Journal of Machine Learning Research 26 (2025) 1-39 Submitted 9/24; Revised 8/25; Published 10/25

On Probabilistic Embeddings in Optimal Dimension Reduction

Ryan Murray RWMURRAY@NCSU.EDU
Department of Mathematics
North Carolina State University
Raleigh, NC 27695 USA

Adam Pickarski APPICKAR@NCSU.EDU
Division of Mathematics
North Carolina State University
Raleigh, NC 27695 USA

Editor: Quentin Berthet

Abstract

Dimension reduction algorithms are essential in data science for tasks such as data exploration, feature selection, and denoising. However, many non-linear dimension reduction algorithms are poorly understood from a theoretical perspective. This work considers a generalized version of multidimensional scaling, which seeks to construct a map from high to low dimension which best preserves pairwise inner products or norms. We investigate the variational properties of this problem, leading to the following insights: 1) Particle-wise descent methods implemented in standard libraries can produce non-deterministic embeddings, 2) A probabilistic formulation leads to solutions with interpretable necessary conditions, and 3) The globally optimal solutions to the relaxed, probabilistic problem is only minimized by deterministic embeddings. This progression of results mirrors the classical development of optimal transportation, and in a case relating to the Gromov-Wasserstein distance actually gives explicit insight into the structure of the optimal embeddings, which are parametrically determined and discontinuous on smooth surfaces. Our results also imply that a standard computational implementation for this problem learns sub-optimal mappings, and we discuss how the embeddings learned in that context have highly misleading clustering structure, underscoring the delicate nature of solving this problem computationally.

Keywords: dimension reduction, optimal transport, calculus of variations, gromov-wasserstein, multi dimensional scaling

1. Introduction

A central task in data science is to find efficient representations of high-dimensional data. One form of this task is known as *dimension reduction*, in which one seeks to construct a mapping from a high-dimensional space to a low-dimensional space which approximately preserves features of an input distribution. Dimension reduction serves many purposes: it aids in data visualization and exploration, feature construction, and denoising. Dimension reduction is often stated in terms of some optimization problem, and naturally the proper-

ties and computational tractability are dependent upon the particular dimension reduction objective.

In this work, we consider dimension reduction problems corresponding to optimization problems of the form

$$\min_T \sum_{ij} c(X_i, X_j, T(X_i), T(X_j)),$$

where we are considering the $X_i \in \mathbb{R}^d$ to be data points in a high-dimensional feature space, and $T : \mathbb{R}^d \to \mathbb{R}^m$ represents a mapping, or embedding, into a lower dimensional space. A simple mnemonic here is that 'd' is for "data" and 'm' is for "embedding". In order to accommodate both finite data sets and large sample or population limits, we consider a generalized objective of the form

$$\mathcal{J}(T) := \iint c(x, x', T(x), T(x'))\mu(dx)\mu(dx'), \tag{1.1}$$

where we will assume that $\mu \in \mathscr{P}(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$. Throughout this work we make very few assumptions upon μ: it could be supported on a discrete point cloud, a low dimensional manifold, or a continuous probability distribution. We call this problem the *second-order dimension reduction problem*, where by second-order we mean that the objective function considers pairwise, or second-order, interactions between points. This problem encompasses many common dimension reduction problems, see Section 4 for examples. Variants of this general problem have also been considered under the heading of multi-dimensional scaling and quadratic assignment problems. While not all dimension reduction algorithms can be written in this second-order form, such algorithms generally serve as building blocks for many commonly used methods, see Section 1.1 for more discussion.

Perhaps the simplest version of this form of problem is *Classical Multidimensional Scaling* (cMDS), which, in the discrete setting and with $\sum_i X_i = \sum_i T(X_i) = 0$, seeks to minimize the objective function

$$\min_{\{Y_\ell\}_{\ell=1}^n} \sum_{ij} (\langle X_i, X_j \rangle - \langle Y_i - \mathbb{E}[Y], Y_j - \mathbb{E}[Y] \rangle)^2. \tag{1.2}$$

Alternatively, this can be written, again assuming that $\mathbb{E}(X) = \mathbb{E}[T(X)] = 0$,

$$\min_{T:\mathbb{R}^d \to \mathbb{R}^m} \iint (\langle x, x' \rangle - \langle T(x), T(x') \rangle)^2 \mu(dx)\mu(dx'). \tag{1.3}$$

In both versions of this problem the minimizer is known to be a linear mapping, implying that the minimizer is parametrically determined and smooth. Furthermore, this minimizer can be described as the projection onto the m-dimensions of greatest variance of μ, and is equivalent to PCA. This approach to dimension reduction is prevalent in many contexts.

However, in some settings linear embeddings are too restrictive to capture important structures in data. For this reason a host of different cost functions have been proposed for dimension reduction, each emphasizing distinct priorities. In many contexts these algorithms are able to flexibly capture important features of high-dimensional distributions inaccessible to linear embeddings, but this flexibility comes at a price: non-linear dimension

reduction problems generally can only be resolved via optimization routines, and their solutions do not admit transparent parametric representation formulas. As such, in many cases theoretical properties of the solutions to these problems are poorly understood. In particular, in the setting where μ is a continuum distribution, i.e. the large data or population limit, and when c is non-convex, it is not clear whether the problem (1.1) even admits a minimizer. We will discuss negative results in the mathematical literature along these lines in Section 2, but in simplified terms for non-convex energies it is possible for approximate minimizers to converge towards a limit which is not a function. While the issue of existence is often straightforward in the finite data setting, the lack of a meaningful population limit raises significant issues for optimization and interpretability of minimizers: we highlight this issue with a simple numerical experiment in Example 1.

Similar issues were long-standing in the theory of optimal transportation, and our approach in this paper mirrors that literature. In that context, the Monge formulation of optimal transportation seeks to minimize

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^d, T_\sharp\mu=\nu} \int c(x, T(x))\mu(dx). \tag{1.4}$$

Here ν is an output distribution and $T_\sharp$ denotes the pushforward measure. Demonstrating that Monge's problem has a solution was a major open problem for many years, and while the dimension reduction problem notably lacks the output distribution constraint, the overall lack of convexity with respect to T still engenders a similar type of issue.

The technical solution to this issue in optimal transportation is to instead consider a relaxed version of the problem, namely

$$\min_{\pi\in\Pi(\mu,\nu)} \int c(x, y)\pi(dx\, dy), \tag{1.5}$$

where $\Pi(\mu, \nu)$ is the set of probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ, ν: such probability measures in $\Pi(\mu, \nu)$ are called transportation plans and are multi-valued generalizations of the transportation map T sought for in the Monge problem. In short, this formulation relaxes the requirement that x is mapped "deterministically" to a single $T(x)$, and instead permits a single x to be mapped probabilistically to multiple outputs. Demonstrating that this problem has a solution using "soft" analytical methods is straightforward. Subsequently, one can establish structural properties of such relaxed solutions. Using tools such as cyclical monotonicity and convex analysis, one can demonstrate that under mild assumptions minimizers of (1.5) are actually induced by a mapping, which means that the original Monge problem possesses a solution. We can similarly pose a relaxed version of the MDS problem by seeking to minimize

$$\mathcal{J}(\pi) := \left\{ \iint c(x, x', y, y')\pi(dx\, dy)\pi(dx'dy'),\ \pi \in \Pi(\mu) \right\} \tag{1.6}$$

where we let $\Pi(\mu)$ denote the set of distributions on $\mathbb{R}^d \times \mathbb{R}^m$ which have marginal μ in the first d coordinates and refer to this as the set of *embedding plans*. Throughout this article, we often use the notation $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^m$ to avoid confusion about which space we are embedding to. In addition, we write $\mathrm{proj}_\mathcal{Y}$ to denote the canonical projection $\mathrm{proj}_\mathcal{Y} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ defined by $\mathrm{proj}_\mathcal{Y}(x, y) = y$.

In order to make the problem more concrete, we focus on two basic examples which encompass a broad family of practical situations. The first models the interactions of the embedded variables by an inner product, the second by a squared norm: these costs are known in the literature for multidimensional scaling as *similarity* and *dissimilarity* costs respectively. To clarify this upfront to the reader, we state our first assumptions on the structure of the cost. The remaining assumptions will be introduced in Section 2.2.

$$\begin{aligned} &(\mathbf{IP}) \quad c(x,x',y,y') = \tilde{c}\big(\langle x,x'\rangle, \langle y,y'\rangle\big) \\ &(\mathbf{N}^2) \quad c(x,x',y,y') = \tilde{c}\big(|x-x'|^2, |y-y'|^2\big) \end{aligned}$$

The optimal transportation problem is inherently one of linear programming, whereas the dimension reduction problem is more aptly seen as a non-convex quadratic program (see Example 3). We mention that there is a quadratic programming variant of optimal transportation. In particular the Gromov-Wasserstein metric between distributions μ, ν, supported respectively on $\mathbb{R}^d$ and $\mathbb{R}^m$, is defined by the minimization problem (see Sturm, 2012; Mémoli, 2011)

$$d_{GW_{p,q}}(\mu,\nu)^p = \min_{\pi\in\Pi(\mu,\nu)} \iint \Big||x-x'|^q - |y-y'|^q\Big|^p \pi(dx\,dy)\pi(dx'dy').^1 \tag{1.7}$$

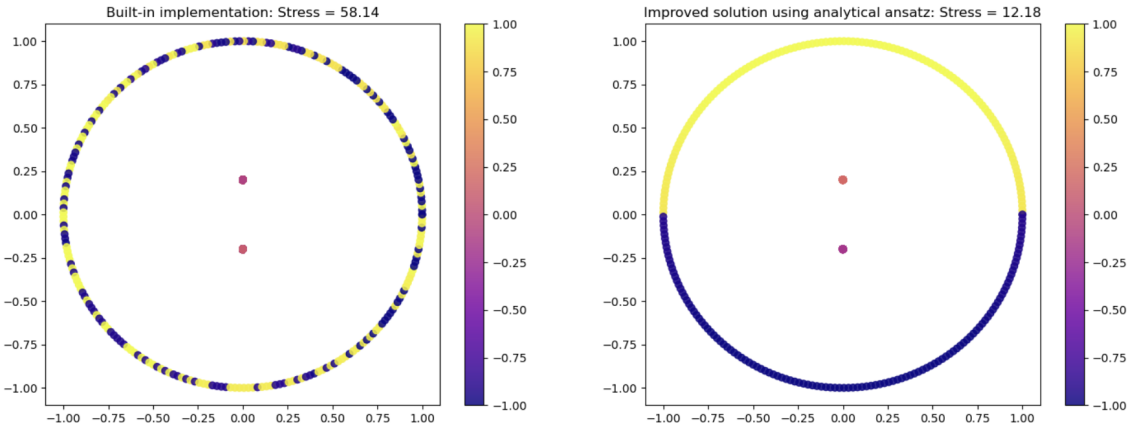
In the Gromov-Wasserstein problem one generally has two marginal constraints, whereas in the dimension reduction problem there is only a single marginal constraint. As such, we can cast the dimension reduction problem as a projection problem in the Gromov-Wasserstein space: namely if we let $c(x,x',y,y') = \Big||x-x'|^q - |y-y'|^q\Big|^p$ then we have that $\min_{\pi\in\Pi(\mu)} \mathcal{J}(\pi) = \min_\nu d_{GW_{p,q}}(\mu,\nu)^p$.

The question of whether minimizers of the Gromov-Wasserstein problem are always induced by transportation maps has recently been studied in (Dumont et al., 2024; Vayer, 2020). In particular, their work shows that in the specific case when $p = q = 2$, there exists an optimal plan which is a "2-map", i.e. supported on the graph of two functions. Our work expands on this in the case of the GW projection problem by showing that the optimal plan is *necessarily* deterministic for a wide range of costs.

A natural question in the context of dimension reduction is whether optimal plans are necessarily maps, or in other words whether solutions to the relaxed problem (1.6) are always solutions of the original problem (1.1). The following example demonstrates that for numerically constructed local minimizers, this is not always the case.

Example 1 [2] *We consider the problem of embedding a particular point cloud in $\mathbb{R}^2$ into $\mathbb{R}$. The point cloud that we choose has 1,000 points placed at $(0, \pm.2)$, as well as 250 points placed randomly upon the unit circle. When we utilize the built-in algorithm for metric multidimensional scaling in Scikit-learn (Pedregosa et al., 2011), the embedding which is found is very discontinuous: this is illustrated in Figure 1a. Indeed, changes around the boundary*

1. The Gromov-Wasserstein metric generally is defined between two metric measure spaces, but we restrict our attention here to distributions on two different Euclidean spaces due to the connection with dimension reduction.
2. The computation in this example was discovered in collaboration with Brian Swenson, and work about computational aspects of this problem is ongoing.



(a) Scikit-Learn Embedding

(b) Globally optimal embedding?

Figure 1: An example in dimension reduction where standard algorithms find solutions that are plans. Here the position of the points represents the original features in $\mathcal{X} = \mathbb{R}^2$, whereas the color represents the learned embedding in $\mathcal{Y} = \mathbb{R}$. For the setup described in Example 1, the first graph shows the embedding learned by the implementation of metric MDS in Scikit-learn, and the second graph shows the embedding Scikit-learn finds if given an analytically-motivated initial guess. The stress values, normalized by the the number of points squared, is also displayed, with a clear improvement in the second image. Code generating this figure is available on the first author's github.

of the unit circle do not have a discernible pattern, and appears to be non-deterministic. The reason for this behavior is that due to the larger clusters near the origin, points on the unit circle are energetically favorable at either ± 1, in the sense that both are local minimizers when other points are held fixed. These local minimizers are both nearly global minimizers as well, as the relative costs of being at either plus or minus one are comparable: this is due to the fact that the two larger clusters are relatively close together. The behavior of the solutions found indeed suggests that non-deterministic embedding plans can be local minimizers of the energy, at least if perturbations are only considered in the sense of small changes to particle positions.

However, working by hand we would expect that the optimal embedding should be much more principled, and should map halves of the circle deterministically to different sides of the real line, according to the cluster they are closer to. Figure 1b, uses this ansatz to construct an initial guess for the same optimization routine in Scikit-learn. The learned embedding, while still having a jump discontinuity, is more interpretable and also obtains a significantly lower cost.

The previous example is, in the authors' opinion, rather arresting from the practical point of view. The embedding constructed by standard libraries has found four well-separated clusters, but two of the clusters were constructed by breaking up the unit sphere

in a completely arbitrary fashion; considering those two clusters as useful features or groups is clearly misleading at best. In fact, later in Example 18 we substantiate this observation theoretically by demonstrating that global minimizers of the quartic MDS problem are necessarily discontinuous. Furthermore, as illustrated in Example 29, non-convex variational problems often exhibit discontinuous solutions—for this reason we expect the discontinuity of global minimizers to persist for many costs of type $\mathbf{N}^2$ which, as we show later, are non-convex.

The stark difference between figures 1a and 1b reveals that the initialization can lead to vastly different locally minimizing embeddings, some of which are potentially problematic. Similar behavior, in terms of performance with respect to different classes of initial conditions, had previously been observed in (Kobak and Linderman, 2021) in the context of tSNE and UMAP.

This computational example also highlights potential mathematical challenges which may arise in establishing the equivalence between the original and the relaxed dimension reduction problems, that is between problems (1.1) and (1.6). Indeed, the embedding learned by the standard implementation ought to be a local minimizer in some sense, suggesting that it may be possible to find local minimizers of (1.6) which are not mappings.

This work aims to address these questions, in certain contexts, through the following contributions:

1. (Proposition 4) We show that the dimension reduction energy (1.1) is not weakly lower semi-continuous in any L^p space for many natural choices of c, meaning that existence of minimizers cannot be established using the direct method of the calculus of variations. In practice, this can lead to highly oscillatory (i.e. non-deterministic) solutions and poor local minima during gradient descent, as demonstrated in Example 1.

2. (Theorems 7 & 8) Under appropriate conditions, we first show that for costs of the form $c(\langle x, x'\rangle, \langle y, y'\rangle)$ and $c(|x - x'|^2, |y - y'|^2)$, the relaxed problem (1.6) has a minimizer. This is mostly a consequence of standard arguments from the calculus of variations.

3. (Theorem 10) For the same class of costs, we demonstrate that any minimizer, π, of (1.6) is essentially supported on the set

$$\{y : J_\pi(y|x) \text{ is minimized in } y\}, \qquad J_\pi(y|x) := \int c(x, x', y, y')\pi(dx'dy'). \tag{1.8}$$

   We call the problem of minimizing $J_\pi(y|x)$ the *Marginal Problem*. This problem in many cases provides a significant constraint upon the form of π. The argument here relies upon a construction of localized perturbations inspired by needle perturbations from control theory.

4. (Corollary 16) We show that for costs in which $\langle y, y'\rangle \mapsto c(\langle x, x'\rangle, \langle y, y'\rangle)$ is convex, there will be deterministic minimizers of (1.6): in the jargon of optimal transportation such solutions are maps. These solutions will furthermore have smoothness controlled by the differentiability of c.

5. (Theorem 20) We show that for costs for which $|y-y'|^2 \mapsto c(0,|y-y'|^2)$ has a unique minimum at $y=y'$ (following from ($\mathbf{A1_{N^2}}$)) and also satisfy a differential condition on c (see ($\mathbf{A2_{N^2}}$) in Section 2.2) that minimizers of (1.6) will necessarily be deterministic. This is a novel result in the theory of the Calculus of Variations since existence of global minimizers for this problem was previously unknown. We refer the reader to the related works section to see the previous state of the art results in this context.

6. (Examples 17, 18, & 19) We discuss in depth the example of a quartic cost in $|y-y'|$ stemming from Gromov-Wasserstein spaces, which is known to give non-linear embeddings. In that context we can additionally show that minimizers admit a parametric representation and have discontinuities along specific hyperplanes.

These results have direct consequences for computational dimension reduction and their applications for practitioners, which we further discuss in Section 5.

The remainder of the work is organized as follows: in Section 1.1 we discuss literature from related fields, including various methods for dimension reduction and optimal transportation. In Sections 2, 3, and 4, we prove the main results for generic costs, namely in Section 2 we prove the existence of solutions to the relaxed problem (1.6), in Section 3 we demonstrate that the support of optimal plans is determined by the Marginal Problem and that similarity costs which are convex in the inner product necessitate deterministic minimizers, and in Section 4 we describe how to obtain a similar result for normed squared costs. In Section 3 we investigate some finer properties of the Gromov-Wasserstein projection problem which also serves to motivate the theoretical considerations in Section 4. In Section 5 we discuss ramifications of these results, as well as some further questions.

1.1 Related Work

Dimension reduction, and specifically Multidimensional Scaling (MDS), has a long history: we refer the reader to the books (Cox and Cox, 2001; Borg and Groenen, 2005) for an in-depth classical statistical treatment of MDS. We mention here that MDS has extensions to a variety of settings, such as the setting where the original points belong to a metric space, or even where we only have access to a matrix of similarities or dissimilarities between our x's. In certain applied fields, such as psychology (Kruskal, 1964), MDS has been utilized extensively for group identification, and is cited in (Borg and Groenen, 2005) as an important tool for data exploration. In the case of classical Multidimensional Scaling, which is equivalent to PCA, the explicit representation of solutions has facilitated many theoretical works, see for example (Li et al., 2020) and the references therein. Several computational approaches have also been developed for speeding up the computation of MDS embeddings. Some references on the topic include (de Leeuw and Mair, 2009; Yang et al., 2015).

On the other hand, in the last twenty years there has been extensive development of new dimension reduction techniques within the context of data science. A standard introductory reference for many of those types of algorithms is (James et al., 2013) Chapter 14, and an in-depth comparison of various non-linear dimension reduction techniques can be found in (Van Der Maaten et al., 2009). These algorithms take a variety of approaches for preserving either global or local structure. Some notable examples include local linear embeddings, isomap, spectral embeddings, Sammon mapping, Multidimensional Scaling, and stochastic

neighborhood embeddings (Roweis and Saul, 2000; Tenenbaum et al., 2000; Belkin and Niyogi, 2003; Sammon, 1969; Kruskal, 1964; Hinton and Roweis, 2002).

While the examples in this paper are fairly general, there are dimension reduction methods which go beyond our framework as they utilize locally adaptive kernels, for example tSNE, UMAP, or LLE (Van der Maaten and Hinton, 2008; McInnes et al., 2018; Roweis and Saul, 2000). There has been recent interest in the mathematical community for identifying simplified models and techniques for understanding tSNE; see for example (Auffinger and Fletcher, 2023) which uses stochastic processes and random matrix theory techniques. There are also some mathematical works which seek to describe specific aspects of finding "good" solutions to SNE (Linderman et al., 2017), in particular by studying early exaggeration techniques commonly used for training. We also remark that variants of the quartic example that we focus on in this work has previously been identified in the statistical learning literature as a particular scaling limit of tSNE (Hinton and Roweis, 2002).

This work has been significantly influenced by the development of the theory of optimal transportation, a good introduction to which can be found in (Villani, 2008). Recent works in the OT literature, such as multi-marginal transport (Pass, 2015) and transport between spaces of unequal dimension (Nenna and Pass, 2020), have also dealt with scenarios similar to ours, but in situations with linear dependence on π.

There has also been a lot of interest recently in the Gromov-Wasserstein distance (Mémoli, 2011), which provides a transportation-based metric between probability measures on two different metric spaces. Very recently multiple authors (Vayer, 2020; Dumont et al., 2024) have studied the question of whether optimal plans in the Gromov Wasserstein problem are in fact realized by mappings. These works attempt to convert the Gromov-Wasserstein problem into an inhomogeneous linear (in π) problem, which then they tackle by using general optimal transportation theory. In particular, in (Dumont et al., 2024) a 2-map for the $p = q = 2$ GW problem is constructed, but the necessity of this solution is still an open question. Furthermore, earlier works such as (Vayer, 2020) show that in the quartic setting, if a certain correlation matrix is non-degenerate then any optimal plan must be induced by a mapping. However, it is unclear how to directly prove that those correlations are in fact non-degenerate. Similarly, in (Arya et al., 2024), a Monge mapping was constructed in the special case between two spheres. In (Vayer et al., 2019) it was claimed that when $d = m = 1$, that optimal solutions admit simple representations (as a monotonic map); however more recent work (Beinert et al., 2023) refuted this claim and provided a counterexample. There has also been several recent works which treat the case where the base measure is finitely supported, and demonstrate that in specific settings the solutions to the Gromov-Wasserstein problem and the MDS problem must be of Monge type: see (Mémoli and Needham, 2022) and (Clark et al., 2024). Our work directly complements this and can be viewed in some ways as extending their result to general probability measures on the input space and a different class of cost functions. We also note that there has been recent work which relaxes the marginal constraint for the Gromov-Wasserstein problem: see (Vincent-Cuaz et al., 2022) wherein the authors consider a graph-centric version of this problem and (Van Assel et al., 2024) wherein the authors define a general framework for relaxing the Gromov-Wasserstein constraint which generalizes both clustering and dimension reduction.

It is important to note the connection between the Gromov-Wasserstein problem and quadratic assignment problems (QAP). In its original formulation (Koopmans and Beckmann, 1957), the quadratic assignment problem describes a variant of the optimal transport problem, wherein the function we minimize is of second degree in the unknown permutation matrix. A notable example of the QAP is the graph matching problem which matches the edges of two graphs in a meaningful way. This can rightly be viewed as a type of Gromov-Wasserstein problem.

We also mention that there has also been a lot of recent work trying to find fast algorithms for GW problems, see for example (Peyré et al., 2016; Vayer et al., 2019; Scetbon et al., 2023). The parametric form we derive for quartic MDS suggests that faster algorithms may also be available for the GW projection problem as well.

Finally, there has been a vein of mathematical literature (Pedregal, 1997; Elbau, 2011; Bellido and Mora-Corral, 2014; Foss et al., 2018) treating the minimization of energies of the form

$$\min_u \mathcal{I}(u), \qquad \mathcal{I}(u) := \iint \Phi(x_1, x_2, u(x_1), u(x_2))\, dx_1\, dx_2.$$

The main focus of these works has been to establish conditions which guarantee the existence of minimizers for energies of this type, by proving weak lower-semicontinuity in an appropriate topology. To our knowledge each of these results requires some form of convexity with respect to Φ. Our work strongly contrasts that line of work, in that 1) we study forms of Φ with specific symmetries, 2) we demonstrate that our energies are *not* weakly lower semicontinuous, and 3) we demonstrate that, in spite of this lack of weak lower semicontinuity, there still exists minimizers of our original dimension reduction problem.

2. Existence of Relaxed Solutions

In this section we consider the problem of existence of minimizers of (1.1) and (1.6). Along the way, we demonstrate that many of the standard techniques from the calculus of variations do not apply to the original problem of finding an embedding map as in (1.1), namely the lack of weak lower semi-continuity. These theoretical observations directly complement the phenomenon observed in Example 1, and demonstrate the difficulty of proving properties of minimizers of the original problem (1.1).

Convexity plays a crucial role in proving existence of minimizers for many variational problems. We begin by demonstrating, through a simple example, why convexity can fail in second-order dimension reduction problems.

Example 2 *We consider, as a running example throughout the paper, the quartic cost* $c(x, x', y, y') = \left(|x - x'|^2 - |y - y'|^2\right)^2$. *Fix* $\varepsilon > 0$ *and let* $T \in C^1(\mathbb{R}^d; \mathbb{R}^m)$ *be a Lipschitz function such that* $\|DT\|_\infty \leq \sqrt{2 - \varepsilon}$. *We consider the effect of interpolating between* $T(x)$ *and* $-T(x)$. *Clearly the midpoint between these two maps is identically zero (we call this map the "zero map" through the paper), namly* $\frac{1}{2}(T(x) - T(x)) \equiv 0$, *and furthermore from the norm structure of the cost we immediately have that* $\mathcal{J}(T) = \mathcal{J}(-T)$. *Hence if* $\mathcal{J}$ *were*

midpoint convex, one would require $\mathcal{J}(T) \geq \mathcal{J}(0)$. However, we have

$$\begin{aligned}\mathcal{J}(T) - \mathcal{J}(0) &= \iint |T(x) - T(x')|^4 - 2|x - x'|^2|T(x) - T(x')|^2 \mu(dx)\mu(dx') \\ &\leq \iint (2-\varepsilon)|x - x'|^2|T(x) - T(x')|^2 - 2|x - x'|^2|T(x) - T(x')|^2 \mu(dx)\mu(dx') \\ &= -\varepsilon \left(\iint |x - x'|^2|T(x) - T(x')|^2 \mu(dx)\mu(dx') \right) \leq 0.\end{aligned}$$

If μ has a direction of non-zero variance, and T is chosen to also vary in that direction, then this inequality is strict: one can find a linear mapping which achieves this goal. Hence $\mathcal{J}$ is not convex with respect to T.

It turns out that the previous observation, which primarily stems from the reflection symmetry of the quartic cost, extends to many second-order costs that have been previously considered for dimension reduction. In particular this symmetry is manifested in both **IP** and $\mathbf{N}^2$ type costs. To this end, we can now restate the non-convexity result above in more generality. Later we will provide suitable assumptions to identify the domain of definition for the dimension reduction problem.

Proposition 3 *If the functionals*

$$\mathcal{J}_{\mathbf{IP}}(T) = \iint c(x, x', \langle T(x), T(x') \rangle) \mu(dx)\mu(dx'),$$

$$\mathcal{J}_{\mathbf{N}^2}(T) = \iint c(x, x', |T(x) - T(x')|^2) \mu(dx)\mu(dx')$$

are finite for functions in $L^p(\mathbb{R}^d; \mathbb{R}^m|\mu)$, and $T \equiv 0$ is not the global minimizer, then $\mathcal{J}_{\mathbf{IP}}$ & $\mathcal{J}_{\mathbf{N}^2}$ are neither convex nor concave on $L^p(\mathbb{R}^d; \mathbb{R}^m|\mu)$.

Proof The proof follows exactly as in Example 2: If $\mathcal{J}(T) = \mathcal{J}(-T) < \mathcal{J}(0)$ for some T then $\mathcal{J}$ cannot be midpoint convex. Furthermore, $\mathcal{J}$ cannot be concave if it is non-constant and positive.

■

As mentioned above, this lack of functional convexity will become a significant theoretical obstacle: this type of obstacle is well-known in the literature for the theory of the Calculus of Variations. For readers that are not familiar with these types of issues, we provide in Appendix A a brief overview of this theory with some examples. A central consideration in this theory is whether or not a functional possesses enough lower semi-continuity to guarantee the existence of minimizers. In the next subsection we show that the dimension reduction problem does not possess this type of lower semi-continuity and hence the standard direct method from the calculus of variations does not apply.

2.1 Failure of lower semi-continuity for the dimension reduction problem

In this section we give our first result for the dimension reduction problem: namely that the energy fails to be weakly lower semi-continuous. This has several immediate implications:

that the existence of minimizers cannot be guaranteed by the direct method, and that a minimizing sequence (for example found numerically) of functions may not converge to a function which minimizes the energy. To begin, we notice that the dimension reduction problem can be restated as

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^m} \int J_T(x,T(x))\mu(dx), \quad \text{with} \quad J_T(x,y) := \int c(x,x',y,T(x'))\mu(dx').$$

As stated in the introduction, the cost function $c(x,x',y,y')$ is often not convex in practice, and in many cases we will not generally have that $y \mapsto J_T(x,y)$ is convex. Thus, by Proposition 31, we suspect that the dimension reduction problem (1.1) will not be weakly lower semi-continuous. The following result demonstrates that this indeed is the case.

Proposition 4 *Consider the dimension reduction problem* (1.1) *in the case where* $c(x,x',y,y') = \tilde{c}(x,x',|y-y'|^2)$ *for some* C^1 *function* $\tilde{c}$ *which is symmetric in* x,x'. *Assume that* μ *has a continuous density on an open and bounded set, and suppose that for all* $x \neq x'$ *we have that* $\frac{d}{dt}\tilde{c}(x,x',t)|_{t=0} < 0$. *Then the dimension reduction problem is not weakly lower semi-continuous.*

Proof Let us choose

$$T_n(x) = v\bigg(\prod_{i=1}^d \operatorname{sign}(\sin(n\pi x_i))\bigg)$$

for some $v \in \mathbb{R}^m$ which will later be specified. First note that clearly $T_n \rightharpoonup 0$. Furthermore, by denoting the sets

$$E_n = \left\{x : \prod_{i=1}^d \operatorname{sign}(\sin(n\pi x_i)) = 1\right\}, \qquad O_n = \left\{x : \prod_{i=1}^d \operatorname{sign}(\sin(n\pi x_i)) = -1\right\},$$

the cost of T_n will be computed as

$$\begin{aligned}
\mathcal{J}(T_n) &= \iint_{E_n\times E_n} c(x,x',0)\mu(dx)\mu(dx') \\
&\quad + \iint_{O_n\times O_n} \tilde{c}(x,x',0)\mu(dx)\mu(dx') \\
&\quad + 2\iint_{E_n\times O_n} \tilde{c}(x,x',2|v|)\mu(dx)\mu(dx') \\
&= \mathcal{J}(0) + 2\iint_{E_n\times O_n} [\tilde{c}(x,x',2|v|) - \tilde{c}(x,x',0)]\mu(dx)\mu(dx').
\end{aligned}$$

Notice that we have, by the Riemann-Lebesgue Lemma,

$$2\iint_{E_n\times O_n} [\tilde{c}(x,x',2|v|) - \tilde{c}(x,x',0)]\mu(dx)\mu(dx') \xrightarrow{n\to\infty} \frac{1}{2}\iint [\tilde{c}(x,x',2|v|) - \tilde{c}(x,x',0)]\mu(dx)\mu(dx').$$

where we have used the fact that

$$\mathbb{1}_{E_n}(x)\mathbb{1}_{O_n}(x') = \frac{(1+\prod_{i=1}^d \operatorname{sign}(\sin(n\pi x_i)))(1+\prod_{i=1}^d \operatorname{sign}(\sin(n\pi x_i')))}{4}$$

along with the continuity of c and the density μ. Thus, given $\varepsilon > 0$ for sufficiently large n, we have that

$$\begin{aligned}\mathcal{J}(T_n) - \mathcal{J}(0) &< \frac{1}{2}\iint [\tilde{c}(x, x', 2|v|) - \tilde{c}(x, x', 0)]\mu(dx)\mu(dx') + \varepsilon \\ &\leq \frac{1}{2}\iint -\phi(x, x')|v| + o(|v|)\mu(dx)\mu(dx') + \varepsilon,\end{aligned}$$

where $\phi(x, x') \geq 0$ with equality only possibly when $x = x'$ by our assumption upon the derivative of $\tilde{c}$. Making v sufficiently small so that we can neglect the $o(|v|)$ term, and taking $\varepsilon \to 0$ then implies that $\liminf_n \mathcal{J}(T_n) < \mathcal{J}(0)$, proving the result. ■

The previous proposition demonstrates that the dimension reduction energy $\mathcal{J}$ is not weakly lower semi-continuous: this implies that information about minimization is lost in limit obtained with that topology. The standard approach to handling this situation is to instead permit limits that are multi-valued: meaning that one x is mapped probabilistically to multiple y values. For example, in the proof of the previous proposition we may write

$$\pi_n(dx\, dy) = \mu(dx)(\mathbb{1}_{E_n}(x)\delta_v(dy) + \mathbb{1}_{O_n}(x)\delta_{-v}(dy)),$$

and then compute

$$\mathcal{J}(T_n) = \iint c(x, x', y, y')\pi_n(dx\, dy)\pi_n(dx'dy').$$

Using the computation with the Riemann-Lebesgue lemma in the proof of the previous proposition, it is straightforward to show that π_n converges (in the sense of weak convergence of measures) to $\pi(dx\, dy) = \mu(dx)(1/2\delta_v(dy) + 1/2\delta_{-v}(dy))$. Hence we have that

$$\mathcal{J}(T_n) \to \iint c(x, x', y, y')\pi(dx\, dy)\pi(dx'dy').$$

Slightly abusing notation, we can then define a *relaxed energy*

$$\mathcal{J}(\pi) := \iint c(x, x', y, y')\pi(dx\, dy)\pi(dx'dy').$$

Here π represents a probabilistic coupling between x's and y's which generalizes a deterministic coupling (or function) mapping each x to a single y. In the context of optimal transportation, the coupling π is sometimes called a *transportation plan*, whereas a deterministic coupling in that context is called a *transportation map*. In the continuum mechanics literature such a probabilistic relaxation is called a *Young measure*. In many contexts the existence of minimizers of the relaxed energy is more straightforward to prove using compactness and continuity arguments: we carry out these standard arguments in the next section.

2.2 Existence of relaxed solutions

In light of the discussion in the previous section, we turn our attention to the problem of existence of minimizers to the relaxed problem (1.6). We begin by giving some definitions. Given $\mu \in \mathscr{P}(\mathbb{R}^d)$ and family $\{\nu(\cdot|x)\}_{x\in\mathbb{R}^d} \subset \mathscr{P}(\mathbb{R}^m)$ for which $x \mapsto \nu(Q|x)$ is a measurable function for all $Q \in \mathcal{B}(\mathbb{R}^m)$, there exists a unique (in measure) probability distribution $\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$ such that for all $P \in \mathcal{B}(\mathbb{R}^d)$ and $Q \in \mathcal{B}(\mathbb{R}^m)$,

$$\pi(P \times Q) = \int_P \nu(Q|x)\mu(dx). \tag{2.1}$$

Let the space of all joint probability measures which can be written in the form above be called $\Pi(\mu)$, more precisely $\Pi(\mu) := \{\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m) \,|\, \mathrm{proj}_{\mathbb{R}^d}\sharp\pi = \mu\}$ which are all the probability measures on $\mathbb{R}^d \times \mathbb{R}^m$ with $\mathcal{X}$-marginal μ. In analogy to optimal transportation, we call $\Pi(\mu)$ the set of *embedding plans* for μ.

As soon as c is itself lower semi-continuous, the function $\pi \mapsto \iint c\, d\pi d\pi$ is automatically lower semi-continuous with respect to weak convergence of probability measures, by Portmanteau's theorem. We recall that a sequence of probability measures $\pi_n \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$ is said to converge weakly to π if, for every bounded, continuous function f, we have that $\int f d\pi_n \to \int f d\pi$. In order to recover sequential compactness for sequences of probability measures $\pi_n \in \Pi(\mu)$, we must introduce the notion of *tightness of measure* and its application on the subspace $\Pi(\mu)$.

Definition 5 (Tightness of Embedding Plans) *A sequence of probability distributions $\{\pi_n\}_{n=1}^\infty \subset \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$ is said to be tight if for every $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathbb{R}^d \times \mathbb{R}^m$ for which $\sup_n \pi_n(K_\varepsilon^c) < \varepsilon$.*

In the case that $\pi_n \in \Pi(\mu)$, we can find a compact set K_d in $\mathbb{R}^d$ so that $\mu(K_d) > 1 - \frac{\varepsilon}{2}$. In turn if we can find a compact set K_m so that $\pi_n(\mathbb{R}^d \times K_m) > 1 - \frac{\varepsilon}{2}$ we can use $K = K_d \times K_m$ and obtain the estimate $\pi_n(K^c) < \varepsilon$: this implies that when $\pi_n \in \Pi(\mu)$ we only need to verify tightness in the marginal over the last m coordinates. In symbols, we write this as

$$\{\pi_n\}_{n=1}^\infty \text{ is tight in } \Pi(\mu) \iff \nu_n := \mathrm{proj}_{\mathcal{Y}}\sharp\pi_n,\ \{\nu_n\}_{n=1}^\infty \text{ is tight in } \mathscr{P}(\mathbb{R}^m).$$

Here we, in a slight abuse of notation, are letting $\nu(Q) = \int_{\mathbb{R}^d \times Q} d\pi(x, y)$: meaning that if we suppress the x-dependence in $\nu(dy|x)$ then we are indicating the marginal distribution in y.

By Prokhorov's theorem, tightness of a sequence of probability measures implies weak compactness. Thus the problem of existence of minimizers to the relaxed problem reduces to establishing tightness of sequences of embedding plans with bounded energy $\mathcal{J}$.

Assumptions

We are now ready to list our assumptions. As stated before, we will consider the following two types of costs:

$$\begin{aligned} (\mathbf{IP}) \quad & c(x, x', y, y') = \tilde{c}_{\mathbf{IP}}\big(\langle x, x'\rangle, \langle y, y'\rangle\big) \\ (\mathbf{N}^2) \quad & c(x, x', y, y') = \tilde{c}_{\mathbf{N}^2}\big(|x - x'|^2, |y - y'|^2\big) \end{aligned}$$

where we make the following assumptions on the function $\tilde{c}_{\mathbf{N}^2}, \tilde{c}_{\mathbf{IP}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$; when writing $\tilde{c}$ without a subscript, the assumption applies to both cases.

- (**A1**) *For every compact set $K \subset \mathbb{R}^d$, there is an unbounded increasing function $f_K : \mathbb{R}^+ \to \mathbb{R}$ such that $x, x' \in K \implies \tilde{c}(\cdot, t) \geq f_K(t) \geq 0$*
- (**A2**) *c is a locally C^2 function in all its variables with derivative values satisfying $|D^2 c| \leq C(1 + c)$*
- ($\mathbf{A1_{IP}}$) *$t \mapsto \tilde{c}_{\mathbf{IP}}(\cdot, t)$ is strictly convex*
- ($\mathbf{A1_{N^2}}$) *$\partial_t \tilde{c}_{\mathbf{N}^2}(0, t) \geq 0$ with equality uniquely attained at $t = 0$.*
- ($\mathbf{A2_{N^2}}$) *$\partial_t \tilde{c}_{\mathbf{N}^2}(0, t) + 2t\, \partial_{tt} \tilde{c}_{\mathbf{N}^2}(0, t) \geq 0$ with equality uniquely attained at $t = 0$.*

Assumption (**A1**) ensures that c is nonnegative as well as provides coercivity. The growth condition (**A2**) will allow us to integrate derivatives in a meaningful way. This assumption on growth conditions of derivatives of c naturally holds for polynomial costs. Assumption ($\mathbf{A1_{IP}}$) will be a sufficient condition under which solutions to the relaxed problem are necessarily deterministic. Similarly, assumptions ($\mathbf{A1_{N^2}}$) and ($\mathbf{A2_{N^2}}$) are needed to quantify the eigenvalues of the mixed Hessian matrix $D_{yy'}c$ which appears in second variation arguments used to obtain maps in the case of $\mathbf{N}^2$ costs.

As a final note, we mention that unless otherwise specified we will drop the tilde on the cost in the above assumptions. For example, we will write $c(|x - x'|^2, |y - y'|^2)$ rather than $\tilde{c}(|x - x'|^2, |y - y'|^2)$. Provided below is a table of several cost functions which can fit into our framework.

Method	$c\big(x, x', y, y'\big)$
PCA	$\big(\langle x, x'\rangle - \langle y, y'\rangle\big)^2$
Kernel PCA (Schölkopf et al., 1997)	$\big(\kappa(x, x') - \langle y, y'\rangle\big)^2$
q-MDS (Shepard, 1962)	$\big(\lvert x - x'\rvert^2 - \lvert y - y'\rvert^2\big)^2$
q-Sammon (Sammon, 1969)	$\frac{\big(\lvert x-x'\rvert^2-\lvert y-y'\rvert^2\big)^2}{\lvert x-x'\rvert^2}$

Table 1: A list of several costs which fit into our framework and satisfy Assumptions. The "q" refers to quartic variants of standard costs used in dimension reduction.

While the assumptions above are widely applicable, they are not universal. The following example illustrates a natural and conceptually appealing choice that falls outside our framework:

Remark 6 *A notable example outside our framework is the squared log-ratio cost*

$$c(s, t) = \left[\log \frac{1 + t}{1 + s}\right]^2,$$

used in the Gromov–Monge embedding objective of (Lee et al., 2025). This choice is scale-aware: the concavity of the logarithm compresses large distances while magnifying the effect of discrepancies at small scales. Such weighting is common in manifold learning, where in high dimensions most pairs of points are far apart, and it is often more important to preserve the relative geometry of nearby points than to match all distances equally well.
In our ($\mathbf{N}^2$) *formulation, this cost satisfies* ($\mathbf{A1_{N^2}}$) *but fails* ($\mathbf{A2_{N^2}}$). *Indeed,*

$$\partial_t c(0,t) = \frac{2\log(1+t)}{1+t},$$

which is 0 *at* $t=0$ *and positive for* $t>0$, *verifying* ($\mathbf{A1_{N^2}}$). *However,*

$$\partial_t c(0,t) + 2t\,\partial_{tt} c(0,t) = \frac{2\big[(1-t)\log(1+t)+2t\big]}{(1+t)^2}$$

is positive for small $t>0$ *but becomes negative for large* t, *so* ($\mathbf{A2_{N^2}}$) *fails. The failure of one or both of* ($\mathbf{A1_{N^2}}$) *and* ($\mathbf{A2_{N^2}}$) *leads to a lack of control on the mixed Hessian* $D^2_{yy'}c(x,x,y,y')$ *which is crucial in Proposition 28.*
The authors use this cost for its favorable optimization properties, namely the Hessian in this formulation is better behaved, yielding significantly faster convergence in their experiments. However, to address the question of existence they impose geometric control directly: they require the embedding $T:\mathcal{X}\to\mathcal{Y}$ *to be* α*-bi-Lipschitz for some* $0<\alpha\le 1$, *meaning*

$$\alpha\,|x-x'| \le |T(x)-T(x')| \le \alpha^{-1}\,|x-x'| \quad \text{for all } x,x'\in\mathcal{X}.$$

This global two-sided bound on distortion rules out the small-scale degeneracies that the ($\mathbf{A2_{N^2}}$) *failure would otherwise permit and in their framework it enables existence to be established via standard compactness–coercivity arguments.*

We now explicitly derive an upper bound which quantifies tightness under the assumption (**A1**). We begin with the inner product case.

Theorem 7 (Inner Product Costs) *Assume* (**IP**) *and* (**A1**) *and that* c *is lower semi-continuous. Let* $\mu\in\mathscr{P}(\mathbb{R}^d)$ *and suppose that* $\inf_{\Pi(\mu)}\mathcal{J}<+\infty$, *where* $\mathcal{J}$ *is given by* (1.6). *Then there exists* $\pi\in\Pi(\mu)$ *such that* $\mathcal{J}(\pi)=\inf_{\Pi(\mu)}\mathcal{J}$.

Proof We consider a sequence π_n so that $\mathcal{J}(\pi_n)\to\inf_{\Pi(\mu)}\mathcal{J}$ and

$$\mathcal{J}(\pi_n) \le 2\inf_{\Pi(\mu)}\mathcal{J}.$$

Notice that if $\langle y,y'\rangle=0$ for $\nu\otimes\nu$-a.e. (y,y'), it must be that the support of ν is concentrated on the singleton $\{0\}$, which would trivially give tightness of π_n; thus without loss of generality we may assume that elements of the minimizing sequence have nontrivial support in y.

We now claim that the sequence π_n must be tight: the argument will essentially show that mass far from the origin must be small in order for the previously displayed inequality to hold. As described in the definition of tightness, it suffices to show that ν_n is tight. To begin, we let $\varepsilon>0$ and partition $\mathbb{R}^m$ into a finite number of disjoint cones $C_1,...,C_\ell$

wherein the angle between any two points is at most $\pi/6$ and denote $C_{i,r} = C_i \cap B_r^c(0)$ for $i = 1, ..., \ell$. Let $K_\varepsilon \subset \mathbb{R}^d$ be a compact set such that $\mu(K_\varepsilon^c) < \frac{\varepsilon}{2}$. By the non-negativity of c which follows from (**A1**), we have

$$\mathcal{J}(\pi_n) \geq \sum_{i=1}^{\ell} \iint_{(K_\varepsilon \times C_{i,r})^2} c \, d\pi_n d\pi_n,$$

which, by assumption (**A1**), yields

$$\mathcal{J}(\pi_n) \geq \sum_{i=1}^{\ell} \iint_{(K_\varepsilon \times C_{i,r})^2} f_{K_\varepsilon} \circ |\langle \cdot, \cdot \rangle| \, d\pi_n d\pi_n.$$

Finally, by the construction of our cones, we have that $y, y' \in C_i \implies |\langle y, y' \rangle| \geq |y| \, |y'|/2$, and hence

$$\mathcal{J}(\pi_n) \geq f_{K_\varepsilon}(r^2/2) \sum_{i=1}^{\ell} \big(\pi_n(K_\varepsilon \times C_{i,r})\big)^2 \geq f_{K_\varepsilon}(r^2/2) \frac{\big(\pi_n(K_\varepsilon \times B_r^c(0))\big)^2}{\ell}.$$

The second inequality follows by Jensen's inequality and by virtue of $C_1, ..., C_\ell$ forming a partition. The above considerations hence imply for every element of the minimizing sequence, one has

$$\pi_n(\mathbb{R}^d \times B_r^c(0)) = \nu_n(B_r^c(0)) \leq \sqrt{\frac{2\ell \inf_{\Pi(\mu)} \mathcal{J}}{f_{K_\varepsilon}(r^2/2)}} + \frac{\varepsilon}{2}.$$

By then making r sufficiently large we can make the right hand side smaller than ε, which shows that the ν_n, and subsequently the π_n, are tight. Prokhorov's Theorem gives a subsequence with a weak limit π, and π is a relaxed minimizer by the weak lower semi-continuity of $\mathcal{J}$, as argued above. ■

The same argument, with only slight modifications to the geometry, provides the same result for the norm-based costs.

Theorem 8 (Normed Costs) *Assume* (**N**^{2}) *and* (**A1**) *and that c is lower semi-continuous. Let $\mu \in \mathscr{P}(\mathbb{R}^d)$ and suppose that $\inf_{\Pi(\mu)} \mathcal{J} < +\infty$, where $\mathcal{J}$ is given by* (1.6)*. Then there exists $\pi \in \Pi(\mu)$ such that $\mathcal{J}(\pi) = \inf_{\Pi(\mu)} \mathcal{J}$.*

Proof The main difference in the proof is that one should replace cones, which have aligned inner products, with pairs of halfspaces which are well-separated, and hence have lower bounds on pairwise distances.

Specifically, let $\{\pi_n\}_{n=1}^{\infty}$ satisfy $\mathcal{J}(\pi_n) \to \inf_{\Pi(\mu)} \mathcal{J}$ and $\mathcal{J}(\pi_n) \leq 2 \inf_{\Pi(\mu)} \mathcal{J}$. Since the cost is translation invariant in y, without loss of generality, we may assume that each element in this sequence has the property that for any $k \in 1 \dots m$ we have $\pi_n(\mathbb{R}^d \times H_k^+) = \pi_n(\mathbb{R}^d \times H_k^-) = 1/2$ where $H_k^+ := \{y \in \mathbb{R}^m : y_k > 0\}$ and $H_k^- := \{y \in \mathbb{R}^m : y_k \leq 0\}$. We

also write $H^+_{k,r} = \{y \in \mathbb{R}^m : y_k > r\}$. As before, take $K_\varepsilon \subset \mathbb{R}^d$ to be a compact set for which $\mu(K_\varepsilon^c) < \frac{\varepsilon}{4m}$, and let $\varepsilon < 1/2$. By the non-negativity of c, one has for any $k \in 1 \dots m$

$$\mathcal{J}(\pi_n) \geq \iint_{(K_\varepsilon \times H^+_{k,r}) \times (K_\varepsilon \times H^-_k)} c \, d\pi_n d\pi_n.$$

By again using the bound (**A1**), the monotonicity and unboundedness of f_{K_ε}, and the fact that $(y, y') \in H^+_{k,r} \times H^-_k \implies |y - y'|^2 > r^2$, then gives, for r sufficiently large,

$$\frac{\mathcal{J}(\pi_n)}{f_{K_\varepsilon}(r^2)} \geq \pi_n(K_\varepsilon \times H^+_{k,r})\pi_n(K_\varepsilon \times H^-_k) \geq \left(\nu_n(H^+_{k,r}) - \frac{\varepsilon}{4m}\right)\left(\frac{1}{2} - \frac{\varepsilon}{4m}\right)$$

and in turn, rearranging, summing over k, and using the fact that $\varepsilon < 1/2$, we obtain

$$\nu_n(\cup_{k=1}^m H^+_{k,r}) \leq m \frac{8 \inf_{\Pi(\mu)} \mathcal{J}}{f_{K_\varepsilon}(r^2)} + \frac{\varepsilon}{4}.$$

By repeating the argument for the halfspaces where $y_k < -r$, we then obtain

$$\nu_n(\{|y|_\infty > r) \leq \frac{16m \inf_{\Pi(\mu)} \mathcal{J}}{f_{K_\varepsilon}(r^2)} + \frac{\varepsilon}{2},$$

and by taking r sufficiently large we can then bound $\nu_n(\{|y|_\infty > r) \leq \varepsilon$. This proves tightness of the ν_n, which in turn proves, up to a subsequence, existence of a weak limit π which must be a minimizer. ∎

Note that we have established the existence of relaxed solutions specifically in the case of inner product or norm squared type costs. While this choice simplifies certain aspects of the proof, it is largely a matter of presentation. In fact, it is reasonable to expect that analogous results would hold more generally, without relying on the inner product or norm-squared structure, provided an appropriate coercivity assumption is imposed.

3. The Marginal Problem

As discussed in the introduction, many of the standard tools for existence of transportation maps in optimal transportation fail in the present context due to a lack of convexity in π of the relaxed problem. In particular, the effects of replacing an embedding plan π with $\pi + \gamma$ (such that $\pi + \gamma \in \Pi(\mu)$) are realized as first *and* second-order terms in γ. More precisely, if γ is a signed measure on $\mathcal{X} \times \mathcal{Y}$ such that for all d-dimensional Borel sets A, $\gamma(A \times \mathbb{R}^m) = 0$, one has

$$\mu(A) = \pi(A \times \mathbb{R}^m) = [\pi + \gamma](A \times \mathbb{R}^m),$$

so that adding γ leaves the $\mathcal{X}$-marginal invariant. With this notation along with the symmetry present in either ($\mathbf{N}^2$) or (**IP**) costs, one can succinctly express the change in energy due to the perturbation γ:

$$\mathcal{J}(\pi + \gamma) - \mathcal{J}(\pi) = 2 \underbrace{\iint c \, d\pi d\gamma}_{=:\mathcal{J}(\gamma|\pi)} + \underbrace{\iint c \, d\gamma d\gamma}_{=:\mathcal{J}(\gamma)}, . \tag{3.1}$$

Here, $\gamma \mapsto \mathcal{J}(\gamma|\pi)$ encapsulates the linear contribution while $\gamma \mapsto \mathcal{J}(\gamma)$ represents the quadratic contribution. Further developing this notation, we remark that $\mathcal{J}(\gamma|\pi)$ encodes the fact that the first-order effect should be thought of as a linear programming problem over $(x, y) \mapsto \int c(x, x', y, y')\pi(dx'dy')$ for a fixed embedding plan π. Denoting this map as $J_\pi(y|x)$, we see that the first-order problem can be formally stated: for any fixed $\tilde{\pi} \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$, find π such that

$$\pi \in \arg\min_{\Pi(\mu)} \int J_{\tilde{\pi}}(y|x)\pi(dx\, dy)$$

As we are free to vary the $\mathcal{Y}$-marginal of π, the above formulation strongly suggests that if $\pi(dx\, dy) = \nu(dy|x)\mu(dx)$ is optimal, then the support of $\nu(\cdot|x)$ is concentrated on the minimizers of $J_\pi(\cdot|x)$. This turns out to indeed be the case, but before validating the claim, we give a definition to streamline the proceeding discussion.

Definition 9 *Given a continuous cost c of type* (**IP**) *or* ($\mathbf{N}^2$) *which satisfies assumption* (**A1**) *and a embedding plan $\pi \in \Pi(\mu)$, we define the marginal problem of $\mathcal{J}(\pi)$ by the function*

$$J_\pi(y|x) := \int c(x, x', y, y')\pi(dx'dy'). \tag{3.2}$$

Furthermore, for the set valued map $\lambda : x \mapsto \arg\min J_\pi(\cdot|x)$, we call the set of all pairs $(x, \lambda(x))$ the minimal graph of J_π and denote it with the symbol Λ_π.

Notice that the chosen convention is that calligraphic letters are reserved to functional problems while standard capital letters denote functions on finite dimensional spaces. We also remark that when c is continuous, With this definition in place, we now present the following theorem. It is prudent to remark we have made no assumption on the uniqueness of $\arg\min J_\pi(\cdot|x)$.

Theorem 10 (Marginal Minimization) *Suppose that c is a continuous cost of type* (**IP**) *or* ($\mathbf{N}^2$) *and satisfies assumption* (**A1**). *If $\pi \in \Pi(\mu)$ is a minimizer of* (1.6), *then the support of π is concentrated on the minimal graph of J_π. In other words, π must satisfy the implicit relation*

$$\pi(\Lambda_\pi) = 1. \tag{3.3}$$

From a high level, the theorem tells us that the variational problem (1.6) may be transformed into a finite dimensional one; that of minimizing $J_\pi(\cdot|x)$ for every given x (which implicitly depends on π). This is analogous to the situation in optimal control wherein a value function is found by solving a PDE which implicitly depends on the control u. Once this value function is found, one may pointwise minimize a (finite dimensional) Hamiltonian to find the optimal control.

Continuing the analogy with control, notice that in the absence of a convexity assumption on c, smoothly varying π is likely prone to get 'stuck' in local minima. To this end, the proof of the theorem uses localized perturbations in $\mathcal{X}$ which transport probability mass in $\mathcal{Y}$ across potentially large distances. These perturbations are analogous to needle variations used in the proof of the Pontryagin Maximum Principle.

We now illustrate the proof idea in the discrete case. To this end, suppose $\mu = (1/n)\sum_{i=1}^n \delta_{x_i}$ and $\pi = (1/n)\sum_{i,j=1}^n \pi_{ij}\delta_{(x_i,y_j)}$ where $y_1, y_2, ..., y_n \in \mathbb{R}^m$ constitute the optimal solution to (1.6); each π_{ij} tells what proportion of the $1/n$ mass at point x_i will go to location y_j. Suppose that $y_j \notin \lambda(x_i)$ for some pair (x_i, y_j) with $\pi_{ij} > 0$. Define a perturbation γ which sends y_j to $\tilde{y}_j \in \lambda(x_i)$, that is

$$\gamma = \varepsilon\frac{\pi_{ij}}{n}\Big(\delta_{(x_i,\tilde{y}_j)} - \delta_{(x_i,y_j)}\Big)$$

where $0 < \varepsilon \ll 1$ and let $\tilde{\pi} = \pi + \gamma$. Computing first the effect on the linear term, $\mathcal{J}(\gamma|\pi)$ we have

$$\mathcal{J}(\gamma|\pi) = \frac{\varepsilon}{n}\Big[J_\pi(\tilde{y}_j|x_i) - J_\pi(y_j|x_i)\Big] < 0,$$

by marginal minimality of $\tilde{y}_j$. Further, we have

$$\gamma \otimes \gamma = (\varepsilon^2/n^2)\Big(\delta_{(x_i,\tilde{y}_j)} \otimes \delta_{(x_i,\tilde{y}_j)} - \delta_{(x_i,\tilde{y}_j)} \otimes \delta_{(x_i,y_j)} - \delta_{(x_i,y_j)} \otimes \delta_{(x_i,\tilde{y}_j)} + \delta_{(x_i,y_j)} \otimes \delta_{(x_i,y_j)}\Big),$$

and hence $\mathcal{J}(\gamma) = \varepsilon^2/n^2\big[c(x_i,x_i,y_j,y_j) - 2c(x_i,x_i,y_j,\tilde{y}_j) + c(x_i,x_i,\tilde{y}_j,\tilde{y}_j)\big]$ which is clearly dominated by the linear term when ε is small enough. Thus by Equation (3.1), $\mathcal{J}(\pi+\gamma) < \mathcal{J}(\pi)$ and we obtain a contradiction to the optimality of π. Extending this idea to the continuum case only requires a direct, measure-theoretic argument.

Proof Let π be an optimal solution of (1.6) and suppose for sake of contradiction that $\pi(\Lambda_\pi^c) > 0$. By defining

$$A_{k,r} = \big\{(x,y) : k^{-1} < J_\pi(y|x) - \min J_\pi(\cdot|x)\big\} \cap \{(x,y) : |x|, |y| < r\},$$

it follows that $\Lambda_\pi^c = \bigcup_{k,r=1}^\infty A_{k,r}$ and consequentially, $\pi(A_{k,r}) > 0$ for some $(k,r) \in \mathbb{N}^2$. Define the measure $\pi_{k,r} = \frac{\pi_{|A_{k,r}}}{\pi(A_{k,r})}$ and take $\tilde{\lambda}$ as a measurable selection of λ. This selection exists by the continuity of the marginal problem, J_π, which follows by the continuity of c.[3] Choose $\varepsilon < \min\{2\pi(A_{k,r}), (k\|c\|_{L^\infty(A_{k,r}\times A_{k,r})})^{-1}\}$ to construct the perturbation

$$\gamma = \frac{\varepsilon}{2}\left(\frac{\nu(A_{k,r}|x)}{\pi(A_{k,r})} \cdot \delta_{\tilde{\lambda}(x)} \otimes \mu - \pi_{k,r}\right).$$

where we have used the representation $\pi(A_{k,r}) = \int \nu(A_{k,r}|x)\mu(dx)$. By the first restriction on ε, it follows that $\pi + \gamma$ is a positive measure. Furthermore we can see that this perturbation does not affect the input marginal, that is $\gamma(P \times \mathbb{R}^m) = 0$ for all $P \in \mathcal{B}(\mathbb{R}^d)$.

Tracking the effects of this perturbation, the linear term becomes:

$$\begin{aligned}
\mathcal{J}(\gamma|\pi) &= \frac{\varepsilon}{2\pi(A_{k,r})}\int J_\pi(y|x)\delta_{\tilde{\lambda}(x)}(dy)\nu(A_{k,r}|x)\mu(dx) - \frac{\varepsilon}{2}\int J_\pi(y|x)\pi_{k,r}(dx\,dy) \\
&< \frac{\varepsilon}{2\pi(A_{k,r})}\int \min J_\pi(\cdot|x)\nu(A_{k,r}|x)\mu(dx) - \frac{\varepsilon}{2}\int \big(\min J_\pi(\cdot|x) + k^{-1}\big)\frac{\nu(A_{k,r}|x)\mu(dx)}{\pi(A_{k,r})} \\
&= -\frac{\varepsilon}{2k}
\end{aligned}$$

3. The existence of a minimizing measurable selection of J_π follows from a theorem of Rockafeller (see 14.37 in Rockafellar et al., 2009) as soon as J_π is a Carathéodory function.

where on the second to last line we make use of the lack of dependence on y in the latter integrand. As c is nonnegative, we have the following estimate for the quadratic term:

$$\begin{aligned}\mathcal{J}(\gamma) &\leq \frac{\varepsilon^2}{4}\iint c(x,x',\lambda(x),\lambda(x'))\nu(A_{k,r}|x)\mu(dx)\nu(A_{k,r}|x')\mu(dx') \\ &+ \frac{\varepsilon^2}{4}\iint c(x,x',y,y')\pi_{k,r}(dx\,dy)\pi_{k,r}(dx'dy') \\ &\leq \varepsilon^2\cdot\|c\|_{L^\infty(A_{k,r}\times A_{k,r})}.\end{aligned}$$

Putting the estimates together with (3.1), one has

$$\mathcal{J}(\pi+\gamma)-\mathcal{J}(\pi) < -\frac{\varepsilon}{k}+\varepsilon^2\cdot\|c\|_{L^\infty(A_{k,r}\times A_{k,r})}$$

which is negative by our choice of ε. This is a contradiction to optimality. ■

Note that in the proof we do not, in fact, require the full structural assumptions (**IP**) or ($\mathbf{N^2}$). It suffices to assume merely that $c \geq 0$ and that the symmetry condition $c(x,x',y,y') = c(x',x,y',y)$ holds for the result to go through.

Remark 11 *In the proof presented above we notice that transporting ε-mass to (global) marginal minimizers incurs a gain on the embedding cost regardless of whether or not π is optimal. This is quite different in philosophy from the standard computational approaches which conduct particle-wise gradient descent in $\mathcal{Y}$. As evidenced by Example 1, particle-wise decent potentially gets caught in local minima of the marginal problem. These local minima can lead to highly oscillatory embeddings: in the language of this work this corresponds to probabilistic couplings.*

A different way of casting this observation is that if we are only allowed to perturb a coupling π smoothly in y then there may be local minimizers of $\mathcal{J}$ which are probabilistic in $\mathcal{Y}$. However we shall see in Section 4 that probabilistic couplings are never optimal in our dimension reduction problems. This suggests the need for improved computational algorithms which are capable of executing perturbations which are not smooth in $\mathcal{Y}$.

Remark 12 *The results in this section directly parallel classical results from quadratic programming, which show that the minimizer of a quadratic program under convex constraints will also be the minimizer of the linear program obtained by linearizing the quadratic program about the optimal solution, subject to the same constraints. Indeed, as our variational problem is really an infinite dimensional quadratic program, the previous results can be seen as an infinite dimensional analog of those classical results.*

3.1 Critical point equation

In light of Theorem 10, it is natural to consider the necessary conditions for optimality in y of the marginal problem, and the constraints that they impose upon the optimal solution π. To begin, we consider assumptions under which the marginal problem, which depends implicitly upon the measure π, is differentiable.

Lemma 13 *Let the cost function c be of type* (**IP**) *or* ($\mathbf{N}^2$) *and satisfy assumptions* (**A1**) *and* (**A2**). *Let π be a minimizer of* (1.6). *Then the funtion J_π is C^2 in x, y.*

Proof Formally differentiating we should have the formula

$$D^2 J_\pi(y|x) = \int D^2 c(x, x', y, y')\pi(dx'dy').$$

However, by (**A2**), we can write

$$\iint |D^2 c| d\pi d\pi \leq C(1 + \mathcal{J}(\pi)) < \infty.$$

This in turn implies that $\int D^2 c(x, x', y, y')\pi(dx'dy')$ is integrable (with respect to π), in x, y. A dominated convergence argument, along with continuity of the derivatives, then gives that J_π is C^2 in x, y. ■

We notice, due to assumption (**A1**), the minimizers of the marginal problem at x live on a compact set $K_x \subset \mathcal{Y}$ and thus a necessary condition for optimality is that $\operatorname{Spt} \pi$ must be concentrated on solutions to the nonlinear integral equation in $\mathcal{X} \times \mathcal{Y}$

$$D_y J_\pi(y|x) = \int D_y c(x, x', y, y')\pi(dx'dy') = 0. \tag{3.4}$$

As the goal is to establish that y is deterministically given by x, if $y \mapsto D_y J_\pi(y|x)$ were injective then for every given x the unique solution to $D_y J_\pi(y|x) = 0$ would specify y. However, we do not expect this to be the case in general (see Example 17). In optimal transportation there is a related necessary condition, which involves an additional term called a *Kantorovich Potential*. In particular, our necessary condition $\int D_y c = 0$ is replaced with $D_y c + D_y \psi = 0$, where ψ is known as the Kantorovich potential, and can be viewed as a Lagrange multiplier associated with the marginal constraints.

In OT, it is common to assume that $D_x c(x, \cdot)$ is injective (the so-called *twist condition*), which ensures that the equation $D\psi(x) + D_x c(x, y) = 0$ prescribes y uniquely from x. In our dimension reduction setting, no such term ψ appears, reflecting the absence of the additional marginal constraint; hence there is no analogous guarantee that y is uniquely determined by x.

In special cases, it can happen that the marginal problem is strictly convex as a function of y. We begin with a simple example in the context of classical dimension reduction algorithms.

Example 14 *Let* $c(x, x', y, y') = \big(\langle x, x'\rangle - \langle y, y'\rangle\big)^2$. *Then the marginal problem takes the form*

$$J_\pi(y|x) = x^T \left[\int x' x'^T \mu(dx')\right] x - 2x^T \left[\int x' y'^T \pi(dx'dy')\right] y + y^T \left[\int y' y'^T \nu(dy')\right] y.$$

Clearly, $y \mapsto J_\pi(y|x)$ *is convex and thus* $D_y J_\pi(y|x) = 0$ *will determine* y *given* x. *Writing the critical point equation, we see*

$$\left[\int y' x'^T \pi(dx'dy')\right] x = \left[\int y' y'^T \nu(dy')\right] y, \tag{3.5}$$

indicating that the optimal map is linear, meaning $y = Ax$. If we utilize the singular value decomposition $A = U\Sigma V^T$, we can rewrite the original optimization problem as

$$\mathcal{J}(\pi) = \mathcal{J}(A) = \iint (x^T x' - x^T V \Sigma^T \Sigma V^T x')^2 \mu(dx)\mu(dx') = \iint (x^T V^T (I - \Sigma^T \Sigma) V x')^2 \mu(dx)\mu(dx').$$

This is equivalent, for centered μ, to principal component analysis.

We now further develop our understanding of the strict convexity of the marginal problem.

Lemma 15 *Let π be a minimizer of* (1.6)*, and let c satisfy* (**A2**) *and* ($\mathbf{A1_{IP}}$)*. Define S to be the smallest linear subspace which contains the support of ν. Then when restricted to S we have that $D^2_{yy} J_\pi$ is positive definite.*

Proof We simply note that

$$D^2_{yy} J_\pi(y|x) = \int c_{tt}(\langle x, x' \rangle, \langle y, y' \rangle) y' y'^T \pi(dx' dy').$$

As $c_{tt} > 0$ by ($\mathbf{A1_{IP}}$), and by the definition of S we immediately obtain that $D^2_{yy} J_\pi$ will be positive definite along S. ■

Building upon these facts, we can give the following simple corollary to Theorem 10.

Corollary 16 (Deterministic solutions a.k.a. Monge Maps for (IP) costs) *Suppose that c is of type* (**IP**) *and satisfies Assumptions* (**A1**), (**A2**) *and* ($\mathbf{A1_{IP}}$)*. Then any optimal solution of* (1.6) *is supported on the graph of a function, whose smoothness is controlled by the differentiability of $t \mapsto \tilde{c}(x, x', t)$.*

Proof By Lemma 15, we have that the marginal problem is strictly convex when restricted to S, and therefore has a unique minimizer within S. In turn Theorem 10 implies that is only supported on those unique minimizers, and hence must be concentrated on the graph of a function.

■

This corollary resolves the necessity of optimal solutions to be mappings in many natural contexts, specifically costs which are convex in $\langle y, y' \rangle$. Such costs include classical multidimensional scaling and kernel principal component analysis.

3.2 Structure of marginal minimizers for normed costs

While the inner product case is often straightforward to study using marginal minimization, the case based upon norm comparisons is not as simple. This is due to a loss of convexity in y, which leads to more complicated structure of the marginal minimizers. In this subsection we study the form that these marginal minimizers take for the q-MDS case, highlighting i) the existence of multiple minimizers, ii) their parametric structure, iii) how this induces discontinuities in the learned embeddings. This also helps to explain the structure observed

in Example 1. While this section is important in motivating the difference with the inner product case and the need for distinct mathematical tools, readers can access the results in the balance of the paper without knowing the details in this section.

However, many of the standard costs used in dimension reduction are non-convex in y, and have marginal problems with more complicated structure in their minimizers. We return to our running example which demonstrates that the marginal problem can have multiple minimizers.

Example 17 *In the case of* $c(x,x',y,y') = \left(|x-x'|^2 - |y-y'|^2\right)^2$*, one has a rather explicit formula for the marginal problem:*

$$J_\pi(y|x) = |y|^4 - 2y^T\psi_\pi(x)y - 4\varphi_\pi(x)^T y + \zeta_\pi(x). \tag{3.6}$$

The coefficients of this polynomial equation are implicitly defined by moments of the joint distribution, in particular:

$$\begin{aligned}
\psi_\pi(x) &= \mathrm{Id_m}|x|^2 - \int \Big[2y'y'^T + (|y'|^2 - |x'|^2)\mathrm{Id_m}\Big]\pi(dx'dy') \\
\varphi_\pi(x) &= 2\underbrace{\left(\int y'x'^T\pi(dx'dy')\right)}_{=:\Phi_\pi} x + \int y'(|y'|^2 - |x'|^2)\pi(dx'dy') \\
\zeta_\pi(x) &= |x|^4 + 4x^T\left(\int x'x'^T\mu(dx')\right)x - 2|x|^2\left(\int \left(|y'|^2 - |x'|^2\right)\pi(dx'dy')\right) \\
&\quad + 4\left(\int \left(|y'|^2 - |x'|^2\right)x'^T\pi(dx'dy')\right)x + \int \left(|y'|^2 - |x'|^2\right)^2\pi(dx'dy')
\end{aligned}$$

where we have assumed the distribution in $\mathbb{R}^m$ *has mean zero by using translation invariance. Similarly, we assume that* μ *is also mean zero. We notice that the matrices* ψ_π, φ_π*, and* ζ_π*, which are completely determined by moments of* π*, give a parametric representation for the marginal problem, just as* A *did in the inner product case from the previous example. We believe that this parametric representation should be useful for many unsupervised learning tasks, as it will directly give properties such as statistical consistency and direct extrapolation. Furthermore, it should facilitate more efficient computational algorithms that work in parameter space: this is the subject of current work.*

Let $\eta_1, ..., \eta_m$ *be an orthogonal basis for which*

$$\sum_{j=1}^m \eta_j\eta_j^T = \int \Big[2yy^T + \left(|y|^2 - |x|^2\right)\mathrm{Id}_m\Big]\pi(dx\,dy)$$

so that

$$\psi_\pi(x) = |x|^2\mathrm{Id_m} - \sum_{j=1}^m \eta_j\eta_j^T \tag{3.7}$$

. For simplicity, assume that $|\eta_1| < |\eta_2| < \cdots < |\eta_m|$*. Evaluating the marginal problem along the lines* $r_i(t) = t\frac{\eta_i}{|\eta_i|}$ *one finds*

$$\frac{d}{dt}J_\pi(r_i(t)|x) = 4\left[t^3 - \left(|x|^2 - |\eta_i|^2\right)t - \frac{\varphi_\pi^T(x)\eta_i}{|\eta_i|}\right],$$

which can have multiple solutions along $r_i(t)$ provided $|x| > |\eta_i|$.

This alone is not necessarily a problem under Theorem 10 in that the marginal problem may have several critical points, but as long as there is a unique global minimizer we may still guarantee existence of non-probabilistic solutions for dimension reduction problem. This said, consider the set $\{x \,|\, \varphi_\pi(x) = 0\}$ where the critical point equation can be expressed as

$$\Big(|y|^2 - |x|^2\Big)y + \sum_{j=1}^{m} \eta_j \eta_j^T y = 0.$$

One may readily check that the solutions to the above equation are exhausted by $y = \pm\frac{\eta_j}{|\eta_j|}\sqrt{|x|^2 - |\eta_j|^2}$ for $j = 1, ..., m$ and $y = 0$. The previous observations imply that the former case is only possible when $|x| > |\eta_j|$ which makes the square root well defined. Plugging in each of these critical points into the marginal problem, we find that

$$J_\pi\left(\pm\tfrac{\eta_j}{|\eta_j|}\sqrt{|x|^2 - |\eta_j|^2}\,\middle|\, x\right) = -\big(|x|^2 - |\eta_j|^2\big)^2 + \zeta_\pi(x) \geq -\big(|x|^2 - |\eta_i|^2\big)^2 + \zeta_\pi(x).$$

where i is the largest index for which $|x| \leq |\eta_{i+1}|$. Hence for $\{x \,|\, \varphi_\pi(x) = 0, |\eta_i| < |x| \leq |\eta_{i+1}|\}$, there are two minimizers to the marginal problem: $\pm\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2 - |\eta_i|^2}$. The case devolves further if $|\eta_j|$ is repeated ($|\eta_1| < \cdots < |\eta_j| = \cdots = |\eta_{j+k-1}| < \cdots < |\eta_m|$) and $|x| \leq |\eta_{j+1}|$ where any y on the k-sphere spanned by $\eta_j, ..., \eta_{j+k-1}$ is a minimizer of $J_\pi(\cdot|x)$.

The previous example is meant to demonstrate how pathological the nature of the marginal minimization problem can be: *for simple costs, the marginal minimizers may be comprised of entire sub-manifolds in $\mathbb{R}^m$ for a single x!* In the pursuit of deterministic minimizers (i.e. Monge-type maps), one approach might be to show that these multiple minimizers can only happen on a thin set (in the above example this corresponds to showing that $\varphi_\pi(x)$ is full rank $\mu(dx)$-a.e.), but due to the implicit dependence of the marginal problem on the embedding plan π, taking this route directly has proven particularly difficult.

Another notable consequence which can be observed from the marginal problem framework is that for normed costs, it will be likely that there will be jump discontinuities arising from an analogous phenomenon to that of Example 29. The following example shows that in the case of q-MDS, we can guarantee discontinuities in the optimal solution. We can expect the argument below to persist for any dimension reduction problem for which $\arg\min J_\pi(\cdot|x)$ has multiple values for some x, but this property is implicitly dependent on π as well and thus challenging to verify in practice.

Example 18 *Putting technicalities of the rank of $D_y J_\pi(y|x)$ aside for the moment, Example 17 in the previous section argues that when none of the lengths of $|\eta_j|$ are repeated, there are $m + 1$ distinct regions for which the marginal problem is defined by a different solution. More precisely, for $A_i := \{x \,|\, \varphi_\pi(x) = 0, |\eta_i| < |x| \leq |\eta_{i+1}|\}$, we have a semi-explicit (governed by moments of the optimal solution) formula for the reduction map: $T(x) = \pm\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2 - |\eta_i|^2}$; when x passes from A_i to A_{i+1} the optimal solution abruptly jumps from $\pm\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2 - |\eta_i|^2}$ to $\pm\frac{\eta_{i+1}}{|\eta_{i+1}|}\sqrt{|x|^2 - |\eta_{i+1}|^2}$.*

Beyond this, one also can observe that for any path $x + \varepsilon v$ with $\varphi_\pi(x) = 0$ (and v not in the nullspace of Φ_π) the marginal minimizer has a jump discontinuity at $\varepsilon = 0$. The

intuition here will come from Example 29. Indeed, the previous considerations have implied that there will be multiple minimizers when $\varphi(x) = 0$, *namely* $\pm\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2 - |\eta_i|^2}$ *where* i *is the smallest index such that* $|x| > |\eta_i|$. *By plugging* $\pm\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2 - |\eta_i|^2}$ *into the marginal problem* (3.6) *at* $x + \varepsilon v$. *We see by* (3.7),

$$\begin{aligned}
J_\pi\big(\pm\tfrac{\eta_i}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2}\,\big|\,x+\varepsilon v\big) &= \big(|x|^2-|\eta_i|^2\big)^2 \\
&\quad - 2\tfrac{\eta_i^T}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2}\Big(|x+\varepsilon v|^2\mathrm{Id}_m - \sum_{j=1}^m \eta_j\eta_j^T\Big)\tfrac{\eta_i}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2} \\
&\quad \pm 4\varphi_\pi^T(x+\varepsilon v)\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2} + \zeta_\pi(x+\varepsilon v) \\
&= \big(|x|^2-|\eta_i|^2\big)^2 - 2\Big(|x|^2-|\eta_i|^2\Big)\Big(|x+\varepsilon v|^2-|\eta_i|^2\Big) \\
&\quad \pm 8\varepsilon v^T\Phi_\pi^T\frac{\eta_i}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2} + \zeta_\pi(x+\varepsilon v)
\end{aligned}$$

where on the last line we have used the fact that $\varphi_\pi(x+\varepsilon v) = 2\varepsilon\Phi_\pi v$. *Crucially, we see that in order for the above expression to be minimal, one needs to choose the sign of the order* ε *term to be opposite that of* $v^T\Phi_\pi^T\eta_i$. *In particular, this shows that near a point* x *for which* $\varphi_\pi(x) = 0$, *the optimal map is*

$$T(x+\varepsilon v) = -\mathrm{sign}(\varepsilon v^T\Phi_\pi^T\eta_i)\tfrac{\eta_i}{|\eta_i|}\sqrt{|x|^2-|\eta_i|^2} + \mathcal{O}(\varepsilon)$$

whose limit does not exist at $\varepsilon = 0$.

Now having seen the possibility of multiple minimizers to the marginal problem and how it can cause discontinuities, we illustrate one more useful perspective in the context of dimension reduction. Being that dimension reduction schemes inherently discard information while representing data in the embedded space, there must be some partition of $\mathcal{X}$ such that each element of the partition may be represented by a single value in the embedding. More precisely, for the map outlined in Definition 9, the set $\{x : \lambda(x) = y\}$ represents all of the points in $\mathcal{X}$ which are optimally embedded to the vector y. While these sets can be arbitrary, we expect them to form $d - m$ dimensional manifolds. To illustrate this, we present one more example.

Example 19 *Let us consider a simple example where 1000 datapoints in* $\mathbb{R}^2$ *are such that 500 points are stacked at* $(0, 1)$ *and the other 500 are stacked at* $(0, -1)$. *The optimal embedding for the q-MDS cost into* $\mathbb{R}$ *is clearly realized by projecting the 2 dimensional dataset onto the y-axis. This allows us to explicitly compute*

$$\psi_\pi(x_1, x_2) = x_1^2 + x_2^2 - 2, \qquad \varphi_\pi(x_1, x_2) = 2x_2$$

thus the critical point equation can be written $y^3 - (x_1^2 + x_2^2 - 2)y = 2x_2$. *Imitating the previous computations, we first notice that when* $|x| < \sqrt{2}$, $\psi_\pi(x_1, x_2) < 0$. *This implies that on disk of radius* $\sqrt{2}$, *the marginal problem* (3.6) *has a unique solution. Indeed for* x *in the set* $\{x : |x| < \sqrt{2}\}$,

$$\frac{d^2}{dy^2}J_\pi(y|x) = 12y^2 - 4\psi_\pi(x) > 0.$$

Furthermore, if $|x| \geq \sqrt{2}$*, there are be multiple minimizers along the set* $\{x : \varphi_\pi(x) = 0\}$ *which will lead to a jump discontinuity as predicted in Example 18. The figure below illustrates the level sets of the minimizers,* $\lambda_\pi : \mathbb{R}^2 \to \mathbb{R}$*.*

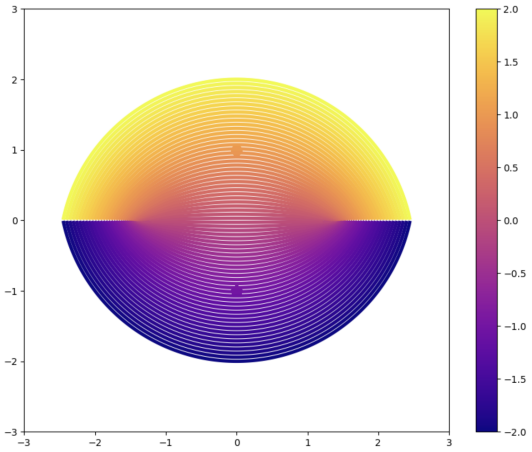


Figure 2: Each curve represents an equivalence class of points in $\mathbb{R}^2$ which all have the same minimizer in $\mathbb{R}$ for the embedding outlined in Example 19. Notice that once $|x| > \sqrt{2}$, the line $x_2 = 0$ has a discontinuity surface.

4. Normed costs: maps via second-order conditions

In this section, we show that for a wide range of normed costs the solution of the dimension reduction problem 1.6 is induced by a map. Namely, we prove the following:

Theorem 20 (Deterministic solutions a.k.a. Monge Maps for $(\mathbf{N}^2)$ **costs)** *Let* $\mu \in \mathscr{P}(\mathbb{R}^d)$ *with cost structure* $(\mathbf{N}^2)$ *satisfying assumptions* $(\mathbf{A1})$*,* $(\mathbf{A2})$*, and* $(\mathbf{A1}_{\mathbf{N}^2})$ *and* $(\mathbf{A2}_{\mathbf{N}^2})$*. Then solutions to* $\min_{\pi \in \Pi(\mu)} \mathcal{J}(\pi)$ *are concentrated on the graph of a function; i.e. there is a measurable* $T : \mathbb{R}^d \to \mathbb{R}^m$ *such that* $\pi(dy|x) = \delta_{T(x)}(dy)$*,* $\mu(dx)$*-almost everywhere. More succinctly, solutions to the dimension reduction problem exist and are necessarily deterministic.*

The main difficulty is that the dimension reduction problem of type $(\mathbf{N}^2)$ is not marginally convex in y (i.e. $y \mapsto J_\pi(y|x)$ is not convex) and thus we expect multiple minimizers to a given marginal problem (take for instance Example 1). This follows since $y \mapsto c(|x - x'|^2, |y - y'|^2)$ need not be convex *even when the function* $t \mapsto c(s,t)$ *is convex.* To surmount this, we track the second-order effect of perturbations, on the level of the dimension reduction plans. We shall see that certain natural structural conditions upon c (namely assumptions $(\mathbf{A1}_{\mathbf{N}^2})$ and $(\mathbf{A2}_{\mathbf{N}^2})$ which we restate for the readers convenience below) will then be sufficient to guarantee that optimal plans are induced by maps.

$(\mathbf{A1}_{\mathbf{N}^2})$ $\quad \partial_t \tilde{c}_{\mathbf{N}^2}(0,t) \geq 0$ *with equality uniquely attained at* $t = 0$*.*

$(\mathbf{A2}_{\mathbf{N}^2})$ $\quad \partial_t \tilde{c}_{\mathbf{N}^2}(0,t) + 2t\, \partial_{tt} \tilde{c}_{\mathbf{N}^2}(0,t) \geq 0$ *with equality uniquely attained at* $t = 0$*.*

As in Section 3, we motivate our proofs by first formally considering the case where the input distribution is realized as a sum of Dirac masses, $\mu = (1/n)\sum_{i=1}^{n} \delta_{x_i}$, for some collection of distinct points $x_1, x_2, ..., x_n \in \mathbb{R}^d$. As in Section 3, we assume that the optimal embedding may be represented discretely by $\pi = (1/n)\sum_{i,j=1}^{n} \pi_{ij}\delta_{(x_i,y_j)}$ for some distinct collection of vectors $y_1, y_2, ..., y_n \in \mathbb{R}^m$. Suppose that in the ith row of π there are at least two nonzero entries and reorder the y's so that $\pi_{ii}, \pi_{ij} > 0$; this essentially encodes the situation where an optimal embedding maps a single x to multiple y's.

By Theorem 10, both $y_i, y_j \in \arg\min J_\pi(\cdot|x_i)$ and thus we can transport the mass stored at (x_i, y_j) to (x_i, y_i) without violating our first-order condition. More precisely, if $\pi_{ii} > 0$ and $\pi_{ij} > 0$, the perturbation

$$\gamma\!\!\!\!\gamma = \min\{\pi_{ii}, \pi_{ij}\}\big[\delta_{(x_i,y_i)} - \delta_{(x_i,y_j)}\big],$$

is well-defined and will have $\mathcal{J}(\gamma\!\!\!\!\gamma|\pi) = 0$, meaning that it will leave the energy unchanged up to second-order variations. When we compute the quadratic term, we have

$$\gamma\!\!\!\!\gamma \otimes \gamma\!\!\!\!\gamma = \min\{\pi_{ii}, \pi_{ij}\}^2 \Big[\delta_{(x_i,y_i)} \otimes \delta_{(x_i,y_i)} - \delta_{(x_i,y_i)} \otimes \delta_{(x_i,y_j)} - \delta_{(x_i,y_j)} \otimes \delta_{(x_i,y_i)} + \delta_{(x_i,y_j)}\delta_{(x_i,y_j)}\Big]$$

and thus

$$\begin{aligned}\mathcal{J}(\gamma\!\!\!\!\gamma) &= \min\{\pi_{ii}, \pi_{ij}\}^2\big[c(|x_i - x_i|^2, |y_i - y_i|^2) - 2c(|x_i - x_i|^2, |y_i - y_j|^2) \\ &\quad + c(|x_i - x_i|^2, |y_j - y_j|^2)\big] \\ &= 2\min\{\pi_{ii}, \pi_{ij}\}^2\big[c(0,0) - c(0, |y_i - y_j|^2)\big].\end{aligned}$$

Crucially, if $t \mapsto c(0,t)$ has a strict global minimum at $t = 0$, which is implied by assumption $(\mathbf{A1_{N^2}})$, then $\mathcal{J}(\gamma\!\!\!\!\gamma) < 0 \iff y_i \neq y_j$. This implies that for each x_i the optimal plan must be supported only on a single y. In spite of the technical difficulties engendered by the loss of lower semicontinuity, this discrete argument suggests a very strong result: that solutions to the relaxed problem (1.6) with normed cost *must be deterministic*, and hence must solve the original problem (1.1). This is quite surprising in light of the examples presented in the introduction suggesting that Young measures can be encountered in practice. This is because the perturbations used in particle-based optimization methods cannot carry out perturbations of the form $\gamma\!\!\!\!\gamma$, and can get stuck in local minima with respect to particle-wise descent.

To extend this argument to the general setting, one must be able to represent solutions to the marginal problem locally in a consistent manner. In particular, it would be ideal to obtain more structure on the nature of the marginally minimizing set-valued map λ as outlined in Definition 9. Leaving technical justification aside for the moment, suppose that locally λ admits a countable representation, i.e. $\lambda(x) = \bigcup_{i=1}^{\infty} \lambda_i(x)$ for a sequence of functions λ_i. We then can leverage the discrete argument between pairs of these functions through the following proposition.

Proposition 21 *Let $\lambda_1, \lambda_2 : B_\delta(x_0) \to \mathcal{Y}$ be measurable functions with $\lambda_1(x) \neq \lambda_2(x)$ for all $x \in B_\delta(x_0)$. Assume that c is a continuous cost of type $(\mathbf{N}^2)$ and satisfies assumptions $(\mathbf{A1})$ and $(\mathbf{A1_{N^2}})$ and assume π is a minimizer of (1.6). Let μ_1, μ_2 be the $\mathcal{X}$-marginal measures of π restricted to the sets $y = \lambda_1(x)$ and $y = \lambda_2(x)$ and $x \in B_\delta(x_0)$. Then μ_1 is mutually singular to μ_2, or in symbols $\mu_1 \perp \mu_2$, meaning that they have disjoint supports.*

Proof We first outline the proof in the case where the λ_i are continuous, and comment on the measurable case at the end. Suppose, for the sake of contradiction, that μ_1 and μ_2 are not mutually singular. Then the measure $\mu_1 \wedge \mu_2 = \mu_1 - (\mu_1 - \mu_2)^+$ is not a zero measure, and we may select a point $\bar{x} \in B_\delta(x_0)$ so that $\mu_1 \wedge \mu_2(B_\varepsilon(\bar{x})) > 0$ for all $\varepsilon > 0$ sufficiently small.

We then construct the perturbation, restricted to $x \in B_\varepsilon(\bar{x})$, via

$$\gamma(dx\, dy) = \mu_1 \wedge \mu_2(dx)\Big[\delta_{\lambda_1(x)}(dy) - \delta_{\lambda_2(x)}(dy)\Big].$$

By construction we have that $\pi + \gamma$ is a probability measure and retains the same $\mathcal{X}$ marginal as π. We also note that γ is not the zero measure by choice of $\bar{x}$.

As long as μ_1, μ_2 are non-trivial, then by Theorem 10 we know that λ_1, λ_2 must be minimizers of the marginal problem on the support of μ_1, μ_2. Using the notation from the proof of Theorem 10 we have that

$$\mathcal{J}(\gamma|\pi) = \int J_\pi(y|x)\gamma(dx\, dy) = \int_{B_\varepsilon(\bar{x})} [\min J_\pi(\cdot|x) - \min J_\pi(\cdot|x)]\mu_1 \wedge \mu_2(dx) = 0.$$

The overall change in the quadratic term is given by,

$$\begin{aligned}\mathcal{J}(\gamma) = &\iint_{B_\varepsilon(\bar{x})\times B_\varepsilon(\bar{x})} c(|x-x'|^2, |\lambda_i(x) - \lambda_i(x')|^2)\mu_1 \wedge \mu_2(dx)\mu_1 \wedge \mu_2(dx') \\ &+ \iint_{B_\varepsilon(\bar{x})\times B_\varepsilon(\bar{x})} c(|x-x'|^2, |\lambda_j(x) - \lambda_j(x')|^2)\mu_1 \wedge \mu_2(dx)\mu_1 \wedge \mu_2(dx') \\ &- 2\iint_{B_\varepsilon(\bar{x})\times B_\varepsilon(\bar{x})} c(|x-x'|^2, |\lambda_i(x) - \lambda_j(x')|^2)\mu_1 \wedge \mu_2(dx)\mu_1 \wedge \mu_2(dx').\end{aligned}$$

By using the continuity of c, λ_1, λ_2, we then estimate

$$\mathcal{J}(\gamma) \leq 2\mu_1 \wedge \mu_2(B_\varepsilon(\bar{x}))^2(c(0,0) - c(0, |\lambda_1(\bar{x}) - \lambda_2(\bar{x})|^2) + \eta(\varepsilon)),$$

where η represents a local modulus of continuity and satisfies $\eta(\varepsilon) \to 0$ as $\varepsilon \to 0$. As $\lambda_1(\bar{x}) \neq \lambda_2(\bar{x})$, and $c(0,t)$ is strictly minimized at $t = 0$ by $(\mathbf{A1_{N^2}})$, we obtain that $\mathcal{J}(\gamma) < 0$, which contradicts the minimality of π. The measurable case follows by extending the proof using Lebesgue point arguments. ■

We can then establish the following corollary.

Corollary 22 *Let $\lambda_i : O_i \to Y$ be continuous functions, where O_i are open sets, and $i \in \{1, \ldots, \infty\}$. Let π be a minimizer of (1.6) for continuous $(\mathbf{N}^2)$ cost satisfying assumptions $(\mathbf{A1})$ and $(\mathbf{A1_{N^2}})$, and let $\tilde{\pi}$ be the restriction of π to the union of the sets $\{(x, \lambda_i(x)) : x \in O_i\}$. Then $\tilde{\pi}$ has support on the graph of a function.*

Proof As there are only countably many $\lambda_i(x)$, we can represent

$$\tilde{\pi}(dx\, dy) = \sum_{i=1}^{\infty} \mu_i(dx)\delta_{\lambda_i(x)}(dy),$$

and where each μ_i is the marginal measure associated with each of the different minimizing branches λ_i. By the previous proposition if $j \neq i$ then $\mu_j \perp \mu_i$. Let $A_i = \text{supp}(\mu_i)$, and note that the A_i are disjoint. In turn we then may define

$$\Lambda(x) = \sum_{i=1}^{\infty} \mathbb{1}_{A_i}(x)\lambda_i(x).$$

Then Λ is a measurable function and $\tilde{\pi}$ is supported on the graph of Λ. ■

The previous proposition offers a direct application to global minimizers of the marginal problem which have non-degenerate Hessian in y; namely those minimizers which are also strict local minimizers. We begin by proving two brief lemmas based upon the implicit function theorem.

Lemma 23 *Let π be a minimizer of 1.6 for cost satisfying $(\mathbf{N}^2)$ and $(\mathbf{A1})$, $(\mathbf{A2})$ and $(\mathbf{A1_{N^2}})$. Suppose that $y_1 \neq y_2$ are global minimizers of the marginal problem at $\bar{x}$, which both satisfy $D^2_{yy}J_\pi(y_i|\bar{x}) > 0$. Then there exists a $\delta > 0$ and C^1 functions $\lambda_i : B_\delta(\bar{x}) \to B_\delta(y_i)$, $i = 1, 2$ so that $\lambda_i(x)$ is the only strict local minimizer of the marginal problem in $B_\delta(y_i)$.*

Proof The minimality of y_1 and y_2 indicate that both $D_y J_\pi(y_1|\bar{x}) = 0$ and $D_y J_\pi(y_2|\bar{x}) = 0$. From the strict non-degeneracy assumption on $D^2_{yy}J_\pi$, the implicit function theorem allows us to construct C^1 maps $\lambda_i : B_\delta(\bar{x}) \to B_\delta(y_i)$ $i = 1, 2$, which uniquely solve $D_y J_\pi(\lambda_i(x)|x) = 0$ on the respective neighborhoods in the product space. We note that without loss of generality δ can be taken small enough to guarantee the strict local minimality of λ_1 and λ_2 since J_π was C^2. ■

Lemma 24 *Assume that c is a cost of type $(\mathbf{N}^2)$ and satisfies assumptions $(\mathbf{A1})$, $(\mathbf{A2})$ and $(\mathbf{A1_{N^2}})$. and let π be a minimizer of (1.6). Then for every x there exists at most a countable number of global minimizers of the marginal problem which satisfy $D^2_{yy}J_\pi(y|x) > 0$.*

Proof First note that since c is C^2 by **(A2)**, it follows from Lemma 13 that the marginal problem is a C^2 function in y. Furthermore, by **(A1)** the minimizers of the marginal problem at a point x must live in a compact set $K_x \subset \mathcal{Y}$. Consider the set $M_\eta \subset K_x$ of global minimizers of the marginal problem at x satisfying $|D^2_{yy}J_\pi(y|x)| \geq \eta$. We notice that M_η will also be compact. As $J_\pi(y|x)$ is C^2, each element of M_η can be surrounded by a ball of some radius $r_\eta > 0$ which contain no other point in M_η: this essentially says that a global minimizer with a lower bound on the Hessian is an isolated minimizer with a quantifiable distance of isolation. As M_η is compact, we then have that it actually must be finite. By taking η to zero, this argument shows that the number of minimizers with non-degenerate Hessian must be at most countable. ■

We now choose to decompose the optimal plan into points where the Hessian is non-degenerate (i.e. rank strictly less than m) and its complement via

$$\pi = \pi_S + \pi_I, \qquad \pi_S = \pi_{|\det(D^2_{yy}J_\pi)=0}, \qquad \pi_I = \pi_{|\det(D^2_{yy}J_\pi)\neq 0}. \tag{4.1}$$

We similarly let μ_S, μ_I, ν_S, and ν_I denote the associated marginal measures. In terms of this decomposition, we can use Corollary 22 along with Lemma 24 to immediately give the following.

Proposition 25 *Let π be a minimizer of (1.6), for c of type $(\mathbf{N}^2)$ satisfying assumptions $(\mathbf{A1})$, $(\mathbf{A2})$ and $(\mathbf{A1_{N^2}})$. Using the decomposition (4.1), then π_I is supported on the graph of a function.*

Remark 26 *In this theorem we notice that there are no requirements on the measure μ, nor on m, d. Furthermore, we notice that in the statement we can say that π is induced by a map on the set where $\rho_s := \frac{d\pi_S(\cdot \times \mathcal{Y})}{d\mu} > 0$, and not just π_S. Hence any part of the support not covered by Proposition 28 will be covered by Proposition 25.*

The only remaining point is to rule out multivaluedness at points where the Hessian of the marginal problem is degenerate. We begin with the following lemma.

Lemma 27 *For costs of type $(\mathbf{N}^2)$, the maximum eigenvalue of the matrix $D^2_{yy'}c(x, x', y, y')$ is a function only of $|x - x'|^2$ and $|y - y'|^2$. Furthermore, under assumptions $(\mathbf{A1_{N^2}})$ and $(\mathbf{A2_{N^2}})$, when $x = x' = x_0$, this eigenvalue $\sigma_{max}(0, |y - y'|^2) \leq 0$ with equality only if $y = y'$.*

Proof Recalling that in the case of $(\mathbf{N}^2)$ we write $c(x, x', y, y') = \tilde{c}_{\mathbf{N}^2}(|x - x'|^2, |y - y'|^2)$. With the notation that $s = |x - x'|^2$ and $t = |y - y'|^2$ we can write

$$D^2_{yy'}c(x, x', y, y') = -2\partial_t \tilde{c}_{\mathbf{N}^2}(|x-x'|^2, |y-y'|^2)\mathrm{Id}_m - 4\,\partial_{tt}\tilde{c}_{\mathbf{N}^2}(|x-x'|^2, |y-y'|^2)(y-y')(y-y')^T.$$

Since this is a rank one perturbation of an identity matrix, it is easily checked that the eigenvalues of the above matrix are

$$\begin{aligned} \sigma_1 &= -2\partial_t \tilde{c}_{\mathbf{N}^2}(|x - x'|^2, |y - y'|^2) - 4\partial_{tt}\tilde{c}_{\mathbf{N}^2}(|x - x'|^2, |y - y'|^2)|y - y'|^2, \\ \sigma_2, ..., \sigma_m &= -2\partial_t \tilde{c}_{\mathbf{N}^2}(|x - x'|^2, |y - y'|^2). \end{aligned}$$

Plugging in $x = x' = x_0$, the assumptions $(\mathbf{A1_{N^2}})$ and $(\mathbf{A2_{N^2}})$ it immediately follows that that $\sigma_{max}(0, |y - y'|^2) \leq 0$ with equality only when $y = y'$. ∎

Proposition 28 *Assume c is a cost of type $(\mathbf{N}^2)$ and satisfies assumptions $(\mathbf{A1})$, $(\mathbf{A2})$, $(\mathbf{A1_{N^2}})$, $(\mathbf{A2_{N^2}})$. If π is an optimal plan, then π_S is supported on the graph of a measurable function $\lambda_S : \mathcal{X} \to \mathcal{Y}$.*

Proof Assume that π is an optimal solution of 1.6 and define $E_{deg} := \{(x, y) : \det(D^2_{yy}J_\pi(y|x)) = 0\}$. We then choose a measurable function $\phi : E_{deg} \to \mathbb{S}^{m-1}$ such that

$$D^2_{yy}J_\pi(y|x) \cdot \phi(x, y) = 0.$$

The existence of such a function can be justified using measurable selections of the multi-function encoding the nullspace of $D^2_{yy}J_\pi(y|x)$, see for example (Rockafellar et al., 2009). We will consider a perturbation which, for every $y \in \mathrm{Spt}\,(\mathrm{proj}_{\mathcal{Y}}\sharp\pi)$, pushes y in the direction

of $\phi(x,y)$ (a degenerate direction of $D^2_{yy}J_\pi$) - in some sense this perturbation modifies the support so that it is using more of the available space in the embedded dimension. We extend ϕ to be equal to zero off of the set E_{deg}.

Construct a measure on $\mathcal{X}$ by $\mu_S(A) = \pi_S(A \times \mathcal{Y})$ and note that $\mu_S \ll \mu$. Denote $\rho_S(x) := \frac{d\mu_S}{d\mu}(x)$ as the Radon-Nikodym derivative of μ_S with respect to μ. Since $\mu_S(\mathcal{X}) = \pi(E_{deg}) > 0$, (otherwise there would be nothing to show) we know there is a set of positive μ-measure for which $\rho_S > 0$; let x_0 be a density point (a point where $\lim_{\delta\to 0}\frac{\mu_S(B_\delta(x_0))}{\mu(B_\delta(x_0))} = \rho_S(x_0)$) in this set.

We notice that by the same argument as in Lemma 24, we have that for all $(x,y),(x',y')$ in the support of π such that $x,x' \in B(x_0,1)$ we will have that y,y' live in a compact set. Thus using (A1) we may bound $\|D_{y,y'}c(x,x',y,y')\|_2 \le M$ for all $(x,y),(x',y')$ belonging to the support of π_S with $x,x' \in B(x_0,1)$, where here we are using $\|\cdot\|_2$ to denote the 2-operator norm.

Next, we note that for every $\theta > 0$, there exists a unit vector v for which

$$\limsup_{\delta\to 0^+}\frac{\pi_S(\left[B_\delta(x_0)\times\mathcal{Y}\right]\cap\{(x,y):\phi(x,y)\cdot v > \cos(\theta)\})}{\mu(B_\delta(x_0))} =: \tilde{\rho}(x_0) > 0.$$

Such a vector exists since one may cover $B_\delta(x_0)$ with a finite number of cones with apex x_0 and opening angle 2θ; by the positivity of $\rho_S(x_0)$ and a pidgeonhole argument the limsup must be strictly positive for at least one of those cones.

We let $E_{align} = \{(x,y):\phi(x,y)\cdot v > \cos(\theta)\}$. For $\delta > 0$, define $E_\delta = \mathrm{Spt}\,(\mu_S)\cap B_\delta(x_0)\times\mathcal{Y}]\cap E_{align}$ and for every $\varepsilon > 0$, define $\varphi^\varepsilon_x(y) := y + \varepsilon \mathbb{1}_{E_\delta}(x,y)\phi(x,y)$ and an associated family of plans π_ε by writing

$$\pi_\varepsilon(dx\,dy) = \varphi^\varepsilon_x\sharp\nu(dy|x)\mu(dx)$$

where $\pi(dx\,dy) = \nu(dy|x)\mu(dx)$ by disintegration. To assess the effect of this perturbation, we compute $\mathcal{J}(\pi_\varepsilon) - \mathcal{J}(\pi)$

$$\begin{aligned}
\mathcal{J}(\pi_\varepsilon) - \mathcal{J}(\pi) &= \iint [c(x,x',y+\varphi^\varepsilon_x(y),y'+\varphi^\varepsilon_{x'}(y')) - c(x,x',y,y')]\pi(dx\,dy)\pi(dx'dy') \\
&= \varepsilon\iint_{E_\delta\times(\mathcal{X}\times\mathcal{Y})} D_y c(x,x',y,y')\phi(x,y)\pi(dx\,dy)\pi(dx'dy') \\
&+ \varepsilon\iint_{(\mathcal{X}\times\mathcal{Y})\times E_\delta} D_{y'} c(x,x',y,y')\phi(x',y')\pi(dx\,dy)\pi(dx'dy') \\
&+ \frac{\varepsilon^2}{2}\iint_{E_\delta\times(\mathcal{X}\times\mathcal{Y})} \phi^T(x,y)D^2_{yy}c(x,x',y,y')\phi(x,y)\pi(dx\,dy)\pi(dx'dy') \\
&+ \frac{\varepsilon^2}{2}\iint_{(\mathcal{X}\times\mathcal{Y})\times E_\delta} \phi^T(x',y')D^2_{y'y'}c(x,x',y,y')\phi(x',y')\pi(dx\,dy)\pi(dx'dy') \\
&+ \varepsilon^2\iint_{E_\delta\times E_\delta} \phi^T(x,y)D^2_{yy'}c(x,x',y,y')\phi(x',y')\pi(dx\,dy)\pi(dx'dy') + o(\varepsilon^2),
\end{aligned}$$

where we have used the C^2 part of assumption (A2) to obtain the $o(\varepsilon^2)$ estimate. The $\mathcal{O}(\varepsilon)$ terms are the same by the symmetry of the cost. Applying Fubini's Theorem yields

$$\iint_{E_\delta \times (\mathcal{X}\times\mathcal{Y})} D_y c(x,x',y,y')\phi(x,y)\pi(dx\,dy)\pi(dx'dy') = \int_{E_\delta} D_y J_\pi(y|x)\phi(x,y)\pi(dx\,dy)$$

which vanishes as the optimality of π implies that $D_y J_\pi(y|x) = 0$ π-a.e according to Theorem 10. Similarly, the former two $\mathcal{O}(\varepsilon^2)$ terms are also the same by symmetry, so again by Fubini one has

$$\begin{aligned}&\iint_{E_\delta \times (\mathcal{X}\times\mathcal{Y})} \phi^T(x,y) D^2_{yy} c(x,x',y,y')\phi(x,y)\pi(dx\,dy)\pi(dx'dy')\\ =&\int_{E_\delta} \phi^T(x,y) D^2_{yy} J_\pi(y|x)\phi(x,y)\pi(dx\,dy).\end{aligned}$$

Due to the definition of ϕ, this integrand is zero on E_δ. On the other hand we can write

$$\iint_{E_\delta \times E_\delta} \phi^T(x,y) D^2_{yy'} c(x,x',y,y')\phi(x',y')\pi(dx\,dy)\pi(dx'dy') \tag{4.2}$$

$$= \iint_{E_\delta \times E_\delta} \big[v^T D^2_{yy'} c(x,x',y,y')v + 2v^T D^2_{yy'} c(x,x',y,y')(\phi(x,y)-v) \tag{4.3}$$

$$+ (\phi(x,y)-v)^T D^2_{yy'} c(x,x',y,y')(\phi(x',y')-v)\big]\pi(dx\,dy)\pi(dx'dy') \tag{4.4}$$

where we add and subtract v and use the symmetry that $D^2_{yy'}c(x,x',y,y') = D^2_{yy'}c(x',x,y',y)$. Dealing with the latter two terms first, we note that since $(x,y) \in E_\delta \subset E_{align}$, $|\phi(x,y)-v| \leq \sqrt{2-2\cos(\theta)}$, which further implies that

$$(\phi(x,y)-v)^T D^2_{yy'} c(x,x',y,y')(\phi(x',y')-v) \leq M(2-2\cos(\theta)).$$

A similar estimate holds for the middle term. Finally, using the symmetry of $D^2_{yy'}c$, we bound $v^T D^2_{yy'}c(x,x',y,y')v$ from above by $\sigma_{max}(|x-x'|^2, |y-y'|^2)$, the maximum eigenvalue for $D^2_{yy'}c(x,x',y,y')$. By optimality of π, and by taking $\varepsilon \to 0$ in order to neglect the $o(\varepsilon^2)$ term, we have that the expression in (4.4) must be non-negative. Rearranging, we obtain

$$-\iint_{E_\delta \times E_\delta} \sigma_{max}(|x-x'|^2, |y-y'|^2)\pi(dx\,dy)\pi(dx'dy') \tag{4.5}$$

$$\leq \iint_{E_\delta \times E_\delta} 2\sqrt{2}M\sqrt{1-\cos(\theta)}\pi(dx\,dy)\pi(dx'dy'). \tag{4.6}$$

If we divide by $\mu(B_\delta)^2$ and take $\delta \to 0$ in a way that the limsup in the definition of $\tilde{\rho}(x_0)$ is attained, we then obtain

$$\begin{aligned}&\tilde{\rho}^2(x_0)\iint_{E_{x_0}\times E_{x_0}} -\sigma_{max}(0, |y-y'|^2)\nu(dy|x_0)\nu(dy'|x_0)\\ &\leq \tilde{\rho}^2(x_0) 2\sqrt{2}M\sqrt{1-\cos(\theta)}\end{aligned}$$

where we've abbreviated $E_{x_0} = \{y : (x_0, y) \in E_{deg} \cap E_{align}\}$. Here by $\tilde{\rho}$ we mean the density associated with the set E_{x_0}. Dividing by $\tilde{\rho}(x_0)$, which is strictly positive, and then taking

$\theta \to 0$, then implies single-valuedness of π_S on E_{deg}, as from Lemma 27, $\sigma_{max}(0, |y-y'|^2) \leq 0$ with equality only when $y = y'$, which completes the proof. ■

Using Propositions 25 and 28 one obtains two maps $\lambda_I, \lambda_S : \mathbb{R}^d \to \mathbb{R}^m$ which may be supported by π. Applying Proposition 21 once more to these two maps establishes Theorem 20.

In (Dumont et al., 2024), the GW problem with norm squared costs was addressed under both marginal constraints, leading to 2-map solutions. In our case, we remove the second marginal constraint and study the corresponding projection problem. Although this setting is less constrained, we establish the existence of Monge maps and prove that optimality in the embedding problem necessarily forces a deterministic solution — a conclusion that is perhaps unexpected in light of Example 1.

Additionally, the method of tracking second variations appears well-suited for advancing the analysis of the full GW problem, with indications that it may lead to map-based solutions under both marginal constraints with some additional assumptions — a direction we plan to detail in forthcoming work.

5. Conclusion

In this work we have examined theoretical properties of some fundamental dimension reduction algorithms. We have shown that, for natural costs based upon similarities (i.e. inner products), and dissimilarities (i.e. norm differences), that the dimension reduction problem must be minimized by a deterministic mapping, and that any probabilistic behavior is necessarily sub-optimal.

On the other hand, the behavior that we observe in Example 1 raises many difficult questions. Clearly local minimizers found using naive particle descent methods may exhibit probabilistic behavior, which is consistent with the failure of lower-semicontinuity we proved in Proposition 4. On the level of practical applicability, we find such probabilistic behavior highly problematic. For example, it could lead to very misleading clustering in data visualization, where similar points in feature space are probabilistically assigned to distinct clusters.

These issues raise many natural follow up questions, a few of which we list here:

- Are the issues with probabilistic minimizers found via particle descent methods still present in real-world data sets? We have not pursued this issue here because comprehensively addressing this question calls for a detailed study across numerous benchmark data sets.

- What computational methods can be developed to avoid spurious probabilistic behavior in dimension reduction, and how can the necessary conditions identified in this work be used to do so?

- If non-linear dimension reduction algorithms often induce discontinuous embeddings, how greatly can they modify the topology of the data in feature space?

- Can the approach developed here for the projection problem be adapted to the full GW problem with norm squared costs to obtain Monge maps under both marginal constraints, and what additional challenges would need to be addressed in doing so?

Acknowledgements

The authors gratefully acknowledge the support of NSF DMS 2307971 and the Simons Foundation MP-TSM. AP also warmly thanks Peter McGrath and Erik Bates for many helpful discussions about early versions of the work.

Appendix A. Theory for variational problems

The direct method of the calculus of variations seeks to generalize the extreme value theorem in finite dimensions to infinite dimensional optimization problems. It proves the existence of minimizers of a functional $\mathcal{I} : U \to \mathbb{R}$, where U is an infinite-dimensional space, by combining the following assumptions:

1. **Coercivity:** Given some set $B \subset U$ we have that $\mathcal{I}(B^c) > \inf_U \mathcal{I}$.
2. **Compactness:** Under some topology τ we have that B is sequentially compact.
3. **Continuity:** Under that same topology, the functional $\mathcal{I}$ is sequentially lower semi-continuous.

One then directly shows the existence of minimizers by taking the following steps: i) Construct a sequence of functions $u_n \in B$ so that $\lim \mathcal{I}(u_n) = \inf_U \mathcal{I}$, ii) After taking a subsequence, $u_n \to_\tau u^*$, and iii) Using the lower semi-continuity we have that $\mathcal{I}(u^*) \leq \liminf \mathcal{I}(u_n)$, implying that u^* is a minimizer.

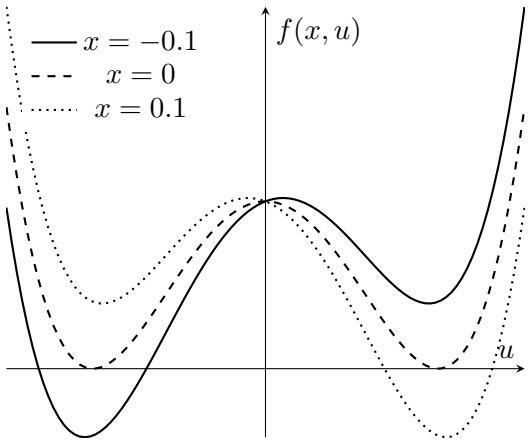
The main challenge in carrying out this approach is that if U is an infinite-dimensional normed space and B is some ball in that norm, then B can never be compact under the same norm. As such, one needs to select a weaker topology that allows compactness. The price to pay is that in weaker topologies continuity of $\mathcal{I}$ is a stronger condition to verify.

In this section, we will primarily focus on L^p type spaces, because for many notable examples we expect minimizers of our variational problem to fail to be continuous. To illustrate why this is the case, we begin with a toy problem demonstrating how non-convex functional optimization can have discontinuous minimizers.

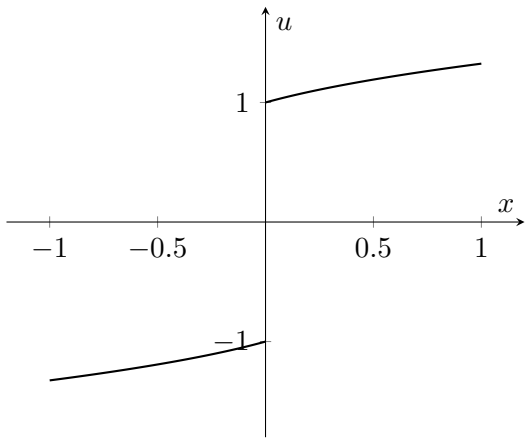
Example 29 (Double-well Potential) *Let* $f(x,u) = \frac{1}{4}(u^2-1)^2 - ux$ *and define the functional*

$$\mathcal{I}(u) = \int_{-1}^{1} f(x,u(x))dx. \tag{A.1}$$

In this simple case, one can directly show that the minimizer of this functional is given by $u^*(x) \in \arg\min f(x,\cdot)$ *for every* $x \in [-1,1]$. *We display this function in Figure 3b, and the discontinuity at* $x = 0$ *is apparent. This occurs because there are two distinct, well-separated, global minima at* $x = 0$. *We notice that for* $x \in \big[-\frac{1}{3^{3/2}} - \frac{1}{3^{1/2}}, \frac{1}{3^{3/2}} + \frac{1}{3^{1/2}},\big]$, *the function* $f(x,\cdot)$ *has 2 local minima, and that the global minima switches from one side to the other at* $x = 0$*: this is illustrated in Figure 3a.*



(a) Notice how as x passes through 0, the global minimizer of $f(x, \cdot)$ jumps between two values.



(b) As a consequence, the minimizer of $\mathcal{I}(u)$ is discontinuous.

Figure 3: Simple example of settings where minimizers of non-convex variational problems move discontinuously.

In the previous example we could immediately verify that u^* is a minimizer, by directly comparing its energy to that of any other function. However, if we did not know the form of u^* we would need to utilize the direct method to prove that a minimizer exists. For the sake of illustration, we will discuss this first in the context of the functional $\mathcal{I}$. As evidenced by the previous example, we need to minimize over a function space that permits discontinuities; we select $L^\infty([-1, 1]; \mathbb{R})$ for simplicity.

When minimizing (A.1) over the space of bounded functions, we notice that continuity of the energy with respect to the strong topology (i.e. the topology induced by the L^∞ norm) is nearly immediate, because

$$|\mathcal{I}(u_1) - \mathcal{I}(u_2)| \leq C \sup_{x \in [-1,1]} |u_1(x) - u_2(x)|.$$

However, the bounded sequences in L^∞ are far from being compact: take for example $\operatorname{sign}(\sin(nx))$ which has no convergent subsequence in L^∞. The standard approach is to weaken the notion of convergence on L^∞ to convergence in duality with L^1, i.e. weak-* convergence, which we denote by $\rightharpoonup^*$. More explicitly, we say that $u_n \rightharpoonup^*_p u \in L^p$ if for every $v \in L^{p^*}$ we have that

$$\int u_n(x)v(x)\,dx \to \int u(x)v(x)\,dx, \qquad \frac{1}{p} + \frac{1}{p^*} = 1.$$

We can directly check that $\operatorname{sign}(\sin(nx)) \rightharpoonup^*_\infty 0$, and indeed we can show that any bounded sequence in L^∞ is weak-* compact. However, the following example shows that upon moving to this topology the functional $\mathcal{I}$ is no longer lower semi-continuous.

Example 30 *Define the sequence $u_n(x) = \operatorname{sign}(\sin(n\pi x))$. As we have said, the sequence has no (strongly) convergent subsequence in $L^\infty([-1, 1]; \mathbb{R})$, but $u_n \rightharpoonup^*_\infty 0$. However, it can*

be checked directly that $\mathcal{I}(u_n) = (-1)^n/n \to 0$*, and that* $\mathcal{I}(0) = 1/2$*. Therefore* $\mathcal{I}$ *is not weakly-* lower semi-continuous. Another way of interpreting this example is to notice that* $f(x, u_n(x))$ *does not converge in the weak-* topology to* $f(x, 0)$*.*

We see in the previous example that the continuity of $\mathcal{I}$ with respect to L^∞ does not imply that it is weak-* lower semi-continuous: the following classical result links this phenomenon with convexity for general integral functionals. For a reference, see Theorems 6.54 & 6.56 in Fonseca and Leoni (2007).

Proposition 31 *Let* $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ *be a continuous function that is bounded below. For* $1 \leq p \leq \infty$ *define* $\mathcal{I} : L^p(\mathbb{R}^d; \mathbb{R}^m) \to \mathbb{R}$ *by*

$$\mathcal{I}(u) = \int f(x, u(x))dx,$$

then I *is weakly lower semi-continuous (weak-* if* $p = \infty$*) if and only if* $u \mapsto f(x, u)$ *is convex.*

References

Shreya Arya, Arnab Auddy, Ranthony Edmonds, Sunhyuk Lim, Facundo Memoli, and Daniel Packer. The Gromov-Wasserstein distance between spheres. *arXiv preprint*, 2024.

Antonio Auffinger and Daniel Fletcher. Equilibrium distributions for t-distributed stochastic neighbour embedding. *arXiv preprint*, 2023.

Robert Beinert, Cosmas Heiss, and Gabriele Steidl. On assignment problems related to Gromov–Wasserstein distances on the real line. *SIAM Journal on Imaging Sciences*, 16 (2):1028–1032, 2023.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

José C Bellido and Carlos Mora-Corral. Existence for nonlocal variational problems in peridynamics. *SIAM Journal on Mathematical Analysis*, 46(1):890–916, 2014.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Ranthony A. Clark, Tom Needham, and Thomas Weighill. Generalized dimension reduction using semi-relaxed Gromov-Wasserstein distance. *arXiv preprint*, 2024.

Trevor Cox and Michael Cox. *Multidimensional scaling*. Chapman and Hall, 2001.

Jan de Leeuw and Patrick Mair. Multidimensional scaling using majorization: Smacof in r. *Journal of Statistical Software, Articles*, 31(3):1–30, 2009.

Théo Dumont, Théo Lacombe, and François-Xavier Vialard. On the existence of Monge maps for the Gromov–Wasserstein problem. *Foundations of Computational Mathematics*, 02 2024.

Peter Elbau. Sequential lower semi-continuity of non-local functionals. *arXiv preprint*, 2011.

Irene Fonseca and Giovanni Leoni. *Modern Methods in the Calculus of Variations: L^p Spaces*. Springer, 01 2007.

Mikil D Foss, Petronela Radu, and Cory Wright. Existing and regularity of minimizers for nonlocal energy functionals. *Differential and Integral Equations*, 31(11-12):807–832, 2018.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Dmitry Kobak and George Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature Biotechnology*, 39:1–2, 02 2021.

Tjalling C. Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25(1):53–76, 1957.

J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Wonjun Lee, Riley C. W. O'Neill, Dongmian Zou, Jeff Calder, and Gilad Lerman. Geometry-preserving encoder/decoder in latent generative models, 2025.

Gongkai Li, Minh Tang, Nicolas Charon, and Carey Priebe. Central limit theorems for classical multidimensional scaling. *Electronic Journal of Statistics*, 14(1):2362 – 2394, 2020.

George C. Linderman, Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. Efficient algorithms for t-distributed stochastic neighborhood embedding. *arXiv preprint*, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*, 2018.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Facundo Mémoli and Tom Needham. Comparison results for Gromov-Wasserstein and Gromov-Monge distances. *arXiv preprint*, 2022.

Luca Nenna and Brendan Pass. Variational problems involving unequal dimensional optimal transport. *Journal de Mathématiques Pures et Appliquées*, 139:83–108, 2020.

Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(6):1771–1790, 2015.

Pablo Pedregal. Nonlocal variational principles. *Nonlinear Analysis: Theory, Methods & Applications*, 29(12):1379–1392, 1997.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

R.T. Rockafellar, M. Wets, and R.J.B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, 1969.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. In *Proceedings of Machine Learning Research*. PMLR, 2023.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

Roger N Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

Karl-Theodor Sturm. The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. *Memoirs of the American Mathematical Society*, 290, 08 2012.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Hugues Van Assel, Cédric Vincent-Cuaz, Nicolas Courty, Rémi Flamary, Pascal Frossard, and Titouan Vayer. Distributional reduction: Unifying dimensionality reduction and clustering with Gromov-Wasserstein. *arXiv preprint*, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

L.P.J. Van Der Maaten, E.O. Postma, and Jaap Van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

Titouan Vayer. *A contribution to Optimal Transport on incomparable spaces*. PhD thesis, Lorient, 2020.

Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced Gromov-Wasserstein. *Proceedings of Machine Learning Research*, 32, 2019.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov-Wasserstein divergence with applications on graphs. In *ICLR 2022-10th International Conference on Learning Representations*, pages 1–28, 2022.

Zhirong Yang, Jaakko Peltonen, and Samuel Kaski. Majorization-minimization for manifold embedding. In *Artificial Intelligence and Statistics*, pages 1088–1097. PMLR, 2015.