that $\mathcal{W}_2(\mu_n, \mu) \to 0$ and $D(\mu_n, \nu) > 0$ in (18) for all $n \in \mathbb{N}$.

The class of domains that are closed under contractions serves as a key example of domains that can be approached from the interior.

DEFINITION 4.3. Let $\lambda_\# \mu$ denote the *dilation* of μ by $\lambda \in \mathbb{R}$, i.e., $\lambda X \sim \lambda_\# \mu$ when $X \sim \mu$. We say that $\mathcal{D}$ is *closed under contractions* if $\lambda_\# \mu \in \mathcal{D}$ for any $\mu \in \mathcal{D}$ and $\lambda \in (0,1)$.

LEMMA 4.4. *If $\mathcal{D}$ is closed under contractions, then for any $\nu \in \mathcal{P}_{2,0}(\Omega)$ and $\delta_0 \neq \mu \in \mathcal{D}$ with $\mu \preceq_K \nu$, there is a sequence $\{\mu_n\} \subset \mathcal{D}$ such that $W_2(\mu_n, \mu) \to 0$ and $D(\mu_n, \nu) > 0$.*

PROOF. If $D(\mu, \nu) > 0$ there is nothing to prove. If not, there exists some $\pi \in \Pi(\mu, \nu)$ with $\int \langle x, y - x \rangle d\pi(x,y) = 0$. Let λ_n be an increasing sequence of positive numbers converging to 1 and choose $\mu_n = \lambda_{n\#}\mu$ and $\pi_n = \left((x,y) \mapsto (\lambda_n x, y)\right)_\# \pi$, so that $\int \langle x, y - x \rangle d\pi_n(x,y) = \lambda_n \int \langle x, y \rangle d\pi(x,y) - \lambda_n^2 \int |x|^2 d\mu(x) > \lambda_n \int \langle x, y \rangle d\pi(x,y) - \lambda_n \int |x|^2 d\mu(x) = 0$. $\square$

The following result illustrates a stability property inherent to problem (22), where $\Omega \subset \mathbb{R}^d$ is not required to be compact.

THEOREM 4.5. *Let $\nu \in \mathcal{P}_{2,0}(\Omega)$ and assume that $\mathcal{D}$ is approachable from the interior w.r.t. ν. Let $\{\nu_n\} \subset \mathcal{P}_2(\Omega)$ such that $\mathcal{W}_2(\nu_n, \nu) \to 0$. Let μ_n be a solution to (22) with ν_n. If $\mathcal{W}_2(\mu_n, \mu) \to 0$ for some μ, then μ is a solution to (22) with ν.*

PROOF. Consider first an arbitrary $\xi \in \mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ satisfying $\mathcal{W}_2^2(\xi, \nu) < \int |y|^2 d\nu(y) - \int |x|^2 d\xi(x)$. By the strict inequality, for all large n we have $\xi \in \mathcal{D} \cap \mathcal{M}_{\nu_n,\Omega}^K$. By optimality of μ_n in (22) with ν_n, we have $\mathrm{Var}(\mu_n) \geq \mathrm{Var}(\xi)$; taking limits then yields $\mathrm{Var}(\mu) \geq \mathrm{Var}(\xi)$.

Now consider any $\xi \in \mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$. Letting $\{\xi_m\} \subset \mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ approximate ξ such that $\mathcal{W}_2^2(\xi_m, \nu) < \int |y|^2 d\nu - \int |x|^2 d\xi_m$, the above argument yields $\mathrm{Var}(\mu) \geq \mathrm{Var}(\xi_m)$; taking limits implies $\mathrm{Var}(\mu) \geq \mathrm{Var}(\xi)$. As $\xi \in \mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ is arbitrary, optimality of μ follows. $\square$

REMARK 8. A natural question is whether a quantitative version of the approachable from the interior condition yields quantitative stability of (22). Namely, suppose we require that for every $\mu \neq \delta_0$ in $\mathcal{D} \cap \mathcal{M}_{\nu,\Omega}^K$ the existence of an approximating sequence $\{\mu_n\} \subset \mathcal{D}$ as in Definition 4.2 with the additional condition that $f(\mathcal{W}_2(\mu_n, \mu)) < f(D(\mu_n, \nu))$, for some suitable strictly increasing continuous function $f : [0, \infty) \to [0, \infty)$, with $f(0) = 0$. Does this then imply that $g(\mathcal{W}_2(\hat{\mu}, \tilde{\mu})) < g(\mathcal{W}_2(\hat{\nu}, \tilde{\nu}))$, where $\hat{\mu}$ and $\tilde{\mu}$ are solutions in (22) for $\hat{\nu}$ and $\tilde{\nu}$, respectively, for an appropriate, strictly increasing continuous functions $g : [0, \infty) \to [0, \infty)$, with $g(0) = 0$?

Like related issues regarding the quantitative stability of optimal transport (see for instance [9]), this question seems very challenging. Our proof of Theorem 4.5 does not seem to provide a particular insight into how one might approach the quantitative stability question, and its resolution will likely require significant new ideas and techniques.

4.2. *Conic domains and equivalence with the relaxed problem.* We establish the equivalence between the problem (22) and the relaxed problem (8) under certain structural assumptions on $\mathcal{D}$. Notably, this result applies even when $\mathrm{spt}(\nu) = \Omega$ is not compact, ensuring the existence and stability of solutions in such cases as well.

DEFINITION 4.6. $\mathcal{D} \subset \mathcal{P}(\mathbb{R}^d)$ is called a *cone* if $\lambda_{\#}\mu \in \mathcal{D}$ for any $\mu \in \mathcal{D}$ and $\lambda \geq 0$. $\mathcal{D}$ is called *translation invariant* if $T_{k\#}\mu \in \mathcal{D}$ for any $\mu \in \mathcal{D}$ and $k \in \mathbb{R}^d$, where $T_k(x) := x + k$.

It is worth noting that problem (22) is stable on cone domains, as these domains are closed under contraction.

Many of the domains we are interested in are cones and translation invariant, including those in Examples 2 and 3, as well as measures supported on lines and multivariate Gaussians as will be explored in Section 4.4 below. Note that the domain in Example 1 is not a cone.

LEMMA 4.7. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, and let $\mu_k := T_{k\#}\mu$ denote the translation of μ by k. Among all translations μ_k, the one that minimizes the $\mathcal{W}_2$ distance to ν is the one for which*

$$(23) \qquad \int_{\mathbb{R}^d} x \, d\mu_k(x) = \int_{\mathbb{R}^d} y \, d\nu(y).$$

PROOF. The $\mathcal{W}_2$ distance between μ_k and ν is given by

$$\mathcal{W}_2^2(\mu_k, \nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |(x + k) - y|^2 d\pi(x, y),$$

where π is an optimal transport plan between μ and ν. Differentiating with respect to k and setting the derivative equal to zero gives $k = \int y \, d\nu(y) - \int x \, d\mu(x)$. Thus, the translation μ_{k^*} that minimizes the $\mathcal{W}_2$ distance satisfies (23), as claimed. $\square$

LEMMA 4.8. *Let $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$ and $\mathcal{D}$ be a cone. For any $\mu^* \in \mathrm{argmax}_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu)$ and $\pi^* \in \mathrm{argmax}_{\pi \in \Pi(\mu^*, \nu)} \int \langle x, y \rangle d\pi(x, y)$, the following equality holds:*

$$(24) \qquad \int \langle x, y \rangle d\pi^*(x, y) = \int |x|^2 d\mu^*(x).$$

If, in addition, $\mathcal{D}$ is translation invariant, then μ^* must be centered; $\int x \, d\mu^*(x) = 0$.

PROOF. By the KDR (18), for any $\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K$ and for any corresponding optimal coupling $\pi^* \in \mathrm{argmax}_{\pi \in \Pi(\mu, \nu)} \int \langle x, y \rangle d\pi(x, y)$, we have the inequality:

$$(25) \qquad \int \langle x, y \rangle d\pi^*(x, y) \geq \int |x|^2 d\mu(x).$$

Suppose this inequality is strict for some $\mu^* \in \mathrm{argmax}_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu)$. As $\mathcal{D}$ is a cone, $\mu_\lambda^* = \lambda_{\#}\mu^* \in \mathcal{D}$ for any $\lambda \geq 0$. Let $\pi_\lambda^* = \big((x, y) \mapsto (\lambda x, y)\big)_{\#}\pi^*$ be the corresponding coupling of μ_λ^* and ν. If the inequality (25) is strict, there exists a $\lambda > 1$ such that:

$$\int \langle x, y \rangle d\pi_\lambda^* = \lambda \int \langle x, y \rangle d\pi^* \geq \lambda^2 \int |x|^2 d\mu^* = \int |x|^2 d\mu_\lambda^*,$$

meaning that $\mu_\lambda^* \preceq_K \nu$. However, the variance of μ_λ^* exceeds that of μ^* (unless $\mu^* = \delta_0$, in which case the result is trivial), contradicting the optimality of $\mu^* \in \mathrm{argmax}_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu)$. We conclude that the inequality (25) must hold with equality for any optimal μ^* and π^*.

To see $\int x d\mu^* = 0$ if $\mathcal{D}$ is translation invariant, let $k = -\int x d\mu^*$ and $\mu_k^* = T_{k\#}\mu^*$ be defined as in Lemma 4.7. Setting a coupling $\pi_k^* = \left((x,y) \mapsto (x+k,y)\right)_\# \pi^*$ of μ_k^* and ν and recalling that ν is centered, the following calculation

$$\int |x|^2 d\mu_k^* = \int |x+k|^2 d\mu^* \leq \int |x|^2 d\mu^* = \int \langle x,y \rangle d\pi^* = \int \langle x+k,y \rangle d\pi^* = \int \langle x,y \rangle d\pi_k^*$$

shows $\mu_k^* \preceq_K \nu$. With this, $\mathrm{Var}(\mu_k^*) = \mathrm{Var}(\mu^*)$ implies μ_k^* is also optimal. Now $k \neq 0$ yields strict inequality above, which contradicts (24). We conclude $k = 0$. $\qquad\square$

We can now extend the solution existence result for the problem (22) when $\Omega = \mathbb{R}^d$.

THEOREM 4.9. *Let $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$, and let $\mathcal{D}$ be a cone. Suppose that $\mathcal{D}$ is either translation invariant or $\mathcal{D} \subset \mathcal{P}_{2,0}(\mathbb{R}^d)$. Then, the problem (22) with $\Omega = \mathbb{R}^d$ is equivalent to the relaxed problem (8), meaning that both formulations have the same set of solutions.*

PROOF. For any $\mu^* \in \underset{\mu \in \mathcal{D}}{\mathrm{argmin}}\, \mathcal{W}_2^2(\mu,\nu)$ and $\pi^* \in \underset{\pi \in \Pi(\mu^*,\nu)}{\mathrm{argmax}} \int \langle x,y \rangle d\pi(x,y)$, we claim

$$(26) \qquad \int \langle x, y-x \rangle \, d\pi^*(x,y) = 0,$$

that is, $\mu^* \preceq_K \nu$. To see this, for any $\mu \in \mathcal{D} \setminus \{\delta_0\}$, consider a rescaling factor λ^* that solves

$$\min_{\lambda \in \mathbb{R}} \mathcal{W}_2^2(\lambda_\# \mu, \nu) = \min_{\lambda \in \mathbb{R}} \left(\lambda^2 \int |x|^2 d\mu - 2\lambda \max_{\pi \in \Pi(\mu,\nu)} \int \langle x,y \rangle d\pi(x,y) + \int |y|^2 d\nu \right).$$

The optimal scaling is attained at:

$$\lambda^* = \frac{\max_{\pi \in \Pi(\mu,\nu)} \int \langle x,y \rangle d\pi(x,y)}{\int |x|^2 d\mu}.$$

Since $\lambda^* = 1$ for the optimal μ^* (no rescaling improves the objective function), (26) follows.

Next, by the assumption on $\mathcal{D}$ and Lemma 4.7, $\int x d\mu^* = \int y d\nu = 0$. This and (26) imply

$$\min_{\mu \in \mathcal{D}} \mathcal{W}_2^2(\mu,\nu) = \min_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K \cap \mathcal{P}_{2,0}(\mathbb{R}^d)} \mathcal{W}_2^2(\mu,\nu)$$

$$= \min_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K \cap \mathcal{P}_{2,0}(\mathbb{R}^d)} \left(\int |y|^2 d\nu - \int |x|^2 d\mu \right)$$

$$(27) \qquad = \mathrm{Var}(\nu) - \max_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu),$$

where we have used Lemma 4.8 to remove $\bigcap \mathcal{P}_{2,0}(\mathbb{R}^d)$ in (27). This shows any solution to the relaxed problem (8) is also a solution to (22) with $\Omega = \mathbb{R}^d$. Conversely, if μ^* solves (22), the equality of values (27) and Lemma 4.8 imply that μ^* solves (8), completing the proof. $\qquad\square$

REMARK 9. Observe that the proofs of Lemma 4.8 and Theorem 4.9 do not rely on full translation invariance of $\mathcal{D}$. Instead, the following weaker condition suffices (recalling that these results assume ν is centered): for any $\mu \in \mathcal{D}$ with $k = -\int x d\mu(x)$, the centered measure $\mu_k = T_{k\#}\mu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$ also belongs to $\mathcal{D}$.

4.3. *Weak optimizer closedness and equivalence to the problem with self-consistency.* We have shown that for appropriate domains, the problem (22) is equivalent to the fully relaxed problem (8). We now consider when (22) is equivalent to the problem with the full self-consistency condition (5). Recall the π-conditional barycentric map from Definition 2.1.

DEFINITION 4.10. Let $\nu \in \mathcal{P}_{2,0}(\Omega)$ and $\mathcal{D} \subset \mathcal{P}_2(\Omega)$. We say that $\mathcal{D}$ is closed under weak optimizers if, for any $\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K$ that solves (22), there exists $\pi \in \Pi(\mu, \nu)$ such that

$$(28) \qquad \int \langle x, y \rangle d\pi(x, y) \geq \int |x|^2 d\mu(x) \quad \text{and} \quad c_{\pi\#}(\mu) \in \mathcal{D}.$$

For instance, the domains $\mathcal{D}^m$ and $\mathcal{D}_u^m$ in Example 3 are closed under weak optimizers.

THEOREM 4.11. *If $\mathcal{D}$ is closed under weak optimizers, then every optimizer μ for the problem (22) satisfies $\mu \preceq_C \nu$, and consequently, μ solves the original problem (7).*

PROOF. Assume that μ solves (22), so there exists $\pi \in \Pi(\mu, \nu)$ satisfying (28). Since $\gamma = \big((x, y) \mapsto (c_\pi(x), y)\big)_\# \pi \in \mathcal{M}(c_{\pi\#}(\mu), \nu)$ is a martingale measure, we have

$$\int |x|^2 dc_{\pi\#}(\mu) - \int \langle x, y \rangle d\gamma = 0 \geq \int |x|^2 d\mu - \int \langle x, y \rangle d\pi.$$

Now, as the barycenter minimizes the expected squared distance, we have the inequality

$$\int \frac{1}{2} |x - y|^2 d\gamma(x, y) \leq \int \frac{1}{2} |x - y|^2 d\pi(x, y),$$

with strict inequality if $c_{\pi\#}(\mu) \neq \mu$. After canceling $\int |y|^2 d\gamma = \int |y|^2 d\pi = \int |y|^2 d\nu$, we get

$$\int \left(-\frac{|x|^2}{2} + \langle x, y \rangle \right) d\gamma \geq \int \left(-\frac{|x|^2}{2} + \langle x, y \rangle \right) d\pi,$$

with strict inequality if $c_{\pi\#}(\mu) \neq \mu$. Adding this to the previous inequality gives

$$\text{Var}(c_{\pi\#}(\mu)) = \int |x|^2 dc_{\pi\#}(\mu) \geq \int |x|^2 d\mu \geq \text{Var}(\mu).$$

The optimality of μ implies equality. Hence $c_{\pi\#}(\mu) = \mu$, which implies $\mu \preceq_C \nu$. □

4.4. *Relationship with principal component analysis.* In this section, we explore the relationship between our formulation (22) and principal component analysis (PCA). Let V_m denote the set of all m-dimensional subspaces of $\mathbb{R}^d$. Consider the following cone domain:

$$\mathcal{D}_m := \{\mu \in \mathcal{P}_{2,0}(\mathbb{R}^d) \,|\, \mu(E) = 1 \text{ for some } E \in V_m\}.$$

Let $p_E : \mathbb{R}^d \to E$ denote the orthogonal projection map onto a subspace E of $\mathbb{R}^d$.

LEMMA 4.12. *Any μ solving the problem $\max\limits_{\mu \in \mathcal{D}_m, \mu \preceq_K \nu} \text{Var}(\mu)$, where $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$, is the orthogonal projection of ν onto some $E \in V_m$.*

PROOF. Theorem 4.9 shows the problem $\max_{\mu \in \mathcal{D}_m, \mu \preceq_K \nu} \text{Var}(\mu)$ is equivalent to the relaxed problem $\min_{\mu \in \mathcal{D}_m} \mathcal{W}_2^2(\mu, \nu)$. Noting $\mathcal{D}_m = \cup_{L \in V_m} \mathcal{D}_m^E$, where $\mathcal{D}_m^E := \{\mu \in \mathcal{D}_m \,|\, \mu(E) = 1\}$, we can decompose the relaxed problem as $\min_{E \in V_m} \min_{\mu \in \mathcal{D}_m^E} \mathcal{W}_2^2(\mu, \nu)$. The lemma follows by the fact that the projection $p_{E\#}\nu$ is the unique minimizer of $\min_{\mu \in \mathcal{D}_m^E} \mathcal{W}_2^2(\mu, \nu)$. □

Lemma 4.12 reveals that PCA can be viewed as a particular case of problem (22) with the domain $\mathcal{D}_1$. Specifically, the first principal component is defined as the direction that maximizes the variance of the projected data. Lemma 4.12 demonstrates that the projected data satisfies the variance maximization problem under the Kantorovich dominance constraint.

We now consider the following version of the PCA in [6, Section 3.3]. Consider the model

$$Y = L^*W + R, \tag{29}$$

where $W \sim \mathcal{N}(0, I_m)$ is an m-dimensional Gaussian vector of latent factors, $L^* \in \mathbb{R}^{d \times m}$ is an unknown factor loading matrix of rank m, and $R \sim \mathcal{N}(0, \sigma^2 I_d)$ represents random noise that is independent of W and cannot be explained by the latent factor. Without loss of generality, assume that $L^* = U^*(\Lambda^*)^{1/2}$, where the columns of $U^* \in \mathbb{R}^{d \times m}$ form an orthonormal set, and $\Lambda^* = \operatorname{diag}[\lambda_1^*, \ldots, \lambda_m^*]$ is a diagonal matrix with $\lambda_1^* \geq \cdots \geq \lambda_m^* > 0$.

Let $\nu = \mathcal{L}(Y)$. We now focus on solving the problem (22) over the following cone domain

$$\mathcal{D} = \{\mu_L = \mathcal{L}(LW) \mid L = U\Lambda^{1/2}, \text{ where the columns of } U \in \mathbb{R}^{d \times m} \text{ are orthonormal,} \tag{30}$$

$$\text{and } \Lambda = \operatorname{diag}[\lambda_1, \ldots, \lambda_m] \text{ is a diagonal matrix with } \lambda_1 \geq \cdots \geq \lambda_m \geq 0\}.$$

Note that the problem (22) remains equivalent to the relaxed problem (8), by Theorem 4.9.

THEOREM 4.13. *Let $\nu_n, \nu \in \mathcal{P}_{2,0}(\mathbb{R}^d)$ and $\mathcal{W}_2(\nu_n, \nu) \to 0$. Then for all n, $\mathcal{D} \cap \mathcal{M}_{\nu_n}^K \neq \emptyset$ with $\mathcal{D}$ in (30), and for any $\mu_{L_n} \in \underset{\mu \in \mathcal{D} \cap \mathcal{M}_{\nu_n}^K}{\operatorname{argmax}} \operatorname{Var}(\mu)$ with $L_n = U_n\Lambda_n^{1/2}$,*

$$L_nL_n^\mathsf{T} - \sigma_n^2 U_nU_n^\mathsf{T} \to L^*L^{*\mathsf{T}} \quad \text{and} \quad \sigma_n^2 \to \sigma^2 \quad \text{as} \quad n \to \infty, \tag{31}$$

where $\sigma_n^2 := \frac{1}{d-m}\int |y - U_nU_n^\mathsf{T}y|^2 d\nu_n(y)$ is an estimator of the noise variance σ^2.

REMARK 10. (31) indicates that the optimization problem $\max_{\mu \in \mathcal{D} \cap \mathcal{M}_{\nu_n}^K} \operatorname{Var}(\mu)$ can recover $\mathcal{L}(L^*W)$ as the empirical distribution ν_n converges to the population distribution ν. This result cannot hold if the full convex order constraint is imposed. In many practical cases, such as when ν_n is discrete (e.g., ν_n is an empirical measure sampled from ν), the convex order condition $\mu \preceq_\mathrm{C} \nu_n$ fails for any Gaussian $\mu \in \mathcal{D}$ unless $\mu = \delta_0$. As a result, the domain $\mathcal{D} \cap \{\mu \preceq_\mathrm{C} \nu_n\}$ reduces to $\{\delta_0\}$.

LEMMA 4.14. *For $\mathcal{D}$ in (30), the set $\mathcal{D} \cap \{\mu \mid \operatorname{Var}(\mu) \leq \delta\}$ is $\mathcal{W}_2$-compact for any $\delta > 0$.*

PROOF. For any $\mu_L \in \mathcal{D}$ with $L = U\Lambda^{1/2}$, $\operatorname{Var}(\mu_L) = \sum_{i=1}^m \lambda_i$. Thus, $\max_i \lambda_i \leq \delta$. □

LEMMA 4.15. *Let $\mu_n = \mathcal{L}(U_n\Lambda_n^{1/2}W)$ and $\mu = \mathcal{L}(U\Lambda^{1/2}W)$ as in (30). If $\mathcal{W}_2(\mu_n, \mu) \to 0$, then $U_n\Lambda_nU_n^\mathsf{T} \to U\Lambda U^\mathsf{T}$ and $\Lambda_n \to \Lambda$. If $\operatorname{rank}(\Lambda) = m$, then $U_nU_n^\mathsf{T} \to UU^\mathsf{T}$ as well.*

PROOF. Assume that $\lambda_1 > \cdots > \lambda_k > 0 = \lambda_{k+1} = \cdots = \lambda_m$ in the diagonal of Λ. Since $\mathcal{W}_2(\mu_n, \mu) \to 0$ implies $\operatorname{Var}(\mu_n)$ is bounded, the sequences $\{U_n\}$ and $\{\Lambda_n\}$ are precompact. For any subsequences $\{U_k\}$ of $\{U_n\}$ and $\{\Lambda_k\}$ of $\{\Lambda_n\}$ converging to $\tilde{U}$ and $\tilde{\Lambda}$, respectively, define $\tilde{\mu} = \mathcal{L}(\tilde{U}\tilde{\Lambda}^{1/2}W)$. Then, $\mathcal{W}_2(\mu_n, \mu) \to 0$ implies $\tilde{\mu} = \mu$, which ensures $\tilde{U}_i\tilde{U}_i^\mathsf{T} = U_iU_i^\mathsf{T}$ for $i = 1, \ldots, k$ where U_i is the ith column of U (notice this yields $\tilde{U}\tilde{U}^\mathsf{T} = UU^\mathsf{T}$ if $\lambda_m > 0$), and $\tilde{\Lambda} = \Lambda$, i.e., $\tilde{\lambda}_i = \lambda_i$ for $i = 1, \ldots, m$. Consequently, $\tilde{U}\tilde{\Lambda}\tilde{U}^\mathsf{T} = U\Lambda U^\mathsf{T}$. The arbitrariness of the subsequence establishes the lemma. The more general case, $\lambda_1 \geq \cdots \geq \lambda_k$, can be similarly proven by taking into account the multiplicity of the singular values $\lambda_1, \ldots, \lambda_k$. □
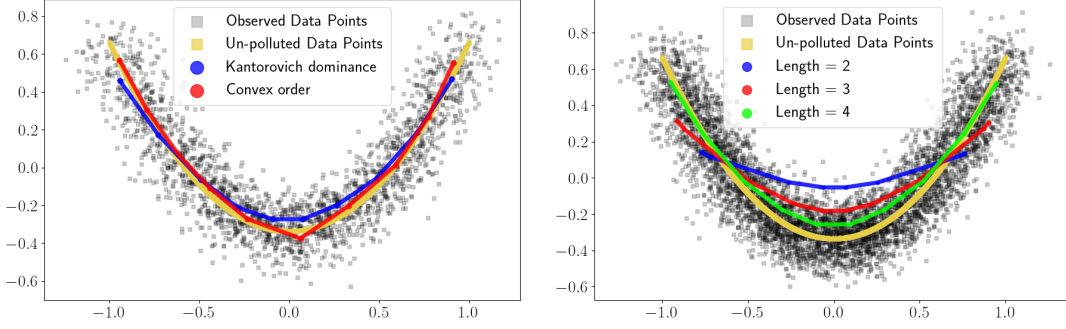
FIG 5. *Left: Optimal measure μ under the Kantorovich dominance and the convex order for 2000 data points. Right: Optimal measure μ under the Kantorovich dominance for 5000 data points.*

PROOF OF THEOREM 4.13. $\mathcal{D} \cap \mathcal{M}_{\nu_n}^K$ is compact by Lemma 4.14, and $\delta_0 \in \mathcal{D} \cap \mathcal{M}_{\nu_n}^K$. Notice the unique solution to $\max_{\mu \in \mathcal{D} \cap \mathcal{M}_\nu^K} \mathrm{Var}(\mu)$ is $\mu^* = \mathcal{N}(0, U^*(\Lambda^* + \sigma^2 I_m)U^{*T})$, since $\nu = \mathcal{N}(0, U^*\Lambda^*U^{*T} + \sigma^2 I_d)$, and the $\mathcal{W}_2$-projection of ν onto $\mathcal{D}$ is clearly μ^*. Then for any $\mu_{L_n} \in \mathrm{argmax}_{\mu \in \mathcal{D} \cap \mathcal{M}_{\nu_n}^K} \mathrm{Var}(\mu)$, Lemma 4.14 and Theorem 4.5 give $\mathcal{W}_2(\mu_{L_n}, \mu^*) \to 0$. Then since $\mu_{L_n} = \mathcal{L}(U_n \Lambda_n^{1/2} W)$ and $\mu^* = \mathcal{L}(U^*(\Lambda^* + \sigma^2 I_m)^{1/2}W)$, Lemma 4.15 gives $U_n \Lambda_n U_n^\mathsf{T} \to U^*(\Lambda^* + \sigma^2 I_m)U^{*\mathsf{T}}$, $\Lambda_n \to \Lambda^* + \sigma^2 I_m$ and $U_n U_n^\mathsf{T} \to U^* U^{*\mathsf{T}}$. This with $\nu_n \to \nu$ implies $\sigma_n^2 \to \sigma^2 = \frac{1}{d-m} \int |y - UU^\mathsf{T} y|^2 d\nu(y)$, yielding $L_n L_n^\mathsf{T} - \sigma_n^2 U_n U_n^\mathsf{T} \to U^*\Lambda^*U^{*\mathsf{T}}$. □

5. Numerical examples.

We provide examples of numerically solving the KDR problem $\max_{\mu \in \mathcal{D}, \mu \preceq_\mathrm{K} \nu} \mathrm{Var}(\mu)$ with $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ and three closely related discrete curve domains.

5.1. *Curves with bounded length.*

Our first example considers the following domain

$$\mathcal{D}_1 = \mathcal{D}_1(m, B) = \left\{ \mu = \sum_{i=1}^m u_i \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, \, u_i \geq 0, \, \sum_{i=1}^m u_i = 1, \, L(\mu) \leq B \right\},$$

where $L(\mu) = \sum_{i=1}^{m-1} \|x_{i+1} - x_i\|$ represents the length of the discrete curve μ. Due to the length bound, $\mathcal{D}_1$ is not a cone. To address the bound, we transform the constrained optimization problem into an unconstrained form by introducing a Lagrangian with multipliers that correspond to each constraint. Details are given in Section B of the Supplementary Material. The method employs gradient ascent to update the position $x = (x_i)_{i=1}^m \in (\mathbb{R}^d)^m$ and weight $u = (u_i)_{i=1}^m$, while applying gradient descent to update the Lagrange multipliers iteratively.

It is possible to numerically solve problem (5), particularly the martingale constraint, using gradient descent. However, this approach involves optimizing over a larger set of unknowns for the coupling π rather than focusing on μ alone, which leads to increased memory requirements for storing variables, greater computational demand for calculating gradients, and potentially longer convergence times. In contrast, although calculating the coupling π is still required to compute the $\mathcal{W}_2$ distance for the KDR problem, it can be done efficiently using the Sinkhorn algorithm [7], [5], which is known for its speed and computational efficiency.

To illustrate our example, we consider ν as a discrete measure, defined over either $n = 2000$ or 5000 points, where $Y = (Z, Z^2) + \varepsilon$, with Z being a one-dimensional variable uniformly distributed over $[-1, 1]$, and ε representing Gaussian noise. The initial measure μ is supported on $m = 10$ points, which are initialized along the first principal component of ν.

Using the proposed method, we optimize the location **x** and the weight **u**. In the case of 2000 data points, we set the length bound $B = 4.0$ and the left side of Figure 5 illustrates

the optimal μ under both the Kantorovich dominance relation and the convex order relation. For the case with 5000 data points, an out-of-memory error occurred when attempting to compute the optimal measure under the convex order. Therefore, we present only the optimal measure under the KDR, with $B = 2.0, 3.0$ and 4.0, as shown in the right side of Figure 5. [4]

Further improvement in performance is achieved in the case of cone domains, as enabled by the following result.

PROPOSITION 5.1. *Assume that $\mathcal{D}$ is a cone and translation invariant. Then, solving the problem* (22) *with $\Omega = \mathbb{R}^d$ is equivalent to solving the following problem:*

$$(32) \qquad P_1 = \max_{\substack{\mu \in \mathcal{D} \cap \mathcal{P}_{2,0}(\mathbb{R}^d), \, \pi \in \Pi(\mu,\nu) \\ \mathrm{Var}(\mu) \leq 1}} \int \langle x, y \rangle d\pi(x,y).$$

PROOF. By Theorem 4.9, the problem (22) is equivalent to the problem $\min_{\mu \in \mathcal{D}} \mathcal{W}_2^2(\mu, \nu)$, which, by Lemma 4.8 and equation (26) in the proof of Theorem 4.9, is also equivalent to:

$$(33) \qquad P_2 = \max_{\substack{\mu \in \mathcal{D} \cap \mathcal{P}_{2,0}(\mathbb{R}^d), \, \pi \in \Pi(\mu,\nu) \\ \int \langle x,y \rangle d\pi(x,y) = \int |x|^2 d\mu}} \int \langle x, y \rangle d\pi(x,y).$$

We will show that any optimizer (μ, π) for (33) induces a solution to (32), and conversely. For $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\pi \in \Pi(\mu, \nu)$ and $\lambda > 0$, define $\mu_\lambda = \lambda_{\#}\mu$ and $\pi_\lambda = (\lambda \times \mathrm{Id})_{\#}\pi \in \Pi(\mu_\lambda, \nu)$. Then for any feasible pair $(\hat{\mu}, \hat{\pi})$ for (32) and $(\check{\mu}, \check{\pi})$ for (33) with $\int \langle x, y \rangle d\hat{\pi} > 0$ and $\check{\mu} \neq \delta_0$,

$$(34) \qquad 1 = \int |x|^2 d\check{\mu}_{\check{\lambda}} = \check{\lambda}^2 \int |x|^2 d\check{\mu} = \check{\lambda}^2 \int \langle x, y \rangle d\check{\pi} = \check{\lambda} \int \langle x, y \rangle d\check{\pi}_{\check{\lambda}}, \text{ and}$$

$$(35) \qquad \hat{\lambda} \int \langle x, y \rangle d\hat{\pi} = \int \langle x, y \rangle d\hat{\pi}_{\hat{\lambda}} = \int |x|^2 d\hat{\mu}_{\hat{\lambda}} = \hat{\lambda}^2 \int |x|^2 d\hat{\mu} \leq \hat{\lambda}^2,$$

where $\check{\lambda} = \sqrt{1/\mathrm{Var}(\check{\mu})}$ so that $\mathrm{Var}(\check{\mu}_{\check{\lambda}}) = 1$, while $\hat{\lambda} > 0$ is the unique constant yielding the second equality in (35). This shows $(\check{\mu}_{\check{\lambda}}, \check{\pi}_{\check{\lambda}})$ is feasible for (32) and $(\hat{\mu}_{\hat{\lambda}}, \hat{\pi}_{\hat{\lambda}})$ for (33), and moreover, if $P_1, P_2 > 0$, then for any optimal pair $(\hat{\mu}, \hat{\pi})$ for (32) and $(\check{\mu}, \check{\pi})$ for (33), we have

$$(36) \qquad\qquad\qquad \hat{\lambda} = 1/\check{\lambda},$$

since for any optimal pair $(\hat{\mu}, \hat{\pi})$ for (32), the constraint $\mathrm{Var}(\hat{\mu}) \leq 1$ must be tight, meaning that the inequality in (35) is satisfied as an equality. Then (36), with (34) and (35), shows that for any optimal $(\hat{\mu}, \hat{\pi})$ for (32), its scaling $(\hat{\mu}_{\hat{\lambda}}, \hat{\pi}_{\hat{\lambda}})$ is optimal for (33), and conversely, for any optimal $(\check{\mu}, \check{\pi})$ for (33), $(\check{\mu}_{\check{\lambda}}, \check{\pi}_{\check{\lambda}})$ is optimal for (32). Finally, the proof also shows $P_1 = 0$ if and only if $P_2 = 0$, in which case $\mu = \delta_0$ serves as the trivial solution to both problems. $\square$

In the following, we apply this result to solve two examples with cone domains[5]. In both cases, we use fixed weights for μ, specifically $\mu = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_i}$, to approximate a dataset of $n = 300$ distributed along a step-shaped curve with added noise.

---

[4]The Python code for Example 5.1 is available at https://github.com/joshuaHiew/Kantorovich_dominance.

[5]Source code for examples 5.2 and 5.3 are available at https://github.com/souza-m/data-denoising.

5.2. *Curves with bounded length-to-standard-deviation ratio.* Here we consider the following modification of the domain $\mathcal{D}_1$:

$$\mathcal{D}_2 = \left\{ \mu = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, \, \frac{L(\mu)}{\mathrm{SD}(\mu)} \le B \right\},$$

where $\mathrm{SD}(\mu) = \sqrt{\mathrm{Var}(\mu)}$ represents the standard deviation of μ. Thus, the bound is now imposed on the ratio between the length and the standard deviation. Since rescaling μ does not change the ratio $L(\mu)/\mathrm{SD}(\mu)$, we see that $\mathcal{D}_2$ is a cone.

For each value of B, we calculate $m = 100$ points $x_1, ..., x_{100}$ and linearly connect them. Figure 6 illustrates the resulting curves for two values of B. We include an optimal solution under the convex order constraint, with size $m = 10$; higher values resulted in longer processing times or even out-of-memory error. Both methods seem to capture the underlying structure of the data reasonably well. Supplementary Material C describes the alternating numerical steps for the solution.
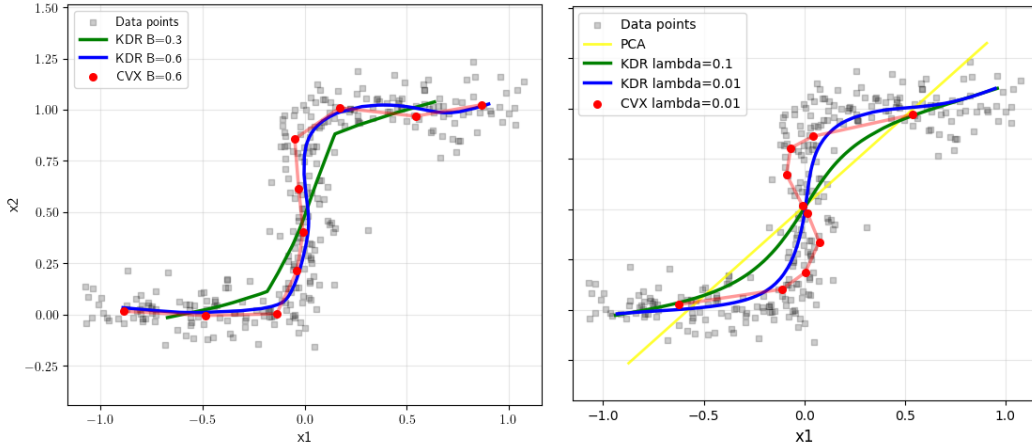

FIG 6. *Left: Curves with bounded length-standard deviation ratios. Right: Curves with bounded curvatures. Curves are formed by connecting points using straight lines.*

5.3. *Curves with bounded curvature.* Our last example considers a domain given by

$$\mathcal{D}_3 = \left\{ \mu = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i} \,\middle|\, x_i \in \mathbb{R}^d, \, \phi(\mu) \le B \right\},$$

where $\phi(\mu) = \sum_{i=2}^{m-1} \cos^2 \frac{\theta_i}{2}$ represents the total curvature, with θ_i being the angle between segments $\overline{x_{i-1}x_i}$ and $\overline{x_i x_{i+1}}$. $\mathcal{D}_2$ is a cone since the angles do not change with scaling of μ, and by Lemma 4.12, $B = 0$ corresponds to the problem of finding the first principal direction.

In the numerical computation, the curvature constraint is handled indirectly via penalization and solved using an alternating method, as detailed in the Supplementary Material C. For each curvature penalty parameter λ, we set $m = 100$ points and connect them linearly to form curves in Figure 6, where the first principal direction is also included for reference. A solution under the convex order constraint (with $m = 10$ as in Section 5.2) is also included. In this example, we observe a slight convergence of the latter towards the barycenter, which can be due to the stricter nature of the convex order or to numerical approximation errors.

The examples in this section demonstrate that the proposed method can efficiently compute the optimal measure under the Kantorovich dominance with large datasets.

REFERENCES

[1] Luigi Ambrosio and Nicola Gigli.
A User's Guide to Optimal Transport, pages 1–155.
Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[2] Mathias Beiglböck and Nicolas Juillet.
On a problem of optimal transport under marginal martingale constraints.
The Annals of Probability, 44(1):42 – 106, 2016.

[3] Stephen Boyd and Lieven Vandenberghe.
Convex optimization.
Cambridge university press, 2004.

[4] Martin Brückerhoff and Nicolas Juillet.
Instability of martingale optimal transport in dimension $d \geq 2$.
Electronic Communications in Probability, 27:1–10, 2022.

[5] Guillaume Carlier.
On the linear convergence of the multimarginal sinkhorn algorithm.
SIAM Journal on Optimization, 32(2):786–794, 2022.

[6] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma.
Spectral methods for data science: A statistical perspective.
Foundations and Trends® in Machine Learning, 14(5):566–806, 2021.

[7] Marco Cuturi.
Sinkhorn distances: Lightspeed computation of optimal transport.
In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[8] Hadrien De March and Nizar Touzi.
Irreducible convex paving for decomposition of multidimensional martingale transport plans.
The Annals of Probability, 47(3):1726–1774, 2019.

[9] Alex Delalande and Quentin Mérigot.
Quantitative stability of optimal transport maps under variations of the target measure.
Duke Math. J., 172(17):3321–3357, 2023.

[10] Tom Duchamp and Werner Stuetzle.
Extremal properties of principal curves in the plane.
The Annals of Statistics, 24(4):1511 – 1520, 1996.

[11] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina.
Principal component analysis.
Nature Reviews Methods Primers, 2(1):100, 2022.

[12] Trevor Hastie and Werner Stuetzle.
Principal curves.
Journal of the American Statistical Association, 84(406):502–516, 1989.

[13] Pierre Henry-Labordère.
Model-free hedging: A martingale optimal transport viewpoint.
Chapman and Hall/CRC, 2017.

[14] I.T. Jolliffe.
Principal Component Analysis.
Springer Series in Statistics. Springer, 2nd ed. edition, 2002.

[15] Seungwoo Kang and Hee-Seok Oh.
Probabilistic principal curves on riemannian manifolds.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[16] Tapas Kanungo, David Mount, Nathan Netanyahu, Christine Piatko, Ruth Silverman, and Angela Wu.
An efficient k-means clustering algorithm: Analysis and implementation.
IEEE transactions on pattern analysis and machine intelligence, 24(7):881–892, 2002.

[17] Balázs Kégl, Adam Krzyzak, Tamás Linder, and Kenneth Zeger.
Learning and design of principal curves.
IEEE transactions on pattern analysis and machine intelligence, 22(3):281–297, 2000.

[18] Dmitry Kramkov and Yan Xu.
An optimal transport problem with backward martingale constraints motivated by insider trading.
The Annals of Applied Probability, 32(1):294–326, 2022.

[19] Jongmin Lee, Jang-Hyun Kim, and Hee-Seok Oh.
Spherical principal curves.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6):2165–2171, 2020.

[20] Stuart Lloyd.
Least squares quantization in pcm.
IEEE transactions on information theory, 28(2):129–137, 1982.

[21] George J. Minty.
Monotone (nonlinear) operators in Hilbert space.
Duke Math. J., 29:341–346, 1962.

[22] Filippo Santambrogio.
Optimal transport for applied mathematicians.
Birkäuser, NY, 55(58-63):94, 2015.

[23] Robert Tibshirani.
Principal curves revisited.
Statistics and computing, 2:183–190, 1992.

[24] Cédric Villani.
Topics in Optimal Transportation.
Number 58. American Mathematical Soc., 2003.

[25] Johannes Wiesel and Erica Zhang.
An optimal transport-based characterization of convex order.
Dependence Modeling, 11(1):20230102, 2023.

SUPPLEMENTARY MATERIAL

A: Solution comparison between the problem (5) and the problem studied in [18].
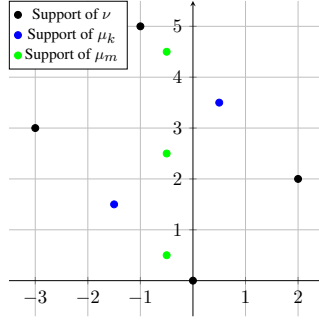
We present an example which shows that the optimal measure for problem (1.2) in [18] is not an optimal measure for our problem (5). We define the measure μ_k, μ_m and ν as follows:

$$\mu_k = \tfrac{1}{2}\delta_{(-3/2,3/2)} + \tfrac{1}{2}\delta_{(1/2,7/2)},$$

$$\mu_m = \tfrac{3}{10}\delta_{(-1/2,1/2)} + \tfrac{2}{5}\delta_{(-1/2,5/2)} + \tfrac{3}{10}\delta_{(-1/2,9/2)},$$

$$\nu = \tfrac{1}{4}\delta_{(0,0)} + \tfrac{1}{4}\delta_{(-3,3)} + \tfrac{1}{4}\delta_{(2,2)} + \tfrac{1}{4}\delta_{(-1,5)}.$$

We define the martingale coupling π_k between μ_k and ν as follows:

$$\pi_k = \tfrac{1}{4}\delta_{((-3/2,3/2),(0,0))} + \tfrac{1}{4}\delta_{((-3/2,3/2),(-3,3))} + \tfrac{1}{4}\delta_{((1/2,7/2),(2,2))} + \tfrac{1}{4}\delta_{((1/2,7/2),(-1,5))},$$

while the martingale coupling π_m between μ_m and ν as follows:

$$\pi_m = \tfrac{1}{4}\delta_{((-1/2,1/2),(0,0))} + \tfrac{1}{20}\delta_{((-1/2,1/2),(-3,3))} + \tfrac{1}{5}\delta_{((-1/2,5/2),(-3,3))}$$

$$+ \tfrac{1}{4}\delta_{((-1/2,9/2),(-1,5))} + \tfrac{1}{20}\delta_{((-1/2,9/2),(2,2))} + \tfrac{1}{5}\delta_{((-1/2,5/2),(2,2))}.$$

For $u = (u_1, u_2)$, $v = (v_1, v_2) \in \mathbb{R}^2$, let $c_k(u, v) = (u_1 - v_1)(u_2 - v_2)$ be the cost function considered in [18]. It is straightforward to check that for any $(x_0, y_0), (x_1, y_1) \in \mathrm{spt}(\pi_k)$,

$$(37) \qquad (1 - t)c_k(x_0, y_0) + tc_k(x_1, y_1) \le t(1 - t)c_k(y_0, y_1), \qquad t \in [0, 1].$$

By [18, Theorem 2.2], (37) shows that π_k is the optimal coupling for their problem. On the other hand, it is also easy to check that π_m provide a better coupling for our problem (5), showing the optimal measure for problem (1.2) in [18] is not the optimal measure for (5).

B: Computational details for Section 5.1.

The Lagrangian we maximize for the example in Section 5.1 is the following:

$$\mathcal{L}(x, u, \lambda_1, \lambda_{1,i}, \lambda_2, \lambda_3) = \left(\sum_i \|x_i\|^2 u_i - \|\sum_i x_i u_i\|^2 \right) - \lambda_1 \left(\sum_i u_i - 1 \right) + \sum_i \lambda_{1,i} u_i$$

$$- \lambda_2 \left(\sum_i \|x_{i+1} - x_i\| - B \right) - \lambda_3 \left(\mathcal{W}_2^2(\mu, \nu) - \left(\sum_j \|y_j\|^2/n - \sum_i \|x_i\|^2 u_i \right) \right),$$

where $\sum_i \|x_i\|^2 u_i - \|\sum_i x_i u_i\|^2$ is the variance of μ, $\lambda_1 \in \mathbb{R}$ is the Lagrange multiplier for the probability constraint $\sum_i u_i = 1$, $\lambda_{1,i} \ge 0$ enforces $u_i \ge 0$, $\lambda_2 \ge 0$ enforces the length constraint $L(\mu) \le B$, and $\lambda_3 \ge 0$ enforces the Kantorovich dominance constraint.

The complementary slackness condition for each constraint ensures that the Lagrange multipliers only contribute when their respective constraints are active. Specifically:

$$\lambda_1\Big(\sum_i u_i - 1\Big) = 0, \quad \lambda_{1,i} u_i = 0 \;\; \forall i, \quad \lambda_2\Big(B - \sum_{i=1}^{m-1} \|x_{i+1} - x_i\|\Big) = 0, \quad \text{and}$$

$$\lambda_3\Big(\sum_j \|y_j\|^2/n - \sum_i \|x_i\|^2 u_i - \mathcal{W}_2^2(\mu,\nu)\Big) = 0.$$

Let π be an optimal coupling corresponding to $\mathcal{W}_2^2(\mu,\nu)$. The gradients with respect to the location $x = (x_i)_{i=1}^m \in (\mathbb{R}^d)^m$ and weight $u = (u_i)_{i=1}^m \in [0,1]^m$ are computed as follows. *Gradient with respect to x:*

$$\frac{\partial \mathcal{L}}{\partial x_i} = 2x_i u_i - 2\Big(\sum_j x_j u_j\Big) u_i - 2\lambda_3 x_i u_i - 2\lambda_3 \sum_j \pi_{ij}(x_i - y_j) + \lambda_2\Big(\frac{x_{i+1} - x_i}{\|x_{i+1} - x_i\|} - \frac{x_i - x_{i-1}}{\|x_i - x_{i-1}\|}\Big).$$

Gradient with respect to u:

$$\frac{\partial \mathcal{L}}{\partial u_i} = \|x_i\|^2 - 2\Big(\sum_j x_j u_j\Big) x_i - \lambda_1 + \lambda_{1,i} - \lambda_3\Big(\|x_i\|^2 + \sum_j \pi_{ij}\|x_i - y_j\|^2\Big).$$

The optimal transport plan (π_{ij}) with marginal μ using current (x_i) and (u_i) and given ν. is calculated using the Sinkhorn algorithm for efficiency.

The multipliers are updated using projected gradient descent to satisfy the KKT conditions, with non-negativity of the multipliers enforced by projecting onto the feasible region. Specifically, each update step is given by:

$$\lambda_1 \leftarrow \lambda_1 + \eta_{\lambda_1}\Big(1 - \sum_i \mu_i\Big),$$

$$\lambda_{1,i} \leftarrow \max\Big(0, \lambda_{1,i} - \eta_{\lambda_{1,i}} \mu_i\Big), \quad \forall i,$$

$$\lambda_2 \leftarrow \max\Big(0, \lambda_2 + \eta_{\lambda_2}\Big(\sum_{i=1}^{m-1} \|x_{i+1} - x_i\| - B\Big)\Big),$$

$$\lambda_3 \leftarrow \max\Big(0, \lambda_3 + \eta_{\lambda_3}\Big(\mathcal{W}_2^2(\mu,\nu) - \sum_j \|y_j\|^2 \nu_j + \sum_i \|x_i\|^2 \mu_i\Big)\Big).$$

The projection operator $\max(0, \cdot)$ ensures that $\lambda_{1,i}, \lambda_2$ and λ_3 remain non-negative, in line with the KKT requirements.

C: Computational details for Sections 5.2 and 5.3.

To approximate a solution to (32), we propose an alternating procedure that splits the problem into two subproblems, optimizing (μ, π) with respect to each variable separately. The variables are $x = (x_1, ..., x_m) \in (\mathbb{R}^d)^m$, where each $x_i = (x_i^1, ..., x_i^d)$, and $\pi = (\pi_{ij}) \in \mathbb{R}_{\geq 0}^{m \times n}$. These examples use the constant weight on the points of μ, that is, we set $\mu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$. π represents a transport plan / coupling between the variable μ and the data $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, subject to the marginal constraints $\sum_j \pi_{ij} = \frac{1}{m}$ for all i, and $\sum_i \pi_{ij} = \frac{1}{n}$ for all j.

The optimization begins with an initial set x^0 satisfying the constraints $\frac{1}{m}\sum_i x_i = 0$ and $\frac{1}{m}\sum_i \|x_i\|^2 \leq 1$ and the domain constraint (e.g., the set of PCA projections in both examples). At each iteration $t \geq 1$, we perform the following two steps, repeating until convergence. The final output is the pair (x, π), where x is rescaled by the factor $\hat{\lambda}$ given in (36).

Step 1. Given x, find π that solves

$$\max_{\pi \in \Pi(1/m, 1/n)} \sum_i \sum_j \pi_{ij} \langle x_i, y_j \rangle \quad \text{st.} \quad \sum_i \sum_j \pi_{ij} x_i = 0 \quad \text{and} \quad \sum_i \sum_j \pi_{ij} \|x_i\|^2 \leq 1.$$

Step 2. Given π, set $\bar{y} = \pi^t y$ and find x that solves

$$\max_{x \in \mathcal{D}} \sum_i \langle x_i, \bar{y}_i \rangle \quad \text{st.} \quad \frac{1}{m} \sum_i x_i = 0 \quad \text{and} \quad \frac{1}{m} \sum_i \|x_i\|^2 \leq 1.$$

In Step 1, note that the constraints on π are satisfied by construction, since the marginal condition imposes $\sum_i \sum_j \pi_{ij} x_i = \sum_i \frac{1}{m} x_i = 0$ and $\sum_i \sum_j \pi_{ij} \|x_i\|^2 = \sum_i \frac{1}{m} \|x_i\|^2 \leq 1$. Thus, the problem falls into the class of traditional optimal transport with fixed marginals. To solve it, in both examples we apply the Sinkhorn method, as noted in Subsection 5.1. Then to solve Step 2 in Subsection 5.2, we explicitly state the domain constraint on x as follows:

$$\max_{x \in (\mathbb{R}^d)^m} \sum_{i=1}^m \langle x_i, \bar{y}_i \rangle \quad \text{st.} \quad \sum_{i=1}^m x_i = 0, \ \sum_{i=1}^m \|x_i\|^2 \leq 1, \ \sum_{i=1}^{m-1} \|x_{i+1} - x_i\| \leq B.$$

Since the second constraint is clearly binding, together with the third constraints it implies that any solution to the above satisfies the domain constraint $\frac{\sum_{i=1}^{m-1} \|x_{i+1} - x_i\|}{\sum_{i=1}^m \|x_i\|^2} \leq B$. We convert this problem into a second-order conic program, which can be solved efficiently by interior point methods – see [3] for an overview. To do that, we define the variable to be optimized as

$$(x_1^1, \ldots, x_1^d, \ldots, x_m^1, \ldots, x_m^d, a_1, \ldots, a_{m-1})$$

where the superscript in x_i denotes the dimensional component and $a_1, \ldots, a_{m-1}$ is an auxiliary variable. The second-moment constraint on x is replaced by the equivalent

$$\sqrt{\sum_{i=1}^m \sum_{k=1}^d (x_i^k)^2} \leq 1.$$

The constraints are equivalently rewritten as

$$x_1^k + \ldots + x_m^k = 0 \qquad\qquad k = 1, \ldots, d,$$

$$\left\| x_1^1 \ldots x_1^d \ldots x_m^1 \ldots x_m^d \right\| \leq 1,$$

$$\|x_{i+1} - x_i\| \leq a_i \qquad\qquad i = 1, \ldots, m-1,$$

$$a_1 + \ldots + a_{m-1} = B.$$

We solve the problem in this format using Python and CVXOPT package.

In Step 2 of Subsection 5.3, we replace the domain constraint $\phi(\mu) \leq B$ by a linear penalization together with another, inner iteration loop, as follows. Call $a_i = x_{i+1} - x_i$ for $i = 1, \ldots, m-1$, and $e_i = \frac{a_i}{|a_i|} - \frac{a_{i-1}}{|a_{i-1}|}$ for $i = 2, \ldots, m-1$. Notice that $\phi(\mu) = \frac{1}{4} \sum_{i=2}^{m-1} \|e_i\|^2$. At iteration t, the problem is solved for ϕ^t defined as

$$\phi^0(x) = 0$$

$$\phi^t(x) = \frac{1}{4} \sum_{i=2}^{m-1} \left\langle e_i^t, \frac{x_{i+i}^{t-1} - x_i}{|a_i|^{t-1}} - \frac{x_i - x_{i-1}^{t-1}}{|a_{i-1}|^{t-1}} \right\rangle = \sum_{i=2}^{m-1} \varphi_i^t x_i + \text{constant},$$

where $\varphi_i^t = -\frac{1}{4} e_i^t \left(\frac{1}{|a_i|^{t-1}} + \frac{1}{|a_{i-1}|^{t-1}} \right)$. The penalized problem is written as

$$\max_{x \in (\mathbb{R}^d)^m} \sum_i \langle x_i, c_i \rangle \quad \text{st.} \quad \sum_i x_i = 0 \quad \text{and} \quad \sum_i \|x_i\|^2 \leq m,$$

where $c_i = \bar{y}_i - \lambda\varphi_i^t$ for some penalization multiplier λ. We solve it through the Lagrangian

$$\mathcal{L}(x, \alpha, \beta) = \sum_{k=1}^{d}\sum_{i=1}^{m} x_i^k c_i^k - \sum_{k=1}^{d}\sum_{i=1}^{m} \alpha^k x_i^k - \beta\Big[\sum_{k=1}^{d}\sum_{i=1}^{m}(x_i^k)^2 - m\Big].$$

The first derivatives are

$$\mathcal{L}_{x_i^k}(x, \alpha, \beta) = c_i^k - \alpha^k - 2\beta x_i^k,$$

$$\mathcal{L}_{\alpha^k}(x, \alpha, \beta) = -\sum_{i=1}^{m} x_i^k,$$

$$\mathcal{L}_{\beta}(x, \alpha, \beta) = m - \sum_{k=1}^{d}\sum_{i=1}^{m}(x_i^k)^2.$$

We can assume that the second-moment condition is binding and $\beta > 0$, since otherwise it would be possible to increase the objective function. The first order conditions imply

$$(i)\ c_i^k - \alpha^k - 2\beta x_i^k = 0 \implies \hat{x}_i^k = \frac{1}{2\beta}\left(c_i^k - \alpha^k\right) \quad \forall k, i,$$

$$(ii)\ \sum_{i=1}^{m} \hat{x}_i^k = \frac{1}{2\beta}\Big(\sum_{i=1}^{m} c_i^k - \alpha^k\Big) = 0 \implies \alpha^k = \sum_{i=1}^{m} c_i^k \quad \forall k,$$

$$(iii)\ \sum_{k=1}^{d}\sum_{i=1}^{m}(\hat{x}_i^k)^2 = \frac{1}{4\beta^2}\sum_{k=1}^{d}\sum_{i=1}^{m}\left(c_i^k - \alpha^k\right) = m$$

$$\implies \beta = \frac{1}{2}\sqrt{\frac{1}{m}\sum_{k=1}^{d}\sum_{i=1}^{m}\left(c_i^k - \alpha^k\right)^2}.$$

Replacing α and β gives x as a function of c. Finally, we update x^t partially at each iteration, as $x^t = \epsilon x_{\text{opt}}^t + (1 - \epsilon) x^{t-1}$ where x_{opt}^t solves the problem for ϕ^t. The loop stops when the sequence (x^t) converges to a fixed point x^*, and we get

$$\phi^*(\mu) = \frac{1}{4}\sum_{i=2}^{m-1}\left\langle e_i^*, \frac{x_{i+i}^* - x_i^*}{|a_i|^*} - \frac{x_i^* - x_{i-1}^*}{|a_{i-1}|^*}\right\rangle = \phi(\mu).$$