Conditional Density Estimation, Latent Variable Discovery and Optimal Transport

Hongkang Yang, Esteban G. Tabak

June 5, 2020

Abstract

A framework is proposed that addresses both conditional density estimation and latent variable discovery. The objective function maximizes explanation of variability in the data, achieved through the optimal transport barycenter generalized to a collection of conditional distributions indexed by a covariate—either given or latent—in any suitable space. Theoretical results establish the existence of barycenters, a minimax formulation of optimal transport maps, and a general characterization of variability via the optimal transport cost. This framework leads to a family of non-parametric neural network-based algorithms, the BaryNet, with a supervised version that estimates conditional distributions and an unsupervised version that assigns latent variables. The efficacy of BaryNets is demonstrated by tests on both artificial and real-world data sets. A parallel drawn between autoencoders and the barycenter framework leads to the Barycentric autoencoder algorithm (BAE).

Keywords: Unsupervised learning, optimal transport, neural network, autoencoders, factor discovery.

AMS Subject classification: 62H25, 62G07, 62M45, 49K30

1 Introduction

In machine learning, one often considers joint distributions of the form $\rho(x, z)$, where x is an observable and z some latent variable, or alternatively z is the source and x the target variable. For instance, in images of human faces, the data space $x \in X$ may have a dimension up to $10^5 \sim 10^6$ if counted in pixels, while a covariate $z \in Z$ consisting of the face orientation is only two-dimensional.

A task of broad applicability is to extract, given data drawn from $\rho(x, z)$, the conditional distributions $\rho(x|z)$. An example in medical studies has $\rho(x|z)$ representing the distribution of blood sugar level conditioned on a patient's age and diet. In generative modeling, $\rho(x|z)$ can represent the distribution of images conditioned on a text description such as z="cat", and one seeks to generate samples from $\rho(x|z)$. A knowledge of $\rho(x|z)$ allows one to estimate the conditional expectation $\mathbb{E}_{\rho(x|z)}[f(x)]$ for any function f of interest.

Alternatively, if one is only given the raw data $\rho(x)$, then one can try to infer a reasonable latent variable z that underlies x. For instance, for the facial images, discovering z as the face orientation explains away a great portion of the data's variability. Whether or not this latent variable has a clear interpretation, it can potentially facilitate data compression and generative modeling.

These two problems are known, respectively, as conditional density estimation and latent variable discovery. The former can be seen as a probabilistic generalization of classification and regression, while the later contains as special cases clustering and dimensionality reduction. They form a pair of supervised/unsupervised problems, such that one learns the dependency of x on z, while the other discovers z. This paper formulates and solves both problems in a single framework based on optimal transport.

1.1 Related work

Existing methods for conditional density estimation generally follow one of two approaches: to directly model $\rho(x, z)$ using kernel smoothing techniques [22, 20], or to model the mapping

$$z \mapsto \rho(x|z)$$

For instance, the Mixture Density Network [7] models $\rho(x|z)$ as a Gaussian mixture with z-dependent parameters, the Conditional GAN [33] models $\rho(x|z)$ by generative adversarial networks (GAN), Deep Conditional Generative Models [44] use variational autoencoders (VAE), and [2] and [48] use normalizing flows. Essentially, these methods apply density estimation techniques for a single distribution to the modeling of all conditional distributions simultaneously. We will introduce an alternative approach such that all $\rho(x|z)$ are represented by a single distribution μ, from which we can easily recover each $\rho(x|z)$, so that we only need to estimate μ.

Existing methods for latent variable discovery are vast and rich. For discrete latent variables z, the problem reduces to clustering, where popular methods include k-means and the EM algorithm [4, 8]. For continuous z, we have dimensionality reduction algorithms such as principal component analysis (PCA), principal curves and surfaces [18], and undercomplete autoencoders (also known as autoassociative neural networks) [4]. Depending on different regularizations on z, there are also the VAE [26], AAE [31], WAE [47], and denoising and sparse autoencoders [4]. We will identify below a parallelism between autoencoders and the algorithms that we propose.

Our theoretical model is based on optimal transport, in particular on the barycenter of probability measures. The idea of applying barycenters to conditional density estimation originates from [46], while the application to latent variable discovery is based on the previous work in [45, 53]. This paper lays the theoretical foundation for the technique of barycenters, and introduces several neural network-based algorithms.

1.2 Sketch of our approach

Intuitively, one of the principles of learning is to reduce uncertainty. Given arbitrary data, an effective way to learn it is to find a representation of it so that some measure of uncertainty is reduced. One prototypical example is the Kolmogorov complexity (or descriptive complexity) [43]: when the data consists of a string such as *ababababab*, it is natural to represent it by $ab \times 5$, so the variability of a long string is reduced to that of a shorter representative. Another instance is PCA, which seeks a low-dimensional representation of high-dimensional data. It maximizes the amount of variance explained by the principal components, thus minimizing the uncertainty remaining.

Clustering provides a similar setting: suppose that we are given the data displayed on the left image of Figure 1, divided into three labeled clusters. We would naturally *learn* the data by memorizing the clusters' common shape and their relative positions. Equivalently, as in the right image, the data can be represented by a common distribution plus the translations that bring the three clusters to it. As we apply these translations to transform the original data, the variance is greatly reduced.

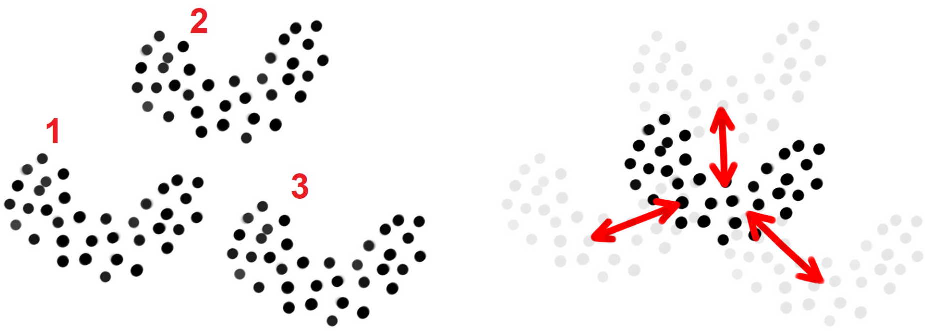

Figure 1: Clusters and their representative

This intuition can be summarized as follows: given a sample space X such as $\mathbb{R}^2$ and a latent variable space Z such as $\{1, 2, 3\}$, an effective way to learn a distribution $\rho(x, z) \in P(X \times Z)$ is to find a representative distribution μ with smaller variability, as well as the transformations between each conditional distribution $\rho(x|z)$ and μ.

In a data-based scenario, we are given a labeled sample $\{x_i, z_i\}_{i=1}^N$. Once we obtain the transformations T_z that send $\rho(x|z)$ to μ and their inverse transformations S_z, the representative μ can be estimated by aggregating all sample points: $\{T_{z_i}(x_i)\}$. Then, given any z, the conditional distribution $\rho(x|z)$ can be sampled via $\{S_z \circ T_{z_i}(x_i)\}$.

We can already see one advantage of this procedure: once the transformations are known, N sample points for $\rho(x, z)$ automatically provide N samples for each conditional $\rho(x|z)$. This is particularly helpful when there are many latent variables z or when these are continuous, so that for most values of z, the conditional $\rho(x|z)$ has very few or zero samples in $\{x_i, z_i\}_{i=1}^N$. Furthermore, this procedure can be used in conjunction with other density estimation algorithms or generative models: the difficult task of modeling many $\rho(x|z)$ (or a high-dimensional $\rho(x, z)$) is simplified into modeling a single μ, and one can, for instance, first train the GAN or VAE on μ and then concatenate it with S_z to model each $\rho(x|z)$.

The theory of optimal transportation is ideal for the formalization of the procedure above. Intuitively, the representative distribution μ should closely resemble each conditional $\rho(x|z)$, that is, μ is the minimizer of some average "distance" between μ and $\rho(x, z) = \rho(x|z)\nu(z)$:

$$\mu = \text{argmin}_{\tilde{\mu}} \int_Z \text{distance}\big(\tilde{\mu}, \rho(x|z)\big) d\nu(z)$$

Thus, we refer to μ as the "barycenter" of $\rho(x, z)$. The optimal transport cost (or "Earth mover's distance") is a good candidate for distance, such that informally the distance between two distributions μ_1 and μ_2 is the minimum "work" required to transport μ_1 (thought of as a pile of sand) to μ_2. When a cost function $c(x, y)$ is given (such as the Euclidean distance $\|x - y\|$), the optimal transport cost is

$$I_c(\mu_1, \mu_2) \approx \inf_T \int c(x, T(z)) d\mu_1,$$

where the infimum is taken over all maps T that transport μ_1 to μ_2. If we consider $c(x, T(x))$ as a measure of pointwise distortion, then $I_c(\rho(x|z), \mu)$ is the distortion or information loss incurred on the original data.

Optimal transportation has several advantages. The optimal transport cost $I_c(\mu_1, \mu_2)$ depends on a user-specified cost c and thus can directly incorporate task-specific information. In particular, if the cost is based on the metric of the space, then $I_c(\mu_1, \mu_2)$ reliably captures our intuitive sense of distance between distributions [5], whereas other measures, such as the Kullback-Leibler divergence and total variation, fail when the distributions have disjoint supports. Also, given that optimal transport minimizes data distortion, it is natural to expect that μ_1 can be easily recovered from its transported image μ_2, that is, the transport maps T_z are invertible. This is a useful property, since our procedure needs to transform back and forth between the conditionals $\rho(x|z)$ and the representative μ.

The greatest advantage, however, is duality. The theory of optimal transport abounds with duality techniques, through which we convert optimization problems over probability measures to problems over functions, and vice versa. A general rule is that, being less restricted, functions are easier to model and optimize than probability distributions, and we perform this conversion whenever possible. The primal problem of conditional density estimation is often intractable, because it is difficult to model directly each of the possibly infinitely many conditionals $\rho(x|z)$. Yet, optimal transport duality converts the primal into an optimization over one transport map $T(x, z)$ and one discriminator $\psi(y, z)$, which can be more easily solved by methods such as neural networks.

The next step is to apply our principle of minimum uncertainty to latent factor discovery, the unsupervised counterpart to conditional density estimation. Recall that in the supervised setting with $\rho(x, z)$, our procedure reduces the uncertainty or "variability" of $\rho(x)$ to the smaller variability of the representative (or barycenter) μ. If one has the freedom to determine the labels z, then the variability of μ can be further reduced. Specifically, what matters is the choice of $\rho(x|z)$, whereas the labels z by themselves are equivalent under permutations and we can arrange them into some prior distribution $v(z)$.

Thus, given an unlabeled data $\rho(x)$, factor discovery should seek a labeling $\rho(x,z)$ that minimizes the variability of its barycenter μ, or equivalently,

$$\max_{\rho(x,z)} \text{Variability}(\rho(x)) - \text{Variability}(\mu).$$

If, for the dataset in Figure 1, one did not know the labels $\{1,2,3\}$, one could assign them. Clearly some labelings are better than others: in the worst scenario, the labels $\{1,2,3\}$ would be distributed uniformly within each cluster, and our procedure would yield a barycenter μ with the same shape and size as the original data, with no variability reduction at all.

We should define "variability" in a way that generalizes variance, so that factor discovery can yield the obvious labeling of Figure 1. Also, variability should depend on the cost $c(x,y)$ in order to incorporate task-specific information. Intuitively, how much we learn is proportional to how much effort we spend learning, or equivalently,

$$\text{Reduced uncertainty} = \text{Work}.$$

So we characterize "variability" as a measurement that satisfies

$$\text{Variability}(\rho(x)) - \text{Variability}(\mu) = \text{Total transport cost } \int I_c\big(\rho(x|z),\mu\big)d\nu(z). \tag{1}$$

In fact, Corollary 4 below shows that definition (1) yields exactly the variance when we use the squared Euclidean distance cost $c = \|x-y\|^2$. Hence, factor discovery becomes

$$\max_{\rho(x,z)} \text{Total transport cost},$$

which differs from conditional density estimation only by the additional maximization.

This paper is structured as follows. Section 2 develops the ideas presented in the introduction, formulating conditional density estimation and latent factor discovery in the framework of optimal transport barycenters. Section 3 addresses the algorithmic aspects, proposing the supervised and unsupervised BaryNet algorithms, which use neural networks. It also discusses BaryNet's relation to existing methods, in particular the autoencoders, and introduces the Barycentric autoencoder (BAE) based on BaryNet. Section 4 tests the performance of the BaryNet algorithms on real-world and artificial data sets, and verifies that they can reliably solve conditional density estimation and latent factor discovery. Finally, Section 5 summarizes the results and discusses possible future work. The proofs of most theorems are provided in an appendix.

2 Theoretical foundation

The ideas presented in the introduction are formalized and proved in this section. We first define optimal transport barycenter and prove its existence. Then, we obtain the conditional transport map $T(x,z)$ from a minimax problem. Finally, we prove the variance decomposition theorem and justify our definitions of variability and latent factor discovery.

2.1 Preliminaries

We denote by X and Z the sample and latent variable spaces, and by Y the space that the barycenter μ belongs to. In practice one often has $X = Y$, but this is not required here.

Most of our results will be presented with (X,Y,Z) Polish spaces, which are complete separable metric spaces. These have enough structure to handle problems of optimal transport, while they are general enough to include most spaces in real-world applications, such as Euclidean spaces $\mathbb{R}^d$, closed subsets of $\mathbb{R}^d$, complete Riemannian manifolds M^d, discrete sets such as $\{1,\dots K\}$, and function spaces such as $C([0,1]), P(\mathbb{R}^d)$.

Given a Polish space X, we denote the space of continuous functions by $C(X)$, the space of bounded continuous functions by $C_b(X)$ and the space of Borel probability measures by $P(X)$.

For clarity, we sometimes write a measure $\rho \in P(X)$ informally as $\rho(x)$ to indicate the space it belongs to, not implying by this that ρ has a density function, unless explicitly declared. For joint probability measures, e.g. $\pi \in P(X \times Y \times Z)$, we denote its marginals by π_X, π_{YZ}, etc. The tensor product of probability measures μ and ν is denoted by $\mu \otimes \nu$.

Given $\rho(x,z) \in P(X \times Z)$, we define the conditional distributions $\rho(x|z)\nu(z) = \rho(x,z)$ using the disintegration theorems [10]. The conditional $\rho(x|z)$ always exists as a Borel measurable map from Z to $P(X)$ in the topology of weak convergence, and it is unique $\nu(z)$-almost surely. Conversely, given $\nu(z)$ and a measurable $\rho(x|z)$, we define the joint distribution $\rho(x,z) := \rho(x|z)\nu(z)$ by

$$\forall \psi \in C_b(X \times Z), \quad \int \psi d\rho(x,z) := \int \psi d\rho(x|z) d\nu(z)$$

2.2 Optimal transport and barycenter

A map $T : X \to Y$ pushes-forward $\rho \in P(X)$ to $\mu \in P(Y)$ (denoted $T\#\rho = \mu$) if

$$\mu(A) = \rho(T^{-1}(A))$$

for all measurable subsets $A \subseteq X$. Monge's original formulation of optimal transport [34]:

$$\inf_{T\#\rho=\mu} \int_X c(x, T(x)) d\rho(x)$$

minimizes over all transport maps T the expected value of a cost function c on $X \times Y$. Kantorovich [24] generalized the transport maps to probabilistic couplings,

$$\Pi(\rho,\mu) := \{\pi \in P(X \times Y),\ \pi_X = \rho, \pi_Y = \mu\},$$

relaxing the optimal transport problem to

$$I_c(\rho,\mu) = \inf_{\pi \in \Pi(\rho,\mu)} \int_{X \times Y} c(x,y) d\pi(x,y). \tag{2}$$

If the minimum of (2) is achieved by some coupling π, we call it an optimal transport plan, or a Kantorovich solution. If π is concentrated on the graph of some function $T : X \to Y$, then T is called an optimal transport map, or a Monge solution.

Inspired by definitions from [46] and [25], we define optimal transport barycenter as follows:

Definition 1 (Barycenter problem)**.** Given a cost function $c(x,y)$, a labeled distribution $\rho(x,z) = \rho(x|z)\nu(z) \in P(X \times Z)$, and any $\mu \in P(Y)$, the *total transport cost* between $\rho(x,z)$ and μ is defined by

$$I_c(\rho(x,z),\mu) = \int_Z I_c(\rho(x|z),\mu) d\nu(z) \tag{3}$$

If the minimum total transport cost

$$\inf_{\mu \in P(Y)} I_c(\rho(x,z),\mu)$$

is achieved by some μ, then we call it the *barycenter* of $\rho(x,z)$.

The notion of a barycenter of finitely many conditionals $\rho(x|z)$ (that is, with finite label space $Z = \{1,\dots K\}$) was introduced in [9, 11, 39], and its existence, uniqueness, and regularity were examined in [3, 37, 25]. Barycenters of infinitely many conditional distributions are studied in [38, 25], which deal with the special case when X is either a Euclidean space or a compact Riemannian manifold and c is the squared distance cost.

We show that the barycenter problem as defined above is well-posed, and the barycenter exists under general conditions:

Theorem 1 (Well-posedness and Existence of Barycenter)**.** Let X, Y, Z be Polish spaces, let $c \in C(X \times Y)$ be a continuous cost that is bounded below ($\inf c > -\infty$), and let $\rho(x, z) = \rho(x|z)\nu(z) \in P(X \times Z)$ be a probability measure. Then,

1. Given any $\mu \in P(Y)$, the total transport cost (3) is well-defined, and

$$\int_Z I_c(\rho(\cdot|z), \mu) d\nu(z) = \min_{\substack{\pi \in P(X \times Y \times Z) \\ \pi_{XZ} = \rho \\ \pi_{YZ} = \mu \otimes v}} \int_{X \times Y \times Z} c(x, y) d\pi(x, y, z), \tag{4}$$

so there exists a Kantorovich solution in the form $\pi \in P(X \times Y \times Z)$.

2. If Assumption 1 from Appendix A holds, then there exists a barycenter $\mu \in P(Y)$. Specifically,

$$\min_{\mu \in P(Y)} \int_Z I_c(\rho(\cdot|z), \mu) d\nu(z) = \min_{\substack{\pi \in P(X \times Y \times Z) \\ \pi_{XZ} = \rho \\ \pi_{YZ} = \pi_Y \otimes \pi_Z}} \int_{X \times Y \times Z} c(x, y) d\pi(x, y, z), \tag{5}$$

and the marginal π_Y of every solution π is a barycenter.

Proof. See Appendix B. Note that $\pi_{YZ} = \pi_Y \otimes \pi_Z$ implies that the barycenter is independent of the latent variable. □

Remark 1. There are pathological examples where the barycenter does not exist: if $X = Y = \mathbb{R}^d$ and $c(x, y) = \exp[-\|x - y\|^2]$, then any barycenter will tend to be pushed arbitrarily far away. Assumption 1 is modeled after the squared distance cost $c = \|x - y\|^2$ and prevents such degeneracy.

2.3 Conditional transport maps

Having shown that the barycenter μ exists, the next step is to find the transport maps T_z and inverse transport maps S_z between each conditional distribution $\rho(x|z)$ and $\mu(y)$.

From Theorem 1, the barycenter problem admits a Kantorovich solution $\pi \in P(X \times Y \times Z)$. If π should also be a Monge solution, that is, π were concentrated on the graph of some transport map $T : X \times Z \to Y$, it would follow that for v-almost all z,

$$T(\cdot, z)\#\rho(x|z) = \mu(y)$$

or equivalently,

$$\tilde{T}\#\rho(x, z) = \mu(y) \otimes \nu(z), \text{ where } \tilde{T}(x, z) := (T(x, z), z) \tag{6}$$

We show that this holds in general:

Theorem 2. Given Polish spaces X, Y, Z, probability $\rho(x, z) \in P(X \times Z)$ and cost $c \in C(X \times Y)$ that is bounded below ($\inf c > -\infty$), under Assumptions 1 and 2 from Appendix A, the barycenter problem has a Monge solution: the minimum total transport cost (5) becomes

$$\min_{\substack{\text{Borel measurable} \\ T : X \times Z \to Y}} \sup_{\substack{\psi_Y(y) \in C_b(Y) \\ \psi_Z(z) \in C_b(Z) \\ \int \psi_Z(z) d\nu(z) = 0}} \int \left[c(x, T(x, z)) - \psi_Y(T(x, z))\psi_Z(z)\right] d\rho(x, z) \tag{7}$$

and every minimizer T is a transport map from $\rho(x, z)$ to a barycenter $\mu(y)$.

Proof. See Appendix C. □

Remark 2. The test function $\psi_Y(y)\psi_Z(z)$ in (7) serves as the "discriminator" that checks that all the conditional distributions $\rho(x|z)$ have been pushed-forward to the same barycenter μ. The technique of discriminator has appeared in [16, 5] to train the generative adversarial networks, and it has been applied to the barycenter problem by [46], which derived test functions of the form

$$\psi(y, z) \text{ such that } \int \psi(y, z) d\nu(z) \equiv 0$$

Theorem 2 improves this technique, because $\psi_Y(y)\psi_Z(z)$ has much less complexity than $\psi(y,z)$. From a data-based perspective, with the distributions given through sample points $\{x_i,y_i,z_i\}_{i=1}^N$, the test function $\psi(y,z)$ becomes a full $N\times N$ matrix, whereas $\psi_Y(y),\psi_Z(z)$ are two $1\times N$ vectors, which can be seen as providing a rank-one factorization of $\psi(y,z)$. Later sections show that all computations are thereby reduced from quadratic to linear time $O(N)$.

An explanation for this improvement is that the barycenter problem has more freedom than the ordinary optimal transport problem. Optimal transport would require the pushforward $\tilde{T}\#\rho(x,z)$ to match a fixed target distribution, so that the dual problem needs to mobilize the entire $C_b(Y\times Z)$ to pin it down. For the barycenter problem, however, Theorem 1 shows that $\pi_{YZ}=\tilde{T}\#\rho(x,z)$ only needs to satisfy the independence condition

$$\pi_{YZ}=\pi_Y\otimes\pi_Z.$$

Correspondingly, the dual problem only requires a small subspace of $C_b(Y\times Z)$.

Regarding the inverse transport maps S_z, one approach is to set $S_z=T_z^{-1}$. This is viable in many scenarios: for instance, by Brennier's Theorem [50], the inverse function T_z^{-1} exists almost surely,

$$T_z^{-1}\circ T_z(x)=x \text{ for } \rho(x|z)\text{-almost all } x,$$

and it is the optimal transport map for the inverse transport,

$$T_z^{-1}\#\mu=\rho(x|z).$$

Then, computing S_z becomes a simple regression problem. Given the labeled data $\{x_i,z_i\}_{i=1}^N$, we first compute the barycenter $\{y_i=T(x_i,z_i)\}$ and then find a map $S:Y\times Z\to X$ that approximates $\{x_i\}$ by $\{y_i,z_i\}$. This is the approach used by our algorithms.

It might be helpful to note that there is a more general approach, which directly solves the optimal transport from $\mu(y)\otimes\nu(z)$ back to $\rho(x,z)$. Assertion 1 of Theorem 1 shows that there is always a Kantorovich solution, while the arguments of Theorem 2 can be applied to show that the inverse transport map $S(y,z)$ can be solved from

$$\min_{\substack{\text{Borel measurable}\\ S:Y\times Z\to X}}\ \sup_{\psi(x,z)\in C_b(X\times Z)}\int\big[c(S(y,z),y)-\psi(S(y,z),z)\big]d\mu(y)d\nu(z)+\int\psi(x,z)d\rho(x,z).$$

Remark 3. As discussed in the introduction, given a labeled sample $\{x_i,z_i\}_{i=1}^N$, we can estimate each conditional distribution $\rho(x|z)$ by the computed sample $\{S_z\circ T_{z_i}(x_i)\}_{i=1}^N$. Yet, when $X=Y$ is Euclidean and $T_z(x)$ is differentiable (e.g. when modeled by a neural net), we can also derive the density function of $\rho(x|z)$: First, estimate the barycenter's density $\mu(y)$ from the computed sample $\{y_i\}_{i=1}^N$ (e.g. by kernel smoothing). Then, estimate the density through

$$\rho(x|z)=|J(T_z(x))|\ \mu(T_z(x)),$$

where $|J|$ is the Jacobian determinant. Then, we can estimate the density function $\rho(x,z)$ through $\rho(x|z)\nu(z)$, where $\nu(z)$ is estimated from the $\{z_i\}$.

2.4 Latent factor discovery

Finally, we justify the definition (1) of variability, from which it follows that the minimization of the barycenter's variability, which is the objective of factor discovery, is equivalent to the maximization of total transport cost. We illustrate this intuition in the case where $X=Y=\mathbb{R}^d$ and $c(x,y)=\|x-y\|^2$. Then, the optimal transport cost I_c becomes W_2^2, where W_2 is the 2-Wasserstein distance. (See [50, 51] for a reference of Wasserstein distance, and [3] for Wasserstein barycenters.)

Theorem 3. Given any measurable space Z and probability measure $\rho(x,z)=\rho(x|z)\nu(z)\in P(\mathbb{R}^d\times Z)$, there exists a Wasserstein barycenter $\mu\in P(\mathbb{R}^d)$ that satisfies

$$Var(\rho(x))-Var(\mu)=\int_Z W_2^2(\rho(x|z),\mu)d\nu(z). \tag{8}$$

Proof. See Appendix D for the proof. To illustrate the proof's intuition, consider the trivial case with Dirac masses:

$$\rho(x,z) = \sum_{k=1}^{K} P_k \delta_{x_k} \otimes \delta_{z_k}$$

where P_k are positive weights. Then, $\rho(x)$ becomes $\sum P_k \delta_{x_k}$, and the unique W_2-barycenter μ is the Dirac mass on the mean of $\rho(x)$. Both sides of (8) reduces to $Var(\rho)$. The rest would be an approximation argument that goes from Dirac masses to general probabilities, and the high-level idea is that the geometric properties of $\mathbb{R}^d$ can be lifted to $(P(\mathbb{R}^d), W_2)$. One result that we apply repeatedly comes from Proposition 3.8 and Remark 3.9 in [3]: Under technical conditions, given weights P_k and conditionals ρ_k, their W_2-barycenter μ, and optimal transport maps T_k from μ back to ρ_k, we have the identity

$$\sum_{k=1}^{K} P_k T_k = Id,\ \mu - a.e.$$

This identity indicates that the W_2-barycenter is exactly the convex sum of the conditionals, just as in the simple setting with Dirac masses illustrated above. It is worth mentioning that this identity can be generalized to compact Riemannian manifolds [25, Theorem 4.4]. □

Corollary 4. Let $V : P(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be any function such that $V(\delta_x) = 0$ for any Dirac mass δ_x. Then, V is the variance Var if and only if for any measurable space Z and any $\rho(x,z) \in P(\mathbb{R}^d \times Z)$, there exists a barycenter μ that satisfies

$$V(\rho(x)) - V(\mu) = \int_Z W_2^2(\rho(x|z), \mu) d\nu(z) \tag{9}$$

Proof. The "only if" part follows from Theorem 3. For the "if" part, set $Z = \mathbb{R}^d$. Given any $\rho(x) \in P(\mathbb{R}^d)$, set $\rho(x,z) = \delta_z(x)\rho(z)$. Then, the X-marginal of $\rho(x,z)$ is $\rho(x)$, and the unique barycenter μ is the Dirac measure at the mean of $\rho(x)$. Then, (9) reduces to $V(\rho(x)) = Var(\rho(x))$. □

Since a Dirac measure δ_x represents a deterministic event without any uncertainty, $V(\delta_x)$ should be zero for any reasonable variability function V. Then, it follows from Corollary 4 that the variability defined by (1) is exactly the variance Var, when the cost is the Euclidean squared distance.

Remark 4. As a further justification, notice that if the cost $c(x,y) = \|x-y\|^2$ is generalized to $(x-y)^T Q (x-y)$ for some positive-definite symmetric matrix Q, then the corresponding variability becomes a "weighted" variance, with a different scaling factor in each eigenspace of Q:

$$V(\rho) = \int (x - \overline{x})^T Q (x - \overline{x}) d\rho(x) = Var(\sqrt{Q}\#\rho)$$

where $\overline{x}$ is the mean and $\sqrt{Q}\#\rho$ is the pushforward by the linear map $\sqrt{Q}$.

Proof. Given any $\rho(x,z)$, formula (5) becomes

$$\begin{aligned}\min_{\substack{\pi \in P(\mathbb{R}^d \times \mathbb{R}^d \times Z) \\ \pi_{XZ} = \rho(x,z) \\ \pi_{YZ} = \pi_Y \otimes \pi_Z}} \int (x-y)^T Q (x-y) d\pi(x,y,z) &= \min_{\substack{\pi \in P(\mathbb{R}^d \times \mathbb{R}^d \times Z) \\ \pi_{XZ} = \rho(x,z) \\ \pi_{YZ} = \pi_Y \otimes \pi_Z}} \int \|x-y\|^2 d(\sqrt{Q}, \sqrt{Q}, Id)\#\pi(x,y,z) \\ &= Var(\sqrt{Q}\#\rho(x)) - Var(\sqrt{Q}\#\mu),\end{aligned}$$

where $\sqrt{Q}\#\mu$ is a barycenter of $(\sqrt{Q}, Id)\#\rho(x,z)$ (under cost $\|x-y\|^2$) that satisfies (8). Then, μ is a barycenter of $\rho(x,z)$ (under cost $(x-y)^T Q (x-y)$) and satisfies (9) with $V(\rho) = Var(\sqrt{Q}\#\rho)$. □

It follows from definition (1) that the variability of the barycenter is complementary to the total transport cost (7) to the barycenter. As argued in the introduction, given any unlabeled data $\rho(x)$, latent factor discovery looks for a labeling $\rho(x,z)$ that minimizes the variability of its barycenter. Then, factor discovery has the following equivalent formulation based on (7):

$$\sup_{\substack{\rho(x,z)\\ \rho_X=\rho(x)}} \min_{\substack{\text{measurable}\\ T:X\times Z\to Y}} \sup_{\substack{\psi_Y(y)\in C_b(Y)\\ \psi_Z(z)\in C_b(Z)\\ \int \psi_Z(z)d\nu(z)=0}} \int \big[c(x,T(x,z)) - \psi_Y(T(x,z))\psi_Z(z)\big] d\rho(x,z). \tag{10}$$

To solve (10), factor $\rho(x,z)$ into $p(z|x)\rho(x)$, where $p(z|x)$ is the conditional label distribution for the sample point x. In particular, when Z is finite, $p(z|x)$ can be seen as a classifier on X. Thus, an effective way to optimize $\rho(x,z)$ and regularize the solution is to parameterize $p(z|x)$: e.g. we can set

$$p_\theta(z|x) = G_\theta(x,\cdot)\#\mathcal{N}(0,I) \tag{11}$$

where G_θ is a neural net with two inputs and $\mathcal{N}(0,I)$ is a unit normal distribution.

Something to be aware of is the $\sup_\rho$ in problem (10), which could lead to degenerate solutions if there is no bound on the "expressivity" of $p(z|x)$. Its behavior is well-controlled when Z is finite and $\rho(x,z)$ is simply a clustering plan of $\rho(x)$. However, theoretical issues could arise when Z is a larger space such as $[0,1]$: we can always find a measurable map $f : [0,1] \to X$ that transports the uniform distribution $U[0,1]$ onto $\rho(x)$ (e.g. via Theorem 1.1 of [23]), and then we can set

$$v(z) := U[0,1],\ \rho(x,z) := \delta_{f(z)}(x)\ v(z) \tag{12}$$

This $\rho(x,z)$ has the right marginal $\rho_X = \rho(x)$, but since every conditional $\rho(x|z)$ is a Dirac mass, the barycenter is also a Dirac mass. So problem (10) leads to a trivial solution compressing all data to a single point.

This situation is analogous to "overfitting" in regression problems, when one intends to learn a rough sketch of the data but the algorithm learns all the fine details instead. Fortunately, such an issue can be avoided in practice by controlling how much the algorithm learns during training. Suppose we adopt an implementation similar to (11). A desirable solution $p(z|x)$ should capture the overall shape of the data $\rho(x)$ but should not become as singular as the solution $p(z|x) = \delta_{f^{-1}(x)}(z)$ from (12). The key is to control the complexity of the function G_θ, so that G_θ does not become as complex as, for instance, the $f(z)$ in (12). This can be achieved by either explicit or implicit regularizations. Specifically, the complexity of G_θ can be characterized by appropriate functional norms such as the Barron norm [13], the number of parameters, or the Fourier spectrum [52]. Explicit regularizations include penalizing the norm [12], bounding the number of parameters, and early stopping [27]. Implicit regularizations include training dynamics that protect against overfitting [1] and the frequency principle [41] that prioritizes learning low frequencies. Even though these regularity results were obtained in settings different from problem (10), they are in a sense universal in neural network training, and hence applicable to our setting. Indeed, none of our experimental results in Section 4 exhibit degenerate solutions.

Besides the unlimited complexity of $p(z|x)$, there is another, rather trivial way for degenerate solutions to arise in problem (10). If X and Z are Euclidean and Z has higher dimension than X (or more generally, $X \subseteq Z$), then we can set

$$p(z|x) = \delta_z(x),\ \rho(x,z) = \delta_x(z)\rho(x)$$

This kind of overfitting is analogous to an autoencoder whose hidden layers have the same size as the input/output layers, so that the network can become the identity function and learn the trivial latent variable $z = x$. We will discuss more about the connection between problem (10) and autoencoders in Sections 3.1.2 and 3.2.

3 Algorithmic design

In the previous section, we converted conditional density estimation, a problem involving probability distributions, to the dual of the barycenter problem (7), which involves only functions. Then, latent factor discovery becomes (10) with an additional maximization over all labelings $\rho(x,z)$ whose marginal ρ_X is the given unlabeled distribution $\rho(x)$.

In practice, we are given a labeled finite sample set $\{x_i, z_i\}_{i=1}^N$ for conditional density estimation, and the objective (7) becomes

$$\inf_\tau \sup_\xi L(\tau, \xi) = \frac{1}{N} \sum_{i=1}^N \left[c\big(x_i, T_\tau(x_i, z_i)\big) - \psi_\xi^Y \big(T_\tau(x_i, z_i)\big) \tilde{\psi}_\xi^Z(z_i) \right], \tag{13}$$

where $T_\tau, \psi_\xi^Y, \psi_\xi^Z$ are maps parameterized by τ, ξ and

$$\tilde{\psi}_\xi^Z(z) := \psi_\xi^Z(z) - \frac{1}{N} \sum_{i=1}^N \psi_\xi^Z(z_i), \tag{14}$$

which is a sample-based version of the constraint $\int \psi^Z dv = 0$ from (7).

For factor discovery, we follow the analysis in Section 2.4 to model the labelings $\rho(x, z)$ via $p_\theta(z|x)\rho(x)$, where $p_\theta(z|x)$ is a parameterized conditional label distribution. Then the objective (10) becomes

$$\begin{aligned} \sup_\theta \inf_\tau \sup_\xi L(\theta, \tau, \xi) &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{p_\theta(z|x_i)} \left[c\big(x_i, T_\tau(x_i, z)\big) - \psi_\xi^Y \big(T_\tau(x_i, z)\big) \tilde{\psi}_\xi^Z(z) \right] \\ \tilde{\psi}^Z(z) &:= \psi^Z(z) - \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{p_\theta(\tilde{z}|x_i)} \psi^Z(\tilde{z}). \end{aligned} \tag{15}$$

It could be difficult to compute the expectation $\mathbb{E}_{p_\theta(z|x_i)}$ directly, unless it has an analytical solution or Z is finite. For simplicity, we often restrict to the case $p_\theta(z|x) = \delta_{z_\theta(x)}$, that is, the labeling is given by a deterministic map, $z_i = z_\theta(x_i)$.

In the following sections, we demonstrate the efficacy of (13) and (15) by implementing them through neural networks. We focus on the special case when X, Y are Euclidean spaces, Z is either Euclidean or finite, and the cost c is differentiable.

3.1 BaryNet

Since (13) and the deterministic version of (15) are optimization problems that involve only functions, it is natural to solve them by neural networks. By Theorems 1 and 2 of [21], feedforward neural nets are universal approximators for continuous functions $C(\mathbb{R}^d)$ and measurable functions $L^1(d\rho)$, so they can model the continuous test function $\psi^Y(y)$ (and $\psi^Z(z)$ when Z is Euclidean) and the measurable transport map $T(x, z)$. We can also model the conditional latent distribution $p_\theta(z|x)$ or the deterministic $z_\theta(x)$ by neural nets, if we require that they depend continuously on their parameters. Hence, (13) and (15) become a collection of interacting networks, an architecture that we refer to as "BaryNet", for barycenter network. As (13) and (15) can be seen as a supervised/unsupervised pair, we call the corresponding networks the supervised/unsupervised BaryNet.

One advantage of neural nets is the ease to control their expressivity. A neural net can approximate any continuous function if either its width [21] or depth [30] goes to infinity, so we can adjust the network's size, or more generally its functional norm [13], to solve problems with varying complexity. In factor discovery, for instance, if we know a priori that the ideal labeling z_θ should approximate the data's principal components, or if we desire simple labelings that are more interpretable, then we can reduce the size of z_θ or penalize its norm.

Another advantage is that the structure of the solution can be easily encoded in the network architecture. For instance, if the ideal solution should be a perturbation to the identity: $f(x) = x + o(|x|)$, then we can model only the perturbation part: $f_\theta(x) = x + g_\theta(x)$. This approach, known as "residual network" [19], makes the network easier to optimize and increases the likelihood to reach optimal solutions. It turns out that this residual design resembles the structure of the transport map $T(x, z)$.

3.1.1 Transport and inverse transport nets

As in residual networks, our transport map can be modeled as

$$T_\tau(x,z) = x + R_\tau(x,z), \tag{16}$$

if the transportation takes place within a single space, $X = Y = \mathbb{R}^d$. A motivation is that solutions to the barycenter problem (7) generally have the following properties:

1. Each transport map $T_z \# \rho(x|z) = \mu$ starts from an identity component x. This holds in general for optimal transport maps in Euclidean spaces, as these are special cases of the transport maps on complete Riemannian manifolds:

   $$T(x) = \exp_x(R(x))$$

   where $R(x)$ is a tangent vector that "points" to the transportation's destination (e.g. see McCann's theorem, Theorem 2.47 of [50]), which in the Euclidean setting reduces to $T(x) = x + R(x)$.

   As a concrete example, Theorem 2.44 of [50] shows that if the cost $c = c(x-y)$ is strictly convex and superlinear, and if the source and target measures are absolutely continuous, then the optimal transport map has the residual form

   $$T(x) = x - \nabla c^*(\nabla \phi(x)),$$

   where c^* is the Legendre transform of c and ϕ is c-concave. Another example is provided in Section 3.3 of [45]: if $\rho(x|z)$ and μ have similar shapes, then the transport map have the form

   $$T(x) \approx x + \beta(z),\ \ \beta(z) = \overline{y} - \overline{x}(z), \tag{17}$$

   where $\overline{x}(z), \overline{y}$ are the means of $\rho(x|z)$ and μ. This $T(x)$ approximates the optimal transport map up to the first moment.

2. Each T_z is invertible: it was argued in Section 2.3 that under general conditions, such as under the hypothesis of Brennier's theorem [50], the transport map T_z is invertible $\rho(x|z)$-almost surely and its inverse T_z^{-1} transports μ back to $\rho(x|z)$. The residual design (16) is an effective way to ensure that T_z is invertible: if the residual term is small, in the sense that $\nabla_x R \approx O$, then the inverse exists locally by the inverse function theorem, and it has the form

   $$S_z(x) = x - R(x,z) + O(\|\nabla_x R(x,z)\| \cdot \|R(x,z)\|). \tag{18}$$

An additional benefit of the residual design is that the Jacobian matrix $\nabla_x T_\tau$ is close to the identity when the residual term is small, thus alleviating the exploding and vanishing gradient problem during training [19].

Regarding the inverse transport map $S(y,z)$, formula (18) suggests that we should also model $S(y,z)$ as a residual network with the same architecture as T_τ. As argued in Section 2.3, after the transport map T_τ is obtained from the barycenter problem (13) or (15), the inverse $S(y,z)$ can be found through a regression problem:

$$\inf_\theta \mathbb{E}_{\rho(x,z)}\Big[c\big(x, S_\theta(T_\tau(x,z),z)\big)\Big] \approx \frac{1}{N}\sum_{i=1}^N c\big(x_i, S_\theta(y_i,z_i)\big). \tag{19}$$

3.1.2 Label net

For the factor discovery problem (15), we focus on two cases, when Z is either finite: $Z = \{1, \dots K\}$, or Euclidean: $Z = \mathbb{R}^k$. For the finite case, the conditional label distribution $p(z|x)$ becomes a probability vector, which can be modeled by the SoftMax function

$$p_\theta(z|x) = \mathrm{SoftMax}(p_\theta(x)) = \left[\frac{e^{p_\theta^1(x_i)}}{\sum_{k=1}^K e^{p_\theta^k(x_i)}},\ \dots\ \frac{e^{p_\theta^K(x_i)}}{\sum_{k=1}^K e^{p_\theta^k(x_i)}}\right],$$

where $p_\theta(x) : \mathbb{R}^d \to \mathbb{R}^K$ is some neural net. The test function $\psi^Z(z)$ reduces to a vector $[q_1, \dots q_K]$, and the transport map $T(x,z)$ splits into K maps $T^k(x)$. The objective (15) becomes

$$\begin{aligned} \sup_\theta \inf_\tau \sup_\xi L(\theta,\tau,\xi) &= \frac{1}{N}\sum_{i=1}^N \sum_{k=1}^K p_\theta(k|x_i)\big[c\big(x_i, T^k_\tau(x_i)\big) - \psi^Y_\xi\big(T^k_\tau(x_i)\big)\tilde{q}_k\big], \\ \tilde{q}_k &:= q_k - \sum_{h=1}^K q_h \sum_{i=1}^N \frac{p_\theta(h|x_i)}{N}. \end{aligned} \tag{20}$$

If we consider the conditional distributions $\rho_k := \rho(x|k)$ as clusters, then $p(k|x_i)$ is the membership probability that sample x_i belongs to cluster ρ_k, and $\rho_1, \dots \rho_K$ become a clustering plan for $\rho(x)$ with weights $\frac{1}{N}\sum_{i=1}^N p(k|x_i)$. Hence, problem (20) reduces to clustering (with soft assignments).

Remark 5. In prior work, optimal transport barycenter has been applied to the clustering problem in [45, 53], which study the case with squared Euclidean distance cost and solve directly the primal problem

$$\min_{p(k|x_i)} Var(\text{barycenter}) \tag{21}$$

instead of the dual problem (20). If we simplify the transport maps T_k by their first-moment approximations (17), then [45] shows that (21) produces the k-means algorithm. If we approximate T_k so that it aligns the second moments of $\rho(x|z)$ and μ, then [53] shows that (21) leads to more robust algorithms that recognize non-isotropic clusters. While [45, 53] only compute the membership probabilities $p(k|x_i)$, the dual problem (20) solves for both $p(k|x_i)$ and T_k, without any simplifying assumption on T_k.

For the Euclidean case $Z = \mathbb{R}^k$, we focus on deterministic labelings $p(z|x_i) = \delta_{z_i}$ and model z_i by $z_\theta(x_i)$. Then the objective (15) simplifies into

$$\begin{aligned} \sup_\theta \inf_\tau \sup_\xi L(\theta,\tau,\xi) &= \frac{1}{N}\sum_{i=1}^N \big[c\big(x_i, T_\tau(x_i, z_\theta(x_i))\big) - \psi^Y_\xi\big(T_\tau(x_i, z_\theta(x_i))\big)\tilde{\psi}^Z_\xi\big(z_\theta(x_i)\big)\big], \\ \tilde{\psi}^Z(z) &:= \psi^Z(z) - \frac{1}{N}\sum_{i=1}^N \psi^Z(z_\theta(x_i)). \end{aligned} \tag{22}$$

A useful property of the labeling $\rho(x,z)$ is that it is invariant under bijections of the latent variable space Z, because essentially we are only looking for a disintegration $\rho(x|z)$ regardless of the specific z. This is trivial for the clustering problem, since any permutation of the labels $Z = \{1, \dots K\}$ produces a different but equivalent labeling. In general, given any $\rho(x,z), T(x,z), \psi^Z(z)$ for the factor discovery problem (10) and given any measurable $f : Z \to Z$, the triple

$$\big((Id, f)\#\rho(x,z),\ T,\ \psi^Z\big)$$

produces the same value as

$$\big(\rho(x,z),\ T \circ (Id, f),\ \psi^Z \circ f\big),$$

so they can be considered equivalent solutions.

This invariance suggests that we can reduce the freedom in the architecture of z_θ without affecting the expressivity of the BaryNet. Let $NN(X, Z)$ denote the set of all neural nets mapping X to Z. Originally, z_θ should range in $NN(\mathbb{R}^d, \mathbb{R}^k)$, which is dense in $C(K, \mathbb{R}^k)$ for any compact $K \subseteq \mathbb{R}^d$, but now we can restrict to some smaller family $\mathcal{Z} \subsetneq NN(\mathbb{R}^d, \mathbb{R}^k)$ such that $NN(\mathbb{R}^k, \mathbb{R}^k) \circ \mathcal{Z}$ is dense in $C(K, \mathbb{R}^k)$. For instance, it is straightforward to show that $\mathcal{Z}$ can be the set of bounded Lipschitz neural nets whose last layer is bias-free. Such restriction is helpful for training, as it reduces the size of the search space.

Remark 6. The finite case (20) and the Euclidean case (22) can be combined into a cluster detection task: first, we solve (22) with $Z = \mathbb{R}^k$ and $k \leq 3$ small, so that Z can be visualized. Then, we inspect the latent variables $\{z_i\}$ to see if there are recognizable clusters, and how many. If so, we perform clustering (20) on either the original data $\{x_i\}$ or the processed data $\{z_i\}$.

As discussed in Section 2.4, degenerate solutions $\rho(x,z)$ can arise either because the complexity of $p(z|x)$ goes unbounded or because $\dim Z = k \geq d = \dim X$. For the latter case, we obtain the trivial labeling $z = x$, that is, $\rho(x,z) = \delta_z(x)\rho(z)$. One simple regularization is to impose a bottleneck architecture, setting $k < d$. It is analogous to the *undercomplete autoencoder* [15], whose intermediate layers are smaller than the input/output layers so that the autoencoder cannot pass by learning the identity function. As it turns out, the connection to autoencoders runs deeper than this.

3.2 Relation to autoencoders and generative modeling

The unsupervised BaryNet (15) can be conceptualized in terms of encoders and decoders. The label net z_θ (or more generally $p_\theta(z|x)$) encodes each x_i into a latent code z_i, and to recover x_i, the inverse transport map $S(\cdot,z)$ decodes z_i probabilistically as the conditional distribution $\rho(x|z_i) = S_{z_i}\#\mu$. Then the "reconstruction loss" of the encoding/decoding process should be proportional to the variability of $\rho(x|z)$,

$$\text{Reconstruction loss} \propto \int V\big(\rho(x|z)\big) dv(z). \tag{23}$$

Meanwhile, it is natural to expect that the variability of the barycenter $V(\mu)$ is positively correlated to each $V(\rho(x|z))$, that is, greater variability in the conditional distributions results in greater variability in their representative μ. By combining the two correlations, it appears that $V(\mu)$ behaves like a reconstruction loss, making the factor discovery problem analogous to an autoencoder.

We formalize this intuition in the case when $X = Y = \mathbb{R}^d$ with squared distance cost $c(x,y) = \|x-y\|^2$, and when all conditionals $\rho(x|z)$ are Gaussians. By Corollary 4, the variability V becomes the variance Var. Denote each $\rho(x|z)$ by $\mathcal{N}(\overline{x}(z), S(z))$, where $\overline{x}$ is the mean and S is the covariance matrix. Denote the principal square root matrix by $\sqrt{S}$.

Theorem 5. Given any measurable space Z and any $\rho(x,z) = \rho(x|z)v(z) \in P(\mathbb{R}^d \times Z)$ such that each $\rho(x|z)$ is a Gaussian distribution $\mathcal{N}(\overline{x}(z), S(z))$, if the marginal $\rho(x)$ has finite second moment: $\mathbb{E}_{\rho(x)}\big[\|x\|^2\big] \leq \infty$, then there exists a barycenter μ, which is a Gaussian $\mathcal{N}(\overline{x}, S)$ and satisfies

$$\begin{aligned}\overline{x} &= \int \overline{x}(z) dv(z) = \mathbb{E}_{\rho(x)}[x] \\ S &= \int \sqrt{\sqrt{S} \cdot S(z) \cdot \sqrt{S}}\; dv(z).\end{aligned}$$

Furthermore, if the set of z such that $\rho(x|z)$ is non-degenerate (its covariance $S(z)$ is positive-definite) has positive v-measure, then this is the unique barycenter.

Proof. See Appendix E. □

Theorem 5 implies that the variability of the barycenter $V(\mu) = Var(\mu) = Tr[S]$ is positively correlated to the variability of the conditional distributions $V(\rho(x|z)) = Tr[S(z)]$. A rigorous argument could apply the implicit function theorem on Banach spaces to show that S depends differentiably on $S(z) \in L^1((\mathbb{R}^k, dv) \to \mathbb{R}^{d\times d})$, and that any perturbation to $S(z)$ that increases its eigenvalues would also increase the eigenvalues of S. Nevertheless, the following corollary seems sufficient to justify the positive correlation.

Corollary 6. If we further assume that each $\rho(x|z)$ is an isotropic Gaussian: $S(z) = std^2(z) \cdot Id$, then the unique barycenter is an isotropic Gaussian with a standard deviation of

$$std = \int std(z) dv(z).$$

Proof. Insert $S(z) = std^2(z) \cdot Id$ into Theorem 5. □

It follows that $V(\mu) = std^2$ is proportional to $\int std^2(z)dv = \int V(\rho(x|z))dv(z)$.

Meanwhile, the intuition (23) can be justified by the following calculation:

$$\begin{aligned}\int Var(\rho(x|z))dv(z) &= \frac{1}{2}\iiint \|x-y\|^2 d\rho(x|z)d\rho(y|z)dv(z)\\ &= \frac{1}{2}\iiint \|x-y\|^2 d\rho(y|z)dp(z|x)d\rho(x)\\ &= \frac{1}{2}\mathbb{E}_{\rho(x)}\mathbb{E}_{p(z|x)}\mathbb{E}_{\rho(y|z)}\left[\|x-y\|^2\right]\\ &= \frac{1}{2}\mathbb{E}_{\rho(x)}\mathbb{E}_{\text{encoder}}\mathbb{E}_{\text{decoder}}\left[\text{pointwise reconstruction error } c(x,y)\right],\end{aligned}$$

where we interpret the conditional label distribution $p_\theta(z|x_i)$ as the encoder and the conditional density $\rho(x|z_i) = S_z\#\mu$ as the probabilistic decoding of x_i. The last term above is exactly the reconstruction loss of a stochastic autoencoder, and has been used as objective function by models such as the Wasserstein autoencoder (WAE) [47].

We conclude that the barycenter's variability is positively correlated with the reconstruction loss of the encoder $p_\theta(z|x)$ and decoder $S_z\#\mu$, thus verifying the analogy between unsupervised BaryNet and autoencoders. Nevertheless, a positive correlation does not imply equivalence. For instance, for the clustering problem with $Z = \{1,\dots K\}$, the classical autoencoder's reconstruction loss reduces to the sum of within-cluster variances and the algorithm becomes k-means. Yet, [53] shows empirically that minimizing the barycenter's variance leads to algorithms more robust than k-means.

Remark 7. This correlation was foreshadowed by [45], which studied the primal problem (21) of factor discovery. Suppose that all $\rho(x|z)$ have the same shape (i.e. equal up to translations), then the transport maps T_z are simplified into (17), and factor discovery reduces to finding a principal surface (hypersurface). Specifically, problem (21) can be written as

$$\begin{aligned}\min_{p(z|x)} Var(\mu) &= \min_{p(z|x)} \int Var\big(\rho(x|z)\big)dv(z)\\ &= \min_{p(z|x)}\min_{m(z)} \iint ||x-m(z)||^2 d\rho(x|z)dv(z)\\ &= \min_{p(z|x)}\min_{m(z)} \iint ||x-m(z)||^2 dp(z|x)d\rho(x) \qquad (24)\end{aligned}$$

This minimization problem can be solved by alternating descent, using the following updates:

$$\begin{aligned}p(z|x) &\leftarrow \delta_{z(x)},\ z(x) = \text{argmin}_{z\in Z}||x-m(z)||^2\\ m(z) &\leftarrow \tilde{x}(z) = \mathbb{E}_{\rho(x|z)}[x]\end{aligned}$$

Thus, we recover the principal surface algorithm [18], which produces the hypersurface $m(z)$ that summarizes the data $\rho(x)$. Meanwhile, formula (24) is also the objective function of the undercomplete autoencoder (with a probabilistic encoder) [15]. Hence, the undercomplete autoencoder can be seen as a first-moment approximation of factor discovery (when we only care about the differences in the means of $\rho(x|z)$ and restrict the transport maps to the rigid translations (17)). This equivalence is more pronounced in the linear setting: assuming further that $z(x)$ or $m(z)$ is linear, then factor discovery (17) reduces to principal component analysis [45], while the autoencoder without nonlinearity also becomes PCA [15].

Naturally, the next step is to apply the regularization techniques of autoencoders to BaryNet, as we have already explored the undercomplete autoencoder in Section 3.1.2. One popular regularization requires that the latent distribution $v(z)$ of any labeling $\rho(x,z)$ must match some prescribed distribution P_Z (e.g. a unit Gaussian in $Z = \mathbb{R}^k$). Then, the task of the autoencoder reduces to finding a coupling $\rho(x,z)$ between $\rho(x)$ and P_Z, such that the encoder $\rho(z|x)$ and decoder $\rho(x|z)$ minimize the reconstruction loss. The motivation of this regularization is that $v = P_Z$ becomes easier to sample, and together with the decoder $\pi(x|z)$, it makes possible the random sampling of $\rho(x)$. Often, the requirement that $v(z) = P_Z$ is relaxed, replacing it by a penalty on some distance between $v(z)$ and P_Z [47].

This regularization technique was introduced by the Adversarial autoencoder (AAE) [31], which refers to P_Z as the *prior* and the latent distribution $v(z)$ as the *aggregated posterior*. AAE is based on the Variational autoencoder (VAE) [26], which penalizes the KL-divergence between the conditional latent distribution $p(z|x)$ and P_Z,

$$\mathbb{E}_{\rho(x)}\big[D_{KL}(p(z|x), P_Z)\big]$$

and AAE replaces this penalty by

$$D_{GAN}(v(z), P_Z) := \max_{D(z)} \mathbb{E}_{P_z}\big[\log D(z)\big] + \mathbb{E}_{v(z)}\big[\log(1 - D(z))\big]$$

This penalty is taken from GAN, and $D(z)$ is a discriminator that estimates the likelihood that a given z comes from $v(z)$ or P_z.

By applying the regularization $v(z) = P_z$ to the factor discovery problem, we obtain

$$\min_{\substack{\rho(x,z)\in P(X\times Z)\\ \rho_X=\rho(x), \rho_Z=P_Z}} V(\text{barycenter of } \rho(x, z)).$$

Then, we derive a problem analogous to unsupervised BaryNet (15),

$$\begin{aligned}
\sup_{\theta} \inf_{\tau} \sup_{\xi} L(\theta, \tau, \xi) =& \mathbb{E}_{\rho(x)}\mathbb{E}_{p_\theta(z|x)}\big[c\big(x, T_\tau(x, z)\big) - \psi_\xi^Y\big(T_\tau(x, z)\big)\tilde{\psi}_\xi^Z(z)\big] \\
&- \big[\mathbb{E}_{\rho(x)}\mathbb{E}_{p_\theta(z|x)}\phi_\tau(z) - \mathbb{E}_{P_Z}\phi_\tau(z)\big] \\
\tilde{\psi}^Z(z) :=& \psi^Z(z) - \mathbb{E}_{\rho(x)}\mathbb{E}_{p_\theta(\tilde{z}|x)}\psi^Z(\tilde{z}),
\end{aligned} \tag{25}$$

which we call Barycentric autoencoder (BAE). The additional term in (25) is equivalent to

$$\inf_{\theta} \sup_{\tau} \mathbb{E}_{(v_\theta - P_Z)(z)}\big[\phi_\tau(z)\big]$$

and serves as a discriminator that enforces $v = P_Z$. Alternatively, we can use the f-divergences or the maximum mean discrepancy [17, 47] to penalize the difference between v and P_Z.

Once the barycenter $\mu \approx \{y_i\}_{i=1}^N$ and inverse transport map $S(y, z)$ are computed, $\rho(x, z)$ can be estimated through

$$\rho(x, z) = S\#(\mu \otimes P_Z) \approx \{S(y_i, z_j)\}_{i=1,\dots N}^{j=1,\dots M},$$

where $\{z_j\}_{j=1}^M$ is any sample drawn from P_Z. An advantage of BAE is that since the distribution P_Z is known, there is an unlimited supply of $\{z_j\}$, and thus unlimited samples for $\rho(x, z)$ and $\rho(x)$.

3.3 Semi-supervised factor discovery

Let us mention briefly that BaryNet can be adapted to the semi-supervised setting. Since the supervised (13) and unsupervised (15) differ only in the freedom to choose $p_\theta(z|x)$, it is straightforward to modify $p(z|x)$ to be semi-supervised in the following two scenarios, which we demonstrate in the deterministic case (22) with z_θ.

In the first scenario, only a subset of the sample has labels. In this case, we have a labeled sample $\{x_i^1, z_i\}_{i=1}^N$ and an unlabeled sample $\{x_j^2\}_{j=1}^M$. Then, problem (22) assigns labels or latent variables z_j to the unlabeled sample through

$$\begin{aligned}
\sup_{\theta} \inf_{\tau} \sup_{\xi} L(\theta, \tau, \xi) =& \frac{\lambda}{N}\sum_{i=1}^N \big[c\big(x_i^1, T_\tau(x_i^1, z_i)\big) - \psi_\xi^Y\big(T_\tau(x_i^1, z_i)\big)\tilde{\psi}_\xi^Z(z_i)\big] \\
&+ \frac{1-\lambda}{M}\sum_{j=1}^M \big[c\big(x_j^2, T_\tau(x_j^2, z_\theta(x_j^2))\big) - \psi_\xi^Y\big(T_\tau(x_j^2, z_\theta(x_j^2))\big)\tilde{\psi}_\xi^Z\big(z_\theta(x_j^2)\big)\big] \\
\tilde{\psi}^Z(z) :=& \psi^Z(z) - \frac{\lambda}{N}\sum_{i=1}^N \psi^Z(z_i) - \frac{1-\lambda}{N}\sum_{j=1}^M \psi^Z(z_\theta(x_j^2))
\end{aligned} \tag{26}$$

where $\lambda \in (0,1)$ is a constant that weights the two samples.

In the second scenario, only some of the labels z^1 are provided, while others z^2 are hidden. This scenario was introduced in [45] as factor discovery with confounding variables: we are given a labeled sample $\{x_i, z_i^1\}$ and need to discover a hidden label $\{z_i^2\}$. For instance, x_i can be climate data, z_i^1 can be time of the year, and the uncovered z_i^2 could correspond to a priori unknown climate patterns such as El Niño. Then, (22) assigns the label z^2 through

$$\sup_{\theta} \inf_{\tau} \sup_{\xi} L(\theta, \xi, \tau) = \frac{1}{N} \sum_{i=1}^{N} \Big[c\big(x_i, T_\tau(x_i, (z_i^1, z_\theta^2(x_i)))\big) - \psi_\xi^Y\left(T_\tau(x_i, (z_i^1, z_\theta^2(x_i)))\right) \tilde{\psi}_\xi^Z\left((z_i^1, z_\theta^2(x_i))\right) \Big]$$

$$\tilde{\psi}^Z(z) := \psi^Z(z) - \frac{1}{N} \sum_{i=1}^{N} \psi^Z\big((z_i^1, z_\theta^2(x_i))\big).$$

3.4 Optimization

We train BaryNet by gradient descent. Since (13) and (15) are min-max problems, we need optimization algorithms that are guaranteed to converge to the saddle points. Note that naïve methods such as gradient descent-ascent fails to converge [14] even for elementary problems such as $\inf_x \sup_y x \cdot y$. Among the known saddle point algorithms, we adopt the optimistic mirror descent (OMD) [32], because it is straightforward to implement, and the quasi implicit twisted descent (QITD) [14], because it automatically adjusts the learning rate and can accelerate training. For reference, the algorithms OMD and QITD are listed in Appendix F.

It is common for saddle point algorithms to assume some convexity condition, e.g. the objective function is (locally) quasiconvex-quasiconcave [14]. Then, minimax theorems such as Sion's [42] imply that (locally) the minimization and maximization can be interchanged. In fact, one can check that neither OMD nor QITD discriminate between

$$\inf_x \sup_y F(x, y)$$

and

$$\sup_y \inf_x F(x, y) = -\inf_y \sup_x \left(-F(x, y)\right),$$

since their update steps for the two problems are the same. Hence, whenever we apply these saddle point algorithms, we are allowed to exchange the inf and sup, so the factor discovery problem (15) can be modified into

$$\sup_{\theta} \inf_{\tau} \sup_{\xi} L(\theta, \tau, \xi) = \inf_{\tau} \sup_{\theta, \xi} L(\theta, \tau, \xi) \tag{27}$$

Remark 8. It is possible to avoid the min-max in problem (13) by using the maximum mean discrepancy (MMD) [17]. The max in (13) arises from the discriminators $\psi^Y(y)\psi^Z(z)$, which enforce the independence $\pi_{YZ} = \pi_Y \otimes \pi_Z$. We can instead penalize the MMD between π_{YZ} and $\pi_Y \otimes \pi_Z$: let $k(y_1, y_2), h(z_1, z_2)$ be characteristic kernels on Y, Z, then define

$$\begin{aligned} MMD(\pi_{YZ}, \pi_Y \otimes \pi_Z) &= \mathbb{E}_{[\pi_{YZ} - \pi_Y \otimes \pi_Z]^{\otimes 2}(y,z,y',z')} \left[k(y,y')h(z,z')\right] \\ &\approx \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{n=1}^{N} k(y_i, y_n) h(z_i, z_n) - \frac{2}{N^3} \sum_{i=1}^{N} \Big(\sum_{m=1}^{N} k(y_i, y_m) \cdot \sum_{n=1}^{N} h(z_i, z_n) \Big) \\ &\quad + \frac{1}{N^2(N^2-1)} \sum_{i,j=1}^{N} k(y_i, y_m) \cdot \sum_{m,n=1}^{N} h(z_j, z_n) \end{aligned}$$

where $\{y_i, z_i\}_{i=1}^N$ is a sample of π_{YZ}. Thus, (13) simplifies to a minimization problem. A disadvantage, however, is increased computational time. Computing the objective function in (13) takes only $O(N)$ time, whereas the above estimator for MMD takes $O(N^2)$ time. It is possible to use subsampling to reduce both the sample size of $\pi_Y \otimes \pi_Z$ and the cost of MMD down to $O(N)$, giving $O(N)$ time in total, but at the expense of increasing the variance of the estimator.

4 Test results

In the following sections, we test BaryNet on real and artificial datasets. Section 4.1 uses supervised BaryNet (13) on synthetic conditional density estimation problems to verify its effectiveness. Section 4.2 uses unsupervised BaryNet (15) on latent factor discovery problems, discovering meaningful hidden variables in climate and earthquake data. Section 4.3 offers a closer look into the functioning of BaryNet, showing how it learns the dependency of data x_i on label z_i and uncovers patterns in climate data. Finally, Section 4.4 applies the transport maps T_z, S_z to color transfer. All tests were conducted using `PyTorch`. See Appendix G for the network architecture and training parameters of each experiment.

The BaryNets were trained using either QITD or OMD (Appendix F). QITD is a second-order method and can automatically adjust its learning rate to accelerate training, but its time complexity is $O(TD^2)$, where T is training time and D is the dimension of the model's parameters. Whereas OMD is a first-order method with time complexity $O(TD)$. In Sections 4.1 and 4.3, we use QITD algorithm in order to obtain better convergence. In Sections 4.2 and 4.4, we apply OMD because the networks being trained are large so OMD could be more efficient.

4.1 Artificial data

In order to evaluate supervised BaryNet's performance on conditional density estimation, we devise a sample with known conditional distributions. Set $X = Y = \mathbb{R}^2, Z = \mathbb{R}$, and $c = \|x - y\|^2$. The data $\{x_i, z_i\}$ consists of 500 points drawn from the distribution $\rho(x|z)v(z)$ where $v(z)$ is uniform over $[-1, 1]$ and each $\rho(x|z)$ consists of a mixture of two Gaussians,

$$\rho(x|z) = \frac{1}{2}\mathcal{N}\left(\begin{bmatrix} (z+1)/2 \\ -(z+1)/2 \end{bmatrix}, \begin{bmatrix} 0.1 & \\ & 0.1 \end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{bmatrix} -(z+1)/2 \\ (z+1)/2 \end{bmatrix}, \begin{bmatrix} 0.1 & \\ & 0.1 \end{bmatrix}\right).$$

Below are the results of applying supervised BaryNet (13) with the QITD algorithm.

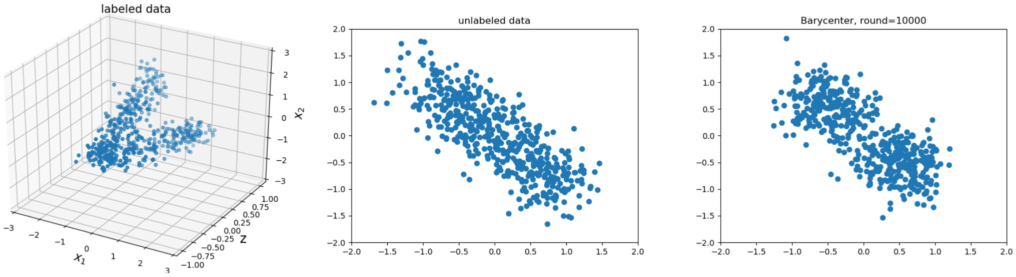

Figure 2: Left: the labeled sample $\rho(x, z) \approx \{x_i, z_i\}$. Middle: the X marginal, $\rho(x) \approx \{x_i\}$. Right: the barycenter $\mu \approx \{y_i = T(x_i, z_i)\}$ produced by BaryNet.

Then, the inverse transport map $S(y, z)$ is computed from (19) using SGD. Below, the conditional distributions $\rho(x|z) \approx \{S(y_i, z)\}$ recovered by BaryNet (in orange) are compared with samples of the same size drawn from the true distribution $\rho(x|z)$ (in blue).

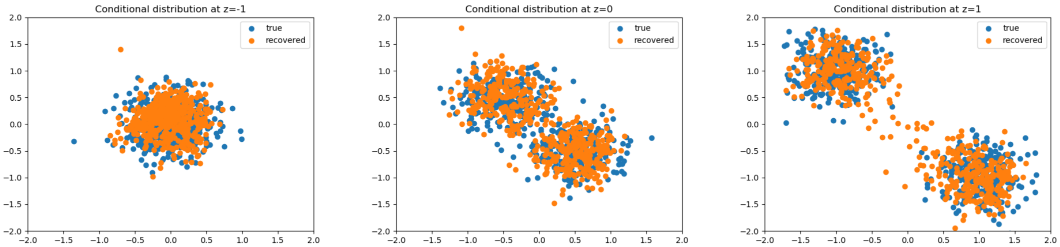

Figure 3: Left: $\rho(x|z=-1)$. Middle: $\rho(x|z=0)$. Right: $\rho(x|z=1)$. All of them show a close match.

We see that BaryNet can reliably recover the conditional distributions $\rho(x|z)$. In particular, its performance does not deteriorate in the extreme cases with $z=\pm 1$, at the two endpoints of the support of $v(z)$.

The plot below displays the evolution of the objective L from (13) during training.

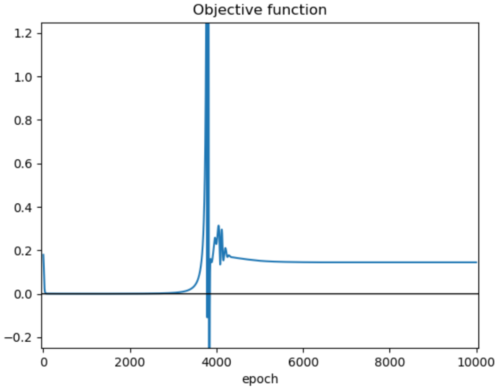

Figure 4: The objective function L during training.

As the test function $\psi^Y \psi^Z$ is initially set to zero, the cost term $\mathbb{E}[c(x, T(x,z))]$ in (13) dominates, so the transport map remains near the identity $T(x,z) \approx x$. Then, as $\psi^Y \psi^Z$ is trained and becomes increasingly discriminative, the objective L rises. The transport map T responds and the training enters a brief oscillatory period, corresponding to a competitive stage of the "game" performed by the map and the discriminator. When BaryNet converges to the right solution, L becomes flat. Faster convergence can be achieved by preconditioning ψ so that it acts from the very beginning, or implementing a two-time-scale training scheme, with the test function $\psi^Y \psi^Z$ in $\inf_T \sup_{\psi^Y \psi^Z} L$ trained faster than T [29].

4.2 Continental climate and seismic belt

To evaluate unsupervised BaryNet's performance on latent factor discovery, we apply it to real-world data that has a meaningful latent variable and test whether BaryNet can discover it. The first dataset is the average daily temperature recorded from 56 stations across U.S. in the ten-year period $[2009, 2019)$, provided by NOAA [35]. The sample space is $X = Y = \mathbb{R}^{56}$ and each x_i represents the temperature distribution in U.S. at a particular date. The cost is set to be $c = \|x-y\|^2$. An intuitive latent variable would be the seasonal effect, represented by the time of the year

$$\cos\left[\frac{2\pi}{365}(\text{date} - n)\right], \tag{28}$$

where n is the coldest day in the year (around January 15 in U.S.). Thus, the latent space is $Z = \mathbb{R}$. We apply unsupervised BaryNet (22) in its min-max formulation (27) and train it by the OMD algorithm.

As argued in Section 3.1.2, we restrict the label net $z_\theta(x)$ to be Lipschitz, using the clamp function introduced by [5], and set its last layer to be bias-free.

Below are the results of BaryNet on the temperature data $\{x_i\}$. The discovered latent variable $\{z_i\}$ exhibits periodicity in time and a strong correlation with (28), with a Pearson correlation of 0.9686, indicating that BaryNet has discovered the seasonal effect, from an input that does not contain any information on time. The non-periodic component of the discovered z, a multiyear modulation with a scale in the order of four years, is consistent with El Niño Southern Oscilation.

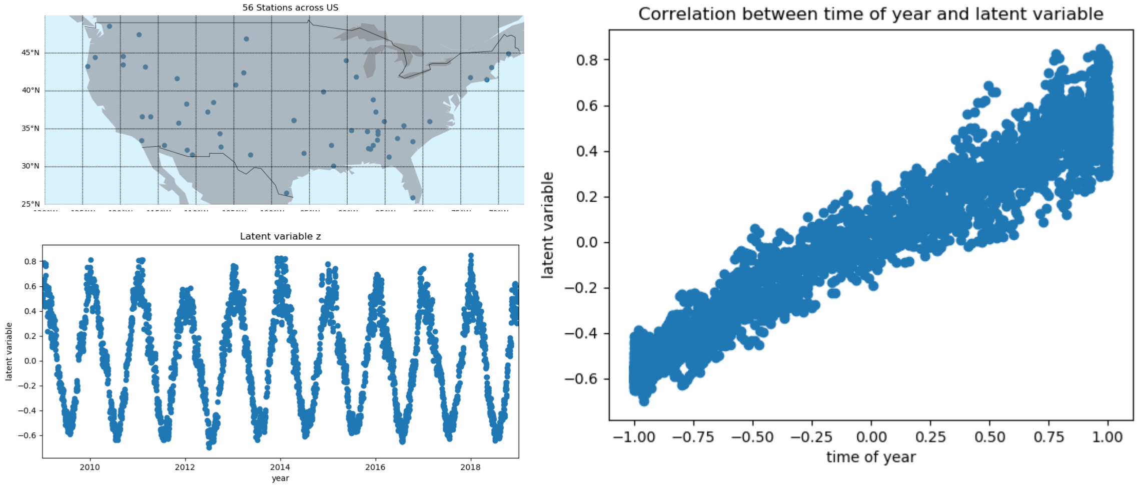

Figure 5: Up left: The chosen weather stations. Down left: The latent variables z_i discovered by BaryNet, plotted against time. Right: Scatter plot between the time of the year (28) and z_i.

The second dataset consists of earthquakes' coordinates. The data is taken from USGS [49], which records historical earthquakes from 1900 to 2008. We focus on earthquakes that occurred on the Peru-Chile Trench (see Figure 6 below). The earthquakes' locations are represented in spherical coordinates, so the sample space has $X = Y = S^2$, and we set the cost function c to be the squared great circle distance

$$c = d^2,\ d\big((x^1, x^2), (x^1_*, x^2_*)\big) = \arccos\big(\sin x^2 \sin x^2_* + \cos x^2 \cos x^2_* \cos(x^1 - x^1_*)\big),$$

where x^1, x^2 are longitude and latitude. Judging from the earthquake plot in Figure 6, since the earthquakes are distributed roughly vertically along the Peru-Chile Trench, it is intuitive that the (one-dimensional) latent variable should be proportional to the earthquake's latitude.

We apply unsupervised BaryNet (27) and train it by OMD. The discovered latent variable $\{z_i\}$ shows a strong correlation with the latitude, with a correlation of 0.9954, indicating that BaryNet has discovered the seismic belt.

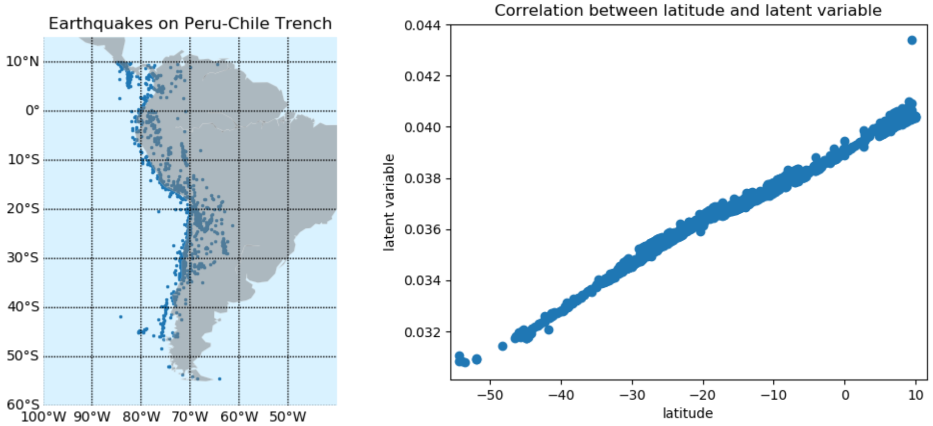

Figure 6: Left: Earthquakes on Peru-Chile Trench. Right: Scatter plot between the earthquake's latitude and the latent variables z_i discovered by BaryNet.

The label nets $z_\theta(x)$ for both the temperature and the earthquake tests are intentionally set to be feedforward. Unlike the residual nets, these are highly non-linear maps without any linear component, and it is difficult for them to learn linear mappings such as $(x^1, x^2) \mapsto x^2$, making it highly unlikely that BaryNet arrived at the desired solutions by chance.

4.3 Hourly and seasonal temperature variation

As discussed in the introduction, supervised BaryNet "learns" the data $\rho(x, z)$ by decomposing it into a representative μ of the conditional densities $\rho(x|z)$ plus the transformations between them. Equivalently, it represents $\rho(x, z)$ as $S\#(\mu \otimes v)$. Thus, BaryNet can be seen as a "probabilistic" generalization of regression, which learns the probabilistic mapping $z \mapsto \rho(x|z)$, approximating it via

$$z \mapsto (S_\theta(\cdot, z) \circ T_\tau)\#\rho = (S_z)_\theta \# \mu_\tau. \tag{29}$$

Its expressivity derives from the nonlinearity of T_τ and S_θ, which enables BaryNet to learn the complex dependency of the data x on the latent variable z.

We demonstrate this intuition using meteorological data where x_i is the average hourly temperature at Ithaca, NY in the ten-year period $[2007, 2017)$, provided by NOAA [36]. The latent variables z_i chosen are the time of day and time of year, represented by

$$\sin(2\pi \text{hour}/24),\ \ \cos(2\pi \text{hour}/24),\ \ \sin(2\pi \text{date}/365),\ \ \cos(2\pi \text{date}/365) \tag{30}$$

Hence $X = Y = \mathbb{R}$, $Z = \mathbb{R}^4$, and we set $c = (x - y)^2$.

BaryNet (13) is trained on $\{x_i, z_i\}$ by the QITD algorithm. To facilitate visualization, the probabilistic regression (29) is displayed through its conditional mean:

$$z \mapsto \mathbb{E}_{(S_z\#\rho)(x)}[x] \approx \frac{1}{N}\sum_{i=1}^{N} S_\theta(y_i, z)$$

We compare BaryNet with mean-square regression on $\{x_i, z_i\}$. The regression problem fits $\rho(x|z)$ under mean-square loss, and the optimizer is also the conditional mean $\mathbb{E}_{\rho(x|z)}[x]$, so it constitutes a valid comparison. Specifically, we perform linear regression, using the latent variables z (30) as features. It may seem unfair to compare the nonlinear BaryNet with a linear model, but the key is that BaryNet is solving the more difficult problem of learning the entire $\rho(x|z)$, whereas regression only learns $\mathbb{E}_{\rho(x|z)}[x]$, so we are satisfied as long as BaryNet performs at least as well as linear regression in terms of the conditional mean.

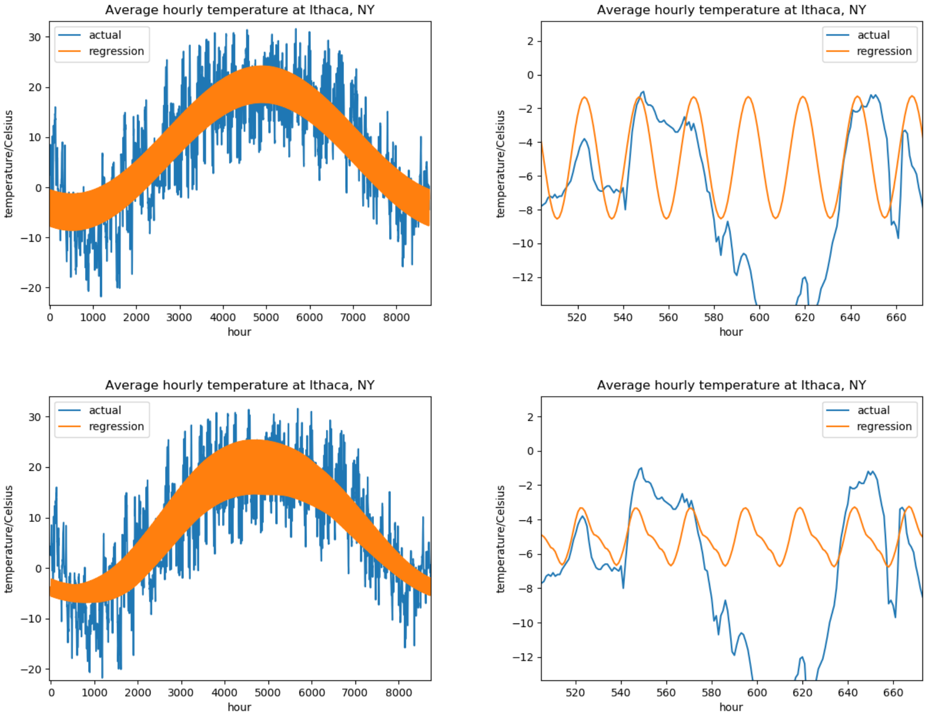

Figure 7: Hourly temperature at Ithaca, NY. Left column: data for the year 2007. Right column: data for the week of Jan 21-28, 2007. Top row: Linear regression on $\{x_i, z_i\}$. Bottom row: Supervised BaryNet. The blue curve represents the temperature data, and the orange curve the regression (for linear regression) and the conditional mean (for BaryNet). (Due to the daily oscillation, the regression curve appears as a thick band in the yearly plot). The regression curves cannot perfectly fit the data, since date and hour alone cannot fully account for the irregularity of weather systems.

BaryNet learns more features of the data than linear regression in both the yearly and weekly plots. In the yearly plot, BaryNet's curve oscillates with greater amplitude during summer than in winter, indicating that the daily temperature during summer has greater variance. In the weekly plot, BaryNet's curve is highly non-sinusoidal, and has greater upward slope during the day than downward slope at night, in agreement with the real daily cycle during winter time.

4.4 Color transfer

We describe briefly here a useful application of the transport maps. As a by-product of BaryNet, the transport map $T_z(x)$ and inverse transport map $S_z(y)$ can be concatenated into a transportation between any pair of conditional distributions:

$$(S_{z_2} \circ T_{z_1})\#\rho(x|z_1) = \rho(x|z_2)$$

Even though we can directly compute a transport map from $\rho(x|z_1)$ to $\rho(x|z_2)$, BaryNet is more efficient if one seeks all pairwise transport maps (just as we may seek all pairwise translations among several languages). If there are K labels (Z=$\{1,\ldots$ K$\}$), then there will be $O(K^2)$ pairwise transport maps, whereas BaryNet only needs to compute $2K$ maps, T_k and S_k. If Z is continuous (e.g. $\mathbb{R}^k$), then BaryNet only needs two maps, $T(x,z)$ and $S(y,z)$, while doing individual pairwise maps is infeasible, not only because there are infinitely many of them, but also because each conditional distribution has one sample point or less.

Instead of languages, we apply this procedure to images. An image can be seen as a 3-dimensional matrix of size $3 \times H \times W$, which represent the RGB color channels, height and width (in terms of pixels). Alternatively, an image can be treated as a sample of $H \times W$ points in $\mathbb{R}^3$. Thus, we view the following images as color distributions, $\rho_1, \rho_2, \rho_3 \in P(\mathbb{R}^3)$, and apply BaryNet to compute the transport maps T_k, S_k.

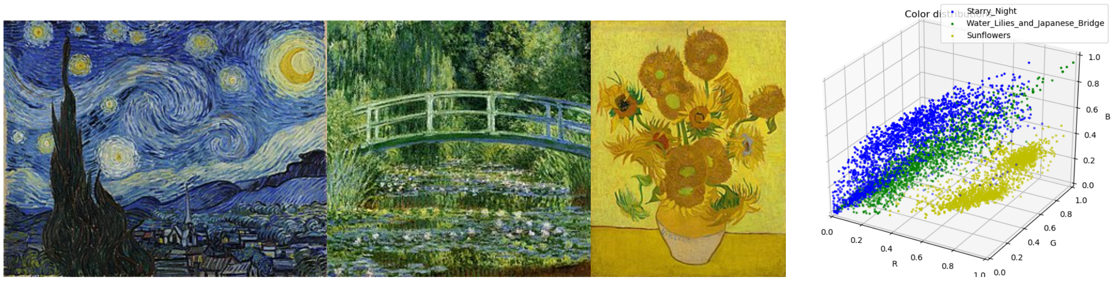

Figure 8: "Starry Night", "Water Lilies and Japanese Bridge", "Sunflowers" by Monet and van Gogh, and their color distributions in $\mathbb{R}^3$.

Then, the coloring style of image j can be transferred to image k by applying pixel-wise the transport map $S_j \circ T_k$ to image k. The results are displayed in Figure 9 below. We used supervised BaryNet (13) with $Z = \{1, 2, 3\}$ and trained it by OMD.

Remark 9. A task very similar to color transfer is *color normalization* [28], that seeks to eliminate the differences in several images' color distributions with minimum distortion. For instance, in medical imaging, we may want to remove irrelevant variations, such as different lighting conditions or staining techniques. BaryNet can easily solve this task by mapping the color distributions ρ_k to their common barycenter, which by definition minimizes total distortion.

5 Conclusions

Conditional density estimation and latent variable discovery are two intimately related problems in machine learning: one learns the dependence of data x on a given variable z, while the other infers a latent variable z from data x. This paper proposes to solve both problems in the framework of optimal transport barycenters. Our method is based on an intuitive principle of minimum uncertainty, that is, the goal of learning is to reduce some measure of uncertainty or variability. Specifically, for latent variable discovery, we begin with some data $\rho(x)$ and our learning ends with a joint distribution $\rho(x, z)$, which assigns the latent variables z to each data point x through $\rho(z|x)$. Our principle leads to maximizing the reduction in variability from $\rho(x)$ to $\rho(x, z)$.

How should we characterize the variability of a joint distribution $\rho(x, z)$? A simple approach is to take the mean variability of the conditional distributions $\rho(x|z)$. Yet, this approach does not always lead to sensible results: the k-means algorithm, for example, minimizes the sum of squared errors, or equivalently the weighted average of each cluster's variance, and it is known that it often fails to recognize clusters with different sizes and shapes [53].

Instead, we seek a distribution μ that can act as a representative of all $\rho(x|z)$, and measure the variability of $\rho(x, z)$ by that of μ. This idea naturally leads to the optimal transport barycenter, which minimizes a user-specified distance between μ and each $\rho(x|z)$. A characterization of variability arises from the barycenter's optimal transport cost, indicating that variability and transport cost are two sides of the same coin. Under simplifying assumptions [45, 53], this definition of $\rho(x, z)$'s variability includes the aforementioned simple approach as special case.

It follows that latent variable discovery should seek assignments $\rho(z|x)$ that minimize the variability of the barycenter of the $\rho(x|z)$. At the same time, conditional density estimation also benefits from the barycenter representation. The difficult task of learning the possibly infinitely many $\rho(x|z)$ is reduced to learning just the barycenter μ, from which we can recover each $\rho(x|z)$.

The contributions of this paper are

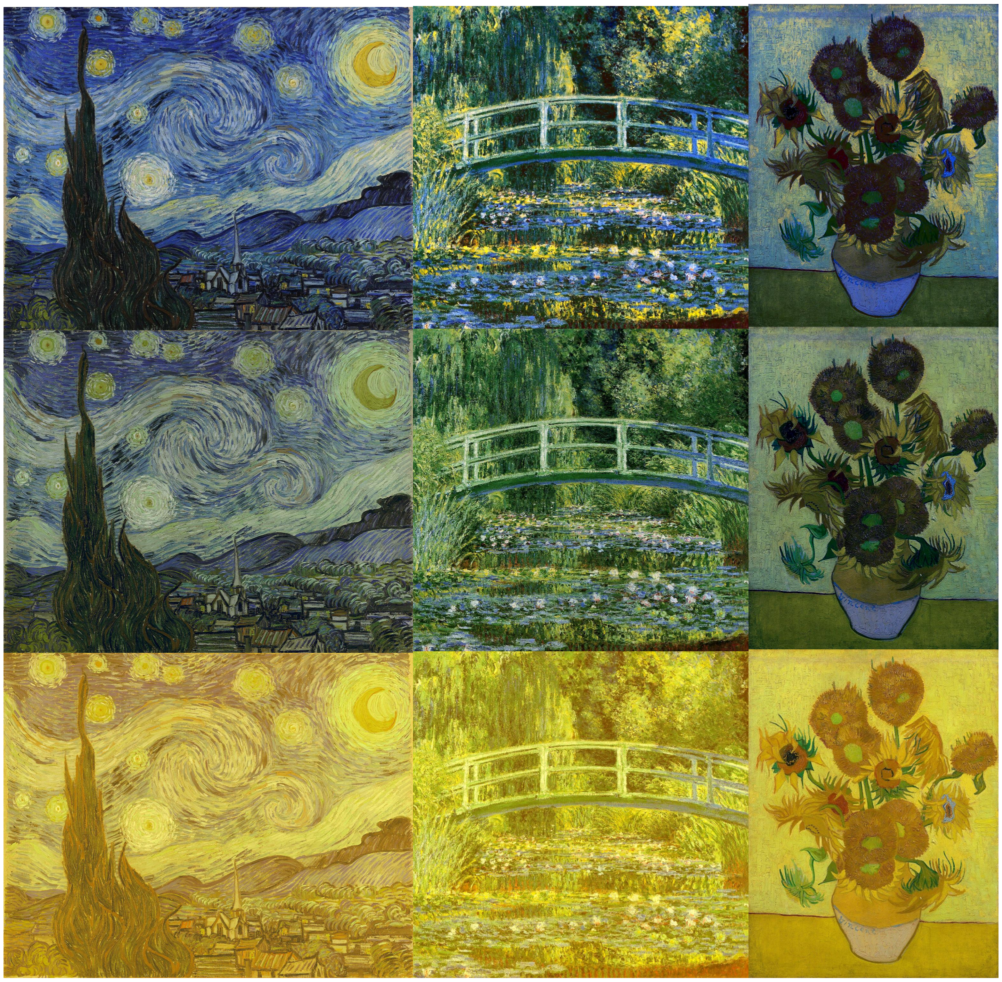

Figure 9: Color transferred images. The k-th column and j-th row is obtained by $S_j \circ T_k \# \rho_k$, so it is image k with the color of image j.

1. The introduction of the BaryNet algorithms, which use neural nets. The unsupervised BaryNet performs latent variable assignment $\rho(z|x)$, while the supervised BaryNet computes the barycenter μ and estimates each conditional distribution $\rho(x|z)$. Their effectiveness is confirmed by tests on artificial and real-world data, with Euclidean and non-Euclidean costs.

2. Enrichment of the theory of optimal transport barycenters. In particular, the existence of Kantorovich and Monge solutions for barycenters with infinitely many $\rho(x|z)$ are studied (Theorems 1 and 2), and geometric properties of the Wasserstein space are discovered that resemble those of Euclidean space (Theorems 3 and 5).

3. An intimate connection between autoencoders and BaryNet is identified. In particular, with squared Euclidean distance cost and the simplifying assumption that $\rho(x|z)$ are equivalent up to translation, BaryNet includes the following algorithms as special cases: k-means, PCA, principle curves and surfaces, and undercomplete autoencoders.

The theoretical framework developed in this article opens up several new directions of research:

1. Parallelism to autoencoders. We proposed the Barycentric autoencoder (BAE) algorithm in Section 3.2, based on the parallelism between autoencoders and unsupervised BaryNet. It would

be interesting to compare the performance of BAE with that of VAE, WAE or AAE. One can also apply the regularizations of denoising autoencoders and sparse autoencoders [15] to BaryNet.

2. Cooperation for density estimation. Given labeled data $\rho(x,z)$ and any density estimation algorithm or generative modeling algorithm, such as WGAN [5], one can check whether the algorithm's performance can be improved by first estimating the density of the barycenter μ, and then transporting to each $\rho(x|z)$ through BaryNet, instead of learning the joint distribution $\rho(x,z)$ directly.

3. Transfer learning and domain adaptation. The semisupervised BaryNet introduced in Section 3.3 can be applied to solve transfer learning problems. For instance, given some unlabeled data $\{x'_j\}$ and a few labeled data $\{x_i, z_i\}$ (such that x_i and x'_j may be drawn from different distributions), we can perform classification on $\{x'_j\}$ based on the information of $\{x_i, z_i\}$. Then, the semisupervised BaryNet (26) produces a labeling in accordance with the principle of minimum uncertainty.

4. Metric learning: When a clustering plan or label assignment $\rho(x,z)$ is provided, we can try to infer what metric or cost function is responsible for that particular assignment of z. As an example, let X be the space of images, and $\rho(x,z)$ be some image dataset whose labels are text descriptions. Since human vision is better tuned to discerning faces than inanimate objects such as buildings, there would be more labels related to faces than to buildings. Specifically, X's subspace of facial images would have a greater density of labels z than the subspace of buildings' images, or equivalently, if label z_1 concerns facial feature (e.g. "smiling face") and z_2 concerns building's feature (e.g. "old building"), then the Euclidean variance of $\rho(x|z_1)$ is most likely smaller than that of $\rho(x|z_2)$. Assume that g is a Riemannian metric on X, such that $\rho(x,z)$ becomes an optimal solution to the factor discovery problem (10) with squared geodesic distance cost $c = d^2(x,y)$ induced by g, then (X,g) should not be Euclidean, and instead g should assign greater distances to the subspace of faces, so that factor discovery adapts to g by making $\rho(x|z_1)$ "smaller" than $\rho(x|z_2)$. Note that, (10) evaluates any assignment $\rho(x,z)$ when given a cost $c(x,y)$; alternatively, it might be modified so as to evaluate any candidate cost or metric when given an assignment.

Acknowledgments

The work of E. G. Tabak was partially supported by NSF grant DMS-1715753 and ONR grant N00014-15-1-2355.

References

[1] Advani, M. S., and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667* (2017).

[2] Agnelli, J. P., Cadeiras, M., Tabak, E. G., Turner, C. V., and Vanden-Eijnden, E. Clustering and classification through normalizing flows in feature space. *SIAM MMS 8* (2010), 1784–1802.

[3] Agueh, M., and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis 43*, 2 (2011), 904–924.

[4] Alpaydin, E. *Introduction to machine learning.* MIT press, 2009.

[5] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875* (2017).

[6] Billingsley, P. *Convergence of probability measures*, 2 ed. Wiley, 1999.

[7] Bishop, C. M. Mixture density networks. Aston University, 1994.

[8] Bishop, C. M. *Pattern recognition and machine learning.* Springer, 2006.

[9] Carlier, G., and Ekeland, I. Matching for teams. *Economic theory 42*, 2 (2010), 397–418.

[10] Chang, J. T., and Pollard, D. Conditioning as disintegration. *Statistica Neerlandica 51*, 3 (1997), 287–317.

[11] Chiappori, P.-A., McCann, R. J., and Nesheim, L. P. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Economic Theory 42*, 2 (2010), 317–354.

[12] E, W., Ma, C., and Wu, L. A priori estimates for two-layer neural networks. *arXiv preprint arXiv:1810.06397* (2018).

[13] E, W., Ma, C., and Wu, L. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039* (2019).

[14] Essid, M., Tabak, E., and Trigila, G. An implicit gradient-descent procedure for minimax problems. *arXiv preprint arXiv:1906.00233* (2019).

[15] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning.* MIT press, 2016.

[16] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems* (2014), pp. 2672–2680.

[17] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research 13*, Mar (2012), 723–773.

[18] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer 27*, 2 (2005), 83–85.

[19] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[20] Holmes, M. P., Gray, A. G., and Isbell, C. L. Fast nonparametric conditional density estimation. *arXiv preprint arXiv:1206.5278* (2012).

[21] Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural networks 4*, 2 (1991), 251–257.

[22] Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics 5*, 4 (1996), 315–336.

[23] Kallenberg, O. *Random measures, theory and applications.* Springer, 2017.

[24] Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)* (1942), vol. 37, pp. 199–201.

[25] Kim, Y.-H., and Pass, B. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics 307* (2017), 640–683.

[26] Kingma, D. P., and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[27] Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680* (2019).

[28] Li, X., and Plataniotis, K. N. A complete color normalization approach to histopathology images using color cues computed from saturation-weighted statistics. *IEEE Transactions on Biomedical Engineering 62*, 7 (2015), 1862–1873.

[29] Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331* (2019).

[30] Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems* (2017), pp. 6231–6239.

[31] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[32] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations* (2019), ICLR.

[33] Mirza, M., and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[34] Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[35] NOAA. Daily temperature data set. `www1.ncdc.noaa.gov/pub/data/uscrn/products/daily01/`, 2019. Accessed: 2019-07-21.

[36] NOAA. Hourly temperature data set. `https://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/`, 2019. Accessed: 2019-07-26.

[37] Pass, B. Multi-marginal optimal transport and multi-agent matching problems: uniqueness and structure of solutions. *arXiv preprint arXiv:1210.7372* (2012).

[38] Pass, B. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis 264*, 4 (2013), 947–963.

[39] Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* (2011), Springer, pp. 435–446.

[40] Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[41] Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. On the spectral bias of neural networks. *arXiv preprint arXiv:1806.08734* (2018).

[42] Simons, S. Minimax theorems and their proofs. In *Minimax and applications*. Springer, 1995, pp. 1–23.

[43] Sipser, M. A Definition of Information. In *Introduction to the Theory of Computation*, 3 ed. Cengage Learning, 2013, ch. 6.4.

[44] Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems* (2015), pp. 3483–3491.

[45] Tabak, E. G., and Trigila, G. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics 71*, 1 (2018), 163–199.

[46] Tabak, E. G., Trigila, G., and Zhao, W. Conditional density estimation and simulation through optimal transport. *Submitted to SIAM Journal on Mathematics of Data Science* (2018).

[47] Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558* (2017).

[48] Trippe, B. L., and Turner, R. E. Conditional density estimation with bayesian normalising flows. *arXiv preprint arXiv:1802.04908* (2018).

[49] USGS. Centennial earthquake catalog. `https://earthquake.usgs.gov/data/centennial/centennial_Y2K.CAT`, 2008. Accessed: 2019-07-30.

[50] Villani, C. *Topics in optimal transportation*. No. 58. American Mathematical Soc., 2003.

[51] Villani, C. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

[52] Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523* (2019).

[53] Yang, H., and Tabak, E. G. Clustering, factor discovery and optimal transport. *arXiv preprint arXiv:1902.10288* (2019).

Appendices

A Assumptions

Assumption 1. The cost function c on $X \times Y$ is *locally uniformly coercive*, that is, for some (and thus every) $y_0 \in Y$ and for every compact $K \subseteq X$,

$$\lim_{d(y,y_0)\to\infty} \inf_{x \in K} c(x,y) = \infty$$

where the limit is taken over all sequences in Y. Also, the space Y satisfies the *Heine-Borel property* that every closed bounded subset is compact.

Most cost functions c used in practice are locally uniformly coercive: for instance, given any metric space (X, d), the l^p distance cost $d(x_1, x_2)^p$ with $p \in (0, \infty)$ trivially satisfies the condition. Also, a broad class of spaces Y satisfy the Heine-Borel property, including Euclidean spaces, complete Riemannian manifolds, and all their closed subsets.

Assumption 2. Assume that for all $\mu \in P(Y)$ and v-almost all z, the optimal coupling between $\rho(x|z)$ and $\mu(y)$ is unique and is a Monge solution, i.e. $\pi(x, y|z) = (Id, T_z)\#\rho(x|z)$ for some transport map T_z.

This assumption holds in most real-world applications. For instance, by Theorem 2.44 of [50], if $X = Y = \mathbb{R}^d$ and the cost c is strictly convex and superlinear, then it holds whenever almost all $\rho(x|z)$ are absolutely continuous. In practice, we are only given samples drawn from unknown probabilities, so we can typically assume that they come from absolutely continuous $\rho(x|z)$.

B Proof of Theorem 1

Since we constructed the conditional distributions $\rho(x|z)$ using disintegration, the map $Z \to P(X) \times P(Y)$, $z \mapsto (\rho(x|z), \mu)$ is automatically measurable (in the topology of weak convergence) for any given μ. Then, Corollary 5.22 of [51] implies that there is a measurable assignment: $z \mapsto \pi(x, y|z)$ such that each $\pi(x, y|z)$ is an optimal transport plan between $\rho(x|z)$ and $\mu(y)$.

Given that c is bounded below, let $c_n := \min(c, n)$ be a sequence of bounded continuous functions that increases pointwise to c. Then, by the monotone convergence theorem, the total transport cost from $\rho(x, z)$ to $\mu(y)$ becomes

$$\int_Z I_c(\rho(\cdot|z), \mu) d\nu(z) = \iint c(x, z) d\pi(x, y|z) d\nu(z) = \lim_{n\to\infty} \iint c_n d\pi(x, y|z) d\nu(z)$$

The last term is well-defined since $z \mapsto \int c_n d\pi(\cdot|z)$ is measurable, so the barycenter problem is well-defined.

Factor the measure $\pi(x, y, z)$ into $\pi(x, y|z)\nu(z)$. Then, integrating π yields 1, showing that π is a probability measure, and integrating π times test functions $\psi \in C_b(X \times Z)$ and $\phi \in C_b(Y \times Z)$ shows that π has the marginals: $\pi_{XZ} = \rho(x, z)$ and $\pi_{YZ} = \mu(y) \otimes \nu(z)$. This proves the $\geq$ side of (4). Since the $\leq$ side of (4) is evident, the first assertion of Theorem 1 is proved.

To prove the second assertion, we need the following lemmas:

Lemma 7. Given any $\mu \in P(Y)$, the total transport cost (3) satisfies the following duality formula

$$\begin{aligned}(3) &= \min_{\substack{\pi\in P(X\times Y\times Z)\\ \pi_{XZ}=\rho\\ \pi_{YZ}=\mu\otimes v}} \int_{X\times Y\times Z} c(x, y) d\pi(x, y, z)\\ &= \sup_{\substack{\phi\in C_b(X\times Z)\\ \psi\in C_b(Y\times Z)\\ \phi+\psi\leq c}} \int_{X\times Z} \phi(x, z) d\rho(x, z) + \int_{Y\times Z} \psi(y, z) d\mu(y) d\nu(z),\end{aligned} \tag{31}$$

so that (3) is lower semi-continuous in μ in the topology of weak convergence of $P(Y)$.

Proof. Define a cost function from $X \times Z$ to $Y \times Z$,

$$\tilde{c}(x, z_1, y, z_2) = c(x, y) + \infty \cdot \delta_{z_1\neq z_2} = \begin{cases} c(x, y) \text{ if } z_1 = z_2 \\ \infty \ \text{ otherwise,} \end{cases}$$

which is lower semi-continuous on $X \times Z \times Y \times Z$. Then, we can apply Kantorovich duality (Theorem 5.10 of [51]) to $\rho(x, y)$ and $\mu \otimes v$ to obtain the duality formula

$$\min_{\substack{\tilde{\pi}\in P(X\times Z\times Y\times Z)\\ \pi_{XZ_1}=\rho\\ \pi_{YZ_2}=\mu\otimes v}} \int \tilde{c}\ d\tilde{\pi} = \sup_{\substack{\phi\in C_b(X\times Z)\\ \psi\in C_b(Y\times Z)\\ \phi+\psi\leq c}} \int_{X\times Z} \phi(x, z) d\rho(x, z) + \int_{Y\times Z} \psi(y, z) d\mu(y) d\nu(z).$$

This is part of the theorem that the minimum on the left side is achieved.

It remains to show that

$$\min_{\substack{\tilde{\pi}\in P(X\times Z\times Y\times Z)\\ \pi_{XZ_1}=\rho\\ \pi_{YZ_2}=\mu\otimes v}} \int \tilde{c}\ d\tilde{\pi} = \min_{\substack{\pi\in P(X\times Y\times Z)\\ \pi_{XZ}=\rho\\ \pi_{YZ}=\mu\otimes v}} \int_{X\times Y\times Z} c\ d\pi. \tag{32}$$

Define the map $P(x, y, z) = (x, z, y, z)$. For the $\leq$ part of (32): given any coupling $\pi(x, y, z)$ that solves the right side, the pushforward $\tilde{\pi} := P\#\pi$ is applicable to the left side. For the $\geq$ part: if the left side is infinite, then we are done. Else, given an optimal coupling $\tilde{\pi}(x, z_1, y, z_2)$,

$$\int \tilde{c}\ d\tilde{\pi} = \int c(x, y) + \infty \cdot \delta_{z_1\neq z_2}\ d\tilde{\pi}(x, y|z_1, z_2) d\tilde{\pi}_{Z_1 Z_2}(z_1, z_2) < \infty.$$

It follows that $\tilde{\pi}_{Z_1 Z_2}\big(\{z_1 \neq z_2\}\big) = 0$, so the measure $\tilde{\pi}$ must be concentrated on the diagonal $\{z_1 = z_2\}$. Then, we can define the pullback $\pi = P^{-1}\#\tilde{\pi}$, which has the correct marginals π_{XZ}, π_{YZ} and is applicable to the right side of (32):

$$\int c\ d\pi = \int c(x, y)\ d\tilde{\pi}(x, z, y, z) = \int \tilde{c}\ d\tilde{\pi}.$$

Hence, we have proved formula (31).

Theorem 2.8 of [6] implies that for any $\psi \in C_b(Y \times Z)$, the map

$$\mu \mapsto \mu \otimes v \mapsto \iint \psi \; dv d\mu$$

is a continuous linear functional in the topology of weak convergence of $\mu \in P(Y)$. So the right side of (31), as a supremum over continuous functions, is lower semi-continuous in μ. □

Lemma 8. Given any $\mu(y) \in P(Y)$, the total transport cost (3) has the lower bound:

$$I_c(\rho(x,z),\mu) \geq I_c(\rho(x),\mu)$$

where we "forget" the labeling z on the right side.

Proof. This is a direct corollary of Theorem 4.8 of [51]. □

Denote the total transport cost (31) by $F(\mu)$. Let $\{\mu^n\} \subseteq P(Y)$ be a minimizing sequence such that $F(\mu^n)$ converges to the optimal transport cost $\inf F$. If $\inf F = \infty$, then any $\mu \in P(Y)$ can serve as a barycenter. Else, assume that $\inf F < \infty$ and choose a constant $C \geq 0$ large enough so that $\sup F(\mu^n) < C$.

We show that Assumption 1 implies that $\{\mu^n\}$ is tight. Since Y is assumed to satisfy the Heine-Borel property, it suffices to show that, for any $\epsilon > 0$, there is some radius R such that

$$\sup_n \mu^n\big(B_R^C(y_0)\big) \leq \epsilon, \tag{33}$$

where $y_0 \in Y$ is some arbitrary point and $B_R^C(y_0) = \{y \in Y, d(y,y_0) \geq R\}$.

Let $K \subseteq X$ be a compact set whose complement has small measure $\rho(K^C) < \epsilon/2$. Since c is assumed to be locally uniformly coercive and bounded below, we can choose a radius R such that

$$\inf_{d(y,y_0)>R} \inf_{x \in K} c(x,y) > 2(C - \inf c)/\epsilon.$$

Assume for contradiction that there is some μ^n such that $\mu^n\big(B_R^C(y_0)\big) > \epsilon$. Then, for any coupling π between $\rho(x)$ and μ^n,

$$\pi\big(K \times B_R^C(y_0)\big) \geq \mu^n\big(B_R^C(y_0)\big) - \rho(K^C) > \epsilon/2$$

that is, any transport plan must move more than $\epsilon/2$ of mass inside $K \subseteq X$ to $B_R^C(y_0) \subseteq Y$, which gives a lower bound on the transport cost:

$$\int c \; d\pi \geq 2(C - \inf c)/\epsilon \cdot \pi\big(K \times B_R^C(y_0)\big) + \inf c \cdot \big[1 - \pi\big(K \times B_R^C(y_0)\big)\big] > C$$

Therefore, the optimal transport cost is bounded by $I_c(\rho(x),\mu^n) \geq C$.

Meanwhile, Lemma 8 implies that $F(\mu^n) \geq I_c(\rho(x),\mu^n)$. Hence,

$$C > F(\mu^n) > C$$

a contradiction. It follows that condition (33) holds, so that $\{\mu^n\}$ is uniformly tight.

By Prokhorov's theorem, $\{\mu^n\}$ is precompact, so some subsequence $\{\mu^{n_k}\}$ converges weakly to some $\mu \in P(Y)$. By Lemma 7, F is lower semi-continuous, so

$$\inf F = \liminf F(\mu^{n_k}) \geq F(\mu) \geq \inf F.$$

It follows that μ minimizes the total transport cost and is a barycenter of $\rho(x,z)$.

Finally, notice that the two minimization problems are equivalent:

$$\min_{\mu \in P(Y)} \min_{\substack{\pi \in P(X \times Y \times Z) \\ \pi_{XZ} = \rho \\ \pi_{YZ} = \mu \otimes v}} \int c \; d\pi = \min_{\substack{\pi \in P(X \times Y \times Z) \\ \pi_{XZ} = \rho \\ \pi_{YZ} = \pi_Y \otimes \pi_Z}} \int c \; d\pi$$

Then formula (5) is proved.

C Proof of Theorem 2

We need the following results:

Lemma 9. Let Y, Z be metric spaces and π a Borel probability measure on $Y \times Z$. The two marginals π_Y, π_Z are independent ($\pi = \pi_Y \otimes \pi_Z$) if and only if for all $f \in C_b(Y), g \in C_b(Z)$,

$$\int_{Y \times Z} f(y)g(z)d\pi(y,z) = \int_Y f(y)d\pi_Y(y) \int_Z g(z)d\pi_Z(z). \tag{34}$$

Proof. The "only if" follows from straightforward integration. For the "if" part, let $U \subseteq Y, W \subseteq Z$ be arbitrary closed subsets, and define the following C_b functions

$$f_n(y) = \max\big(1 - n \cdot d_Y(y,U), 0\big),\ g_n(z) = \max\big(1 - n \cdot d_Z(z,U), 0\big)$$

that descend to the indicator functions of U and W. Apply (34) to $f_n \cdot g_n$ and take the limit $n \to \infty$; the Dominated Convergence Theorem implies that

$$\pi(U \times W) = \pi_Y(U)\pi_Z(W). \tag{35}$$

Let P_Y, P_Z be the collections of all closed subsets of Y, Z. Let S_Y be the collection of all Borel measurable subsets U of Y such that for any closed $W \in P_Z$, the independence formula (35) holds. We seek to apply Dynkin's π-λ theorem. $P_Y \subseteq S_Y$ is closed under intersection and thus is a π-system. Meanwhile, it is straightforward to show that S_Y is closed under set difference and countable union of increasing sequence, so S_Y is a λ-system. It follows from Dynkin's theorem that S_Y contains the σ-algebra of Y.

Similarly, we define S_Z to be the collection of all Borel measurable subsets W of Z such that for any measurable $U \in S_Y$ (not just P_Y), the independence formula (35) holds. Repeating the above argument for S_Z, S_Y shows that S_Z contains the σ-algebra of Z. Hence, (35) holds for all measurable rectangles in $Y \times Z$ and $\pi = \pi_Y \otimes \pi_Z$. □

Corollary 10. Given the same condition as in Lemma 9, the independence $\pi = \pi_Y \otimes \pi_Z$ holds if and only if for all $f \in C_b(Y), g \in C_b(Z)$,

$$\int g(z)d\pi_Z(z) = 0 \to \int_{Y \times Z} f(y)g(z)d\pi(y,z) = 0. \tag{36}$$

Proof. The condition (34) can be rearranged into

$$\int_{Y \times Z} f(y)\big[g(z) - \int g\ d\pi_Z\big]d\pi(y,z) = 0.$$

□

Lemma 11. Let A be any closed subset of a metric space Y. The set of Dirac masses on A:

$$\Delta_A = \{\delta_y,\ y \in A\}$$

is closed in the weak topology of $P(Y)$.

Proof. Suppose a sequence $\{\delta_{y_n}\}$ in Δ_A converges weakly to some $\mu \in P(Y)$. For any $n \geq 1$, let A_n be the closure of the subsequence $\{y_m\}_{m \geq n}$. Then, by weak convergence, $\mu(A_n) = 1$ for all n and thus $\mu(A_\infty) = 1$ where $A_\infty = \cap_n A_n$ is the set of limits of $\{y_n\}$. It follows that the set A_∞ is non-empty. Meanwhile, A_∞ cannot contain more than one point, otherwise it would contradict the weak convergence of $\{\delta_{y_n}\}$. Hence, μ is the Dirac mass δ_y where $y \in A$ is the only point in A_∞. □

Corollary 12. Given a metric space Y, there exists a measureable function $F_\Delta : P(Y) \to Y$ such that $F_\Delta(\delta_y) = y$ for every Dirac mass δ_y.

Proof. We can simply define (for some arbitrary fixed $y_0 \in Y$),

$$F_\Delta(\mu) = \begin{cases} y \text{ if } \mu \in \Delta_Y \text{ and } \mu = \delta_y \\ y_0 \text{ else} \end{cases}$$

Then, for any closed subset $A \subseteq Y$,

$$F_\Delta^{-1}(A) = \begin{cases} \Delta_A \text{ if } y_0 \notin A \\ \Delta_A \cup \big(P(Y) \backslash \Delta_Y\big) \text{ else} \end{cases}$$

It follows from Lemma 11 that $F_\Delta^{-1}(A)$ is a measureable subset. Hence, F_Δ is measureable. □

As argued in the beginning of Section 2.3, we can define the transport maps $T(x, z)$ by formula (6), which is a Monge formulation of the constraints on the Kantorovich solution from (5): given any candidate transport map $T : X \times Z \to Y$, the corresponding transport plan is

$$\pi := (Id, T)\#\rho(x, z) \in P(X \times Z \times Y), \tag{37}$$

The marginal constraint $\pi_{XY} = \rho(x, z)$ is satisfied automatically, while the independence constraint $\pi_{YZ} = \pi_Y \otimes \pi_Z$ can be checked via Corollary 10.

Specifically, define the indicator function:

$$I(T) = \begin{cases} 0 \text{ if there exists } \mu \in P(Y) \text{ such that } (T, Proj_Z)\#\rho(x, z) = \mu \otimes v \\ \infty \text{ otherwise.} \end{cases}$$

Then, Corollary 10 implies that

$$\begin{aligned} I(T) &= \sup_{\psi_Y \in C_b(Y)} \sup_{\substack{\psi_Z \in C_b(Z) \\ \int \psi_Z dv = 0}} \int_{Y \times Z} \psi_Y(y)\psi_Z(z)\ d\tilde{T}\#\rho \\ &= \sup_{\psi_Y \in C_b(Y)} \sup_{\substack{\psi_Z \in C_b(Z) \\ \int \psi_Z dv = 0}} \int \psi_Y(T(x, z))\psi_Z(z)\ d\rho(x, z). \end{aligned}$$

It follows that we have a Monge formulation of the barycenter problem (5):

$$\begin{aligned} &\inf_{\substack{\text{Borel measurable} \\ T: X \times Z \to Y}} \int c(x, T(x, z))\ d\rho(x, z) + I(T) \\ =&\inf_{\substack{\text{Borel measurable} \\ T: X \times Z \to Y}} \sup_{\substack{\psi_Y \in C_b(Y) \\ \psi_Z \in C_b(Z) \\ \int \psi_Z dv = 0}} \int c(x, T(x, z)) - \psi_Y(T(x, z))\psi_Z(z)\ d\rho(x, z). \end{aligned} \tag{38}$$

Essentially, (38) is a minimization over couplings of the form (37), which are Kantorovich solutions concentrated on graphs. Therefore, (38) is bounded below by (5), which minimizes over general Kantorovich solutions. To finish the proof, it suffices to show that (5) can be achieved by some transport map T.

Let π be a Kantorovich solution of (5), which exists by Theorem 1, and let $\mu = \pi_Y$ be the corresponding barycenter. By Assumption 2, for v-almost all $z \in Z$, the unique optimal coupling between $\rho(x|z)$ and μ is concentrated on the graph of some transport map T_z. Define the map $T(x, z) := T_z(x)$. It follows that for v-almost all z,

$$\pi(x, y|z) = (Id, T_z)\#\rho(x|z)$$

or equivalently, for ρ-almost all (x, z),

$$\pi(y|x, z) = \delta_{T(x,z)}$$

Let F_Δ be the map given by Corollary 12. Given that $\pi(y|x,z)$ can be chosen as a measureable function from $X \times Z$ to $P(Y)$, we can conclude that

$$T(x,z) = F_\Delta\big(\pi(y|x,z)\big)$$

is a measureable function from $X \times Z$ to Y.

Hence,

$$(5) = \int c\ d\pi = \int c\ d\pi(y|x,z)d\rho(x,z) = \int c\big(x, T(x,z)\big)d\rho(x,z) \geq (38)$$

D Proof of Theorem 3

If the marginal $\rho(x)$ does not have finite second moment, then both sides of (8) are infinite: either $Var(\mu)$ or $W_2^2(\rho(x),\mu)$ must be infinite and Lemma 8 bounds $\int W_2^2(\rho(x|z),\mu)d\nu(z)$ below by $W_2^2(\rho(x),\mu)$. Hence, in the following proof we can assume that

$$\infty > \mathbb{E}_{\rho(x)}\big[\|x\|^2\big] = \int_{\mathbb{R}^d \times Z} \|x\|^2 d\rho(x,z) = \int_Z \mathbb{E}_{\rho(x|z)}\big[\|x\|^2\big]d\nu(z). \tag{39}$$

We first prove the special case where $\nu(z)$ is finitely-supported and that the conditionals $\rho(x|z)$ are absolutely continuous.

Denote the subset of $P(\mathbb{R}^d)$ that consists of probabilities measures with finite second moments by $P_2(\mathbb{R}^d)$. Denote the support of ν by $\{z_k\}_{k=1}^K \subseteq Z$. Denote the positive numbers $\nu(\{z_k\})$ by P_k and the conditionals $\rho(x|z_k)$ by $\rho_k(x)$. Then, the marginal $\rho(x)$ is the weighted sum $\sum_{k=1}^K P_k\rho_k$. Condition (39) implies that each $\rho_k \in P_2(\mathbb{R}^d)$.

Lemma 13. Given absolutely continuous measures $\rho_k \in P_2(\mathbb{R}^d)$ and weights $P_k > 0$ for $1 \leq k \leq K$, there exists a unique barycenter μ and it satisfies the discrete version of (8):

$$Var(\rho(x)) = Var(\mu) + \sum_{k=1}^K P_k W_2^2(\rho_k, \mu) \tag{40}$$

Proof. By Theorems 3.1 and 5.1 of [25], since ρ_k are absolutely continuous, there exists a unique barycenter $\mu(x)$ and it is also absolutely continuous. Then, Brenier's theorem (Theorem 2.12 [50]) implies that there is a unique optimal transport map T_k from each ρ_k to μ. The transport maps have the form $T_k = \nabla\psi_k$ for some convex functions ψ_k, and they are invertible almost everywhere: let ψ_k^* be the Legendre transform of ψ_k, then

$$\nabla\psi_k^* \circ \nabla\psi_k(x) = x,\ \nabla\psi_k \circ \nabla\psi_k^*(y) = y$$

for ρ_k-almost all x and μ-almost all y. Furthermore, $\nabla\psi_k^*$ serves as the optimal transport map from μ back to ρ_k.

Denote the mean of $\rho(x)$ by $\overline{x}$. Note that $\overline{x}$ is also the mean of the barycenter μ: let π be the (unique) Kantorovich solution given by Theorem 1, let X_k be the random variables of $\rho_k = \pi_{XZ}(x|z_k)$, and let Y be the random variable of $\mu = \pi_Y$. Define the mean $\overline{X} = \sum_{k=1}^K P_k X_k$. Then the barycenter problem's objective (5) becomes

$$\mathbb{E}\sum_{k=1}^K P_k||Y - X_k||^2 = \mathbb{E}||Y - \overline{X}||^2 + \sum_{k=1}^K P_k\mathbb{E}||\overline{X} - X_k||^2. \tag{41}$$

Since Y minimizes the objective, we must have $Y = \overline{X}$, so that $\mathbb{E}[Y] = \overline{x}$. Then

$$\begin{aligned} Var(\rho) = \int_{\mathbb{R}^d} \|x - \overline{x}\|^2 d\rho(x) &= \sum_{k=1}^K P_k \int \|(x - T_k(x)) + (T_k(x) - \overline{x})\|^2 d\rho_k(x) \\ &= \sum_{k=1}^K P_k \int \|x - T_k(x)\|^2 + \|T_k(x) - \overline{x}\|^2 + 2\langle x - T_k(x), T_k(x) - \overline{x}\rangle d\rho_k(x). \end{aligned}$$

The first term is exactly the total transport cost, while the second term is the barycenter's variance:

$$\sum_{k=1}^{K} P_k \int \|T_k(x) - \overline{x}\|^2 d\rho_k(x) = \sum_{k=1}^{K} P_k \int \|y - \overline{x}\|^2 d\mu(y) = Var(\mu).$$

Regarding the third term, we use the fact that $\nabla\psi_k^* \# \mu = \rho_k$ to obtain

$$\begin{aligned} \sum P_k \int \langle x - T_k(x), T_k(x) - \overline{x}\rangle d\rho_k(x) &= \sum P_k \int \langle \nabla\psi_k^*(y) - y, y - \overline{x}\rangle d\mu(y) \\ &= \int \langle \sum_{k=1}^{K} P_k \nabla\psi_k^*(y) - y, y - \overline{x}\rangle d\mu(y). \end{aligned}$$

Remark 3.9 from [3] shows that $\sum P_k \nabla\psi_k^*$ is exactly the identity, so the third term vanishes. It follows that formula (40) holds. □

Now we tackle the general case when $\rho(x,z)$ is an arbitrary probability measure over $\mathbb{R}^d \times Z$. Condition (39) implies that $\rho(x|z) \in P_2(\mathbb{R}^d)$ for v-almost every z, so without loss of generality, $\rho(x|z)$ can be seen as a random variable from Z to $P_2(\mathbb{R}^d)$. We denote its distribution by $\Omega(\eta)$, which belongs to $P(P_2(\mathbb{R}^d))$, the space of probability measures over $P_2(\mathbb{R}^d)$.

Abusing notation, we denote by $P_2(P_2(\mathbb{R}^d))$ the space of probabilities $\Omega'(\eta)$ on $(P_2(\mathbb{R}^d), W_2)$ with finite second moment, that is, for some (and thus any) $\rho_0 \in P_2(\mathbb{R}^d)$,

$$\int_{P_2(\mathbb{R}^d)} W_2^2(\rho_0, \eta) d\Omega'(\eta) < \infty.$$

Then $P_2(P_2(\mathbb{R}^d))$ can be equipped with the 2-Wasserstein metric. Condition (39) implies that our distribution Ω belongs to $P_2(P_2(\mathbb{R}^d))$: for any $\rho_0 \in P_2(\mathbb{R}^d)$,

$$\begin{aligned} \int_{P_2(\mathbb{R}^d)} W_2^2(\rho_0, \eta)\ d\Omega(\eta) &= \int \inf_{\substack{\pi \in P(\mathbb{R}^d \times \mathbb{R}^d) \\ \pi_1 = \rho_0, \pi_2 = \eta}} \int \|x - x'\|^2 d\pi(x, z')\ d\Omega(\eta) \\ &\leq \iint \|x - x'\|^2\ d\rho_0 \otimes \eta(x, x')\ d\Omega(\eta) \text{ by the trivial coupling} \\ &\leq \iint 2(\|x\|^2 + \|x'\|^2) d\rho_0(x) d\eta(x') d\Omega(\eta) \\ &\leq 2\mathbb{E}_{\rho_0(x)}[X^2] + 2\int_{P_2(\mathbb{R}^d)} \mathbb{E}_{\rho(x)}[X^2] d\Omega(\rho) \\ &\leq 2\mathbb{E}_{\rho_0(x)}[X^2] + 2\int_{Z} \mathbb{E}_{\rho(x|z)}[X^2] d\nu(z) \\ &< \infty \text{ by (39)}. \end{aligned}$$

By Theorem 6.18 of [51], both $(P_2(\mathbb{R}^d), W_2)$ and $(P_2(P_2(\mathbb{R}^d)), W_2)$ are Polish spaces, each of whose elements can be approximated by finitely-supported probability measures. Let $\{\Omega^n\}_{n=1}^{\infty} \subseteq P_2(P_2(\mathbb{R}^d))$ be a sequence of finitely-supported measures that converge to Ω in W_2. Then, each Ω^n can be expressed as

$$\Omega^n = \sum_{k=1}^{K^n} P_k^n \delta_{\rho_k^n},$$

where K^n is the size of the support of Ω^n, the positive numbers P_k^n are the weights, and $\delta_{\rho_k^n}$ is the Dirac measure at $\rho_k^n \in P_2(\mathbb{R}^d)$. Define the marginal ρ^n of Ω^n by

$$\rho^n = \sum_{k=1}^{K^n} P_k^n \rho_k^n. \tag{42}$$

It follows that $\rho^n \in P_2(\mathbb{R}^d)$.

In order to apply Lemma 13, we show that these ρ^n_k can be assumed to be absolutely continuous. A nice property of $(P_2(\mathbb{R}^d), W_2)$ is that absolutely continuous measures are dense in it: any measure in $(P_2(\mathbb{R}^d), W_2)$ can be approximated by finitely-supported measures, which can then be approximated by absolutely continuous measures using kernel smoothing. Thus, for each Ω^n, we can construct

$$\tilde{\Omega}^n = \sum_{k=1}^{K^n} P^n_k \delta_{\tilde{\rho}^n_k}, \text{ such that } W_2^2(\rho^n_k, \tilde{\rho}^n_k) < \frac{1}{nK^n}, \text{ so that } W_2^2(\Omega^n, \tilde{\Omega}^n) < \frac{1}{n},$$

so that $\tilde{\Omega}^n$ also converge to Ω in $(P_2(P_2(\mathbb{R}^d)), W_2)$. It follows that ρ^n are also absolutely continuous.

Now given that each Ω^n consists of absolutely continuous ρ^n_k, Lemma 13 implies that each Ω^n has a unique barycenter μ^n and it satisfies

$$Var(\rho^n) = Var(\mu^n) + \int_{P_2(\mathbb{R}^d)} W_2^2(\mu^n, \eta) d\Omega^n(\eta). \tag{43}$$

The following two lemmas show that ρ^n and μ^n enjoy good convergence properties.

Lemma 14. The marginal ρ^n converges to $\rho(x)$ in $(P_2(\mathbb{R}^d), W_2)$.

Proof. We apply condition (iv) of Theorem 7.12 from [50], which shows that for any Polish space X with metric d, a sequence η^n converges to η in $(P_2(X), W_2)$ if and only if

$$\lim_{n\to\infty} \int_X \psi d\eta^n = \int_X \psi d\eta \tag{44}$$

for any continuous function ψ that grows at most quadratically: $|\psi(x)| \leq C(1 + d(x_0, x)^2)$ for some $x_0 \in X$ and $C > 0$. Therefore, it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} \psi d\rho^n = \int \psi d\rho$$

for any ψ with a quadratic bound: $|\psi(x)| \leq C(1 + \|x\|^2)$ for some $C > 0$.

By (42), it is equivalent to

$$\lim_{n\to\infty} \int_{P_2(\mathbb{R}^d)} F_\psi(\eta) d\Omega^n(\eta) = \int F_\psi(\eta) d\Omega(\eta), \tag{45}$$

where $F_\psi(\eta) := \int \psi d\eta$. The function F_ψ is continuous on $(P_2(\mathbb{R}^d), W_2)$ by condition (44). The quadratic bound on ψ translates to a quadratic bound on F_ψ:

$$F_\psi(\eta) \leq C\big(W_2^2(\eta, \delta_0) + 1\big),$$

where δ_0 is the Dirac measure at 0.

Then we can apply Theorem 7.12 [50] on $(P_2(P_2(\mathbb{R}^d)), W_2)$, and (45) follows from the W_2 convergence of Ω^n to Ω. □

Lemma 15. A subsequence of $\{\mu^n\}$ converges in $(P_2(\mathbb{R}^d), W_2)$ to a barycenter μ of Ω.

Proof. First, the total transport cost from Ω^n to its barycenter μ^n can be computed through

$$c^n := \int_{P_2(\mathbb{R}^d)} W_2^2(\eta, \mu^n) d\Omega^n(\eta) = W_2^2(\delta_{\mu^n}, \Omega^n),$$

where the second W_2 belongs to $P_2(P_2(\mathbb{R}^d))$ and δ_{μ^n} is the Dirac measure on μ^n. Then, for any n, m,

$$\begin{aligned} c^n &\leq W_2^2(\delta_{\mu^m}, \Omega^n) \text{ since } \mu^n \text{ minimizes total transport cost} \\ &\leq c^m + W_2^2(\Omega^n, \Omega^m) \text{ by triangle-inequality} \\ \to& |c^n - c^m| \leq W_2^2(\Omega^n, \Omega^m). \end{aligned}$$

Since $W_2^2(\Omega^n, \Omega) \to 0$, the difference $W_2^2(\Omega^n, \Omega^m) \to 0$ as $n, m \to \infty$, so $|c^n - c^m| \to 0$. It follows that $\{c^n\}$ is a Cauchy sequence and thus converges.

Next, we establish some uniform bound on the decay of $\{\mu^n\}$ at infinity. We begin with a weak bound: by Lemma 8 and triangle inequality,

$$\begin{aligned} W_2^2(\rho(x), \mu^n) &\leq W_2^2(\Omega, \delta_{\mu^n}) \\ &\leq c^n + W_2^2(\Omega^n, \Omega). \end{aligned}$$

Denote by $\delta_0 \in P(\mathbb{R}^d)$ the Dirac measure at 0 and by $B_R \subseteq \mathbb{R}^d$ the open ball centered at 0 with radius R. By the triangle inequality, for any R,

$$\begin{aligned} W_2(\rho(x), \mu^n) &\geq |W_2(\mu^n, \delta_0) - W_2(\rho(x), \delta_0)| \\ &\geq \sqrt{\int_{\mathbb{R}^d - B_R} R^2 d\mu^n} - \sqrt{\int_{\mathbb{R}^d} \|x\|^2 d\rho(x)} \\ &\geq R\sqrt{\mu^n(\mathbb{R}^d - B_R)} - \sqrt{\mathbb{E}_{\rho(x)}[X^2]}. \end{aligned}$$

Combining the two inequalities, we obtain for any n and R,

$$\mu^n(\mathbb{R}^d - B_R) \leq \Big(\frac{\sqrt{\mathbb{E}_{\rho(x)}[X^2]} + \sqrt{c^n + W_2^2(\Omega^n, \Omega)}}{R}\Big)^2.$$

Since the second moment $\mathbb{E}_{\rho(x)}[X^2]$ is finite by (39) and $c^n, W_2^2(\Omega^n, \Omega)$ are convergent sequences, there exists some constant C large enough so that

$$\sup_n \mu^n(\mathbb{R}^d - B_R) \leq \frac{C}{R^2}. \tag{46}$$

An immediate consequence is that μ^n is uniformly tight. Then, Prokhorov's theorem implies that $\{\mu^n\}$ has a subsequence $\{\mu^{n_i}\}$ that converges weakly to some limit $\mu \in P(\mathbb{R}^d)$. By Kantorovich duality [51], the optimal transport cost $W_2^2(\cdot, \cdot)$ on $P(\mathbb{R}^d)$ can be expressed as a supremum over bounded continuous functions, and thus is lower semi-continuous in the topology of weak convergence of $P(\mathbb{R}^d)$. Thus,

$$\forall \eta \in P(\mathbb{R}^d),\ W_2^2(\mu, \eta) \leq \liminf_{n_i \to \infty} W_2^2(\mu^{n_i}, \eta).$$

In particular, by setting $\eta = \delta_0$, we have shown that μ has finite second moment: $\mu \in P_2(\mathbb{R}^d)$.

Now we prove that this subsequence μ^{n_i} converges to μ in the stronger topology of $(P_2(\mathbb{R}^d), W_2)$, and we use bootstrapping to improve the tail bound (46). Condition (ii) of Theorem 7.12 of [50] indicates that it is necessary and sufficient to prove that

$$\lim_{R \to \infty} \limsup_{n_i \to \infty} \int_{\mathbb{R}^d - B_R} \|x\|^2 d\mu^{n_i}(x) = 0. \tag{47}$$

We show that this condition holds for the whole sequence μ^n.

Recall the arguments of Lemma 13: Since ρ_k^n, μ^n are absolutely continuous for all n and $1 \leq k \leq K^n$, Brennier's theorem implies that the optimal transport maps $\nabla\psi_k^n, \nabla(\psi_k^n)^*$ between them are invertible almost everywhere and, Remark 3.9 of [3] indicates that $\sum P_k \nabla(\psi_k^n)^* = Id$. Then, μ^n-almost all x can be expressed as the convex combination $x = \sum P_k^n \nabla(\psi_k^n)^*(x)$. It follows from convexity that

$$\|x\|^2 \leq \sum_{k=1}^{K^n} P_k^n \|\nabla(\psi_k^n)^*(x)\|^2.$$

Then, $\nabla(\psi_k^n)^*\#\mu^n = \rho_k^n$ implies that,

$$\begin{aligned}\int_{\mathbb{R}^d - B_R} \|x\|^2 d\mu^n(x) &\leq \sum_{k=1}^{K^n} P_k^n \int_{\mathbb{R}^d - B_R} \|\nabla(\psi_k^n)^*(x)\|^2 d\mu^n(x) \\ &= \sum_{k=1}^{K^n} P_k^n \int_{\nabla(\psi_k^n)^*(\mathbb{R}^d - B_R)} \|x\|^2 d\rho_k^n(x). \end{aligned} \tag{48}$$

In the last line above, we are integrating the measure ρ_k^n restricted to the domain $\nabla(\psi_k^n)^*(\mathbb{R}^d - B_R)$. Equivalently, we are integrating over some measure $\tilde{\rho}_k^n$ (not necessarily a probability measure) such that $0 \leq \tilde{\rho}_k^n \leq \rho_k^n$ (setwise) and

$$\tilde{\rho}_k^n(\mathbb{R}^d) = \rho_k^n\big(\nabla(\psi_k^n)^*(\mathbb{R}^d - B_R)\big) = \mu^n(\mathbb{R}^d - B_R).$$

For convenience, for any R, n and $1 \leq k \leq K^n$, define the following collections of measures:

$$\begin{gathered} M_k^n(R) := \{\tilde{\rho}_k^n \in M^+(\mathbb{R}^d),\ \tilde{\rho}_k^n \leq \rho_k^n \text{ and } \tilde{\rho}_k^n(\mathbb{R}^d) \leq C/R^2\} \\ M^n(R) := \{\tilde{\rho}^n \in M^+(\mathbb{R}^d),\ \tilde{\rho}^n \leq \rho^n \text{ and } \tilde{\rho}^n(\mathbb{R}^d) \leq C/R^2\} \\ M(R) := \{\tilde{\rho} \in M^+(\mathbb{R}^d),\ \tilde{\rho} \leq \rho \text{ and } \tilde{\rho}(\mathbb{R}^d) \leq C/R^2\}, \end{gathered}$$

where $M^+(\mathbb{R}^d)$ is the set of nonnegative Borel measures, and the uniform upper bound C/R^2 comes from (46).

It follows that the restriction of ρ_k^n to $\nabla(\psi_k^n)^*(\mathbb{R}^d - B_R)$ belongs to $M_k^n(R)$ and

$$\int_{\nabla(\psi_k^n)^*(\mathbb{R}^d - B_R)} \|x\|^2 d\rho_k^n(x) \leq \sup_{\tilde{\rho}_k^n \in M_k^n(R)} \int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho}_k^n(x). \tag{49}$$

Given any choice of $\{\tilde{\rho}_k^n\}_{k=1}^{K^n}$, it is straightforward to show that the weighted sum

$$\tilde{\rho}^n := \sum_{k=1}^{K^n} P_k^n \tilde{\rho}_k^n$$

belongs to $M^n(R)$ and

$$\sum_{k=1}^{K^n} P_k^n \sup_{\tilde{\rho}_k^n \in M_k^n(R)} \int \|x\|^2 d\tilde{\rho}_k^n(x) \leq \sup_{\tilde{\rho}^n \in M^n(R)} \int \|x\|^2 d\tilde{\rho}^n. \tag{50}$$

For any n, since ρ^n is absolutely continuous, Brennier's theorem implies that there is an optimal transport map T^n such that $T^n\#\rho^n = \rho$. Given any $\tilde{\rho}^n \in M^n(R)$, it is straightforward to show that $\tilde{\rho} := T_n\#\tilde{\rho} \in M(R)$ and that the optimal transport cost

$$W_2(\tilde{\rho}^n, \tilde{\rho}) \leq W_2(\rho^n, \rho).$$

Denote by $\tilde{\delta}_0$ the Dirac measure at zero with the same mass as $\tilde{\rho}^n$. By the triangle inequality,

$$\begin{aligned} \sqrt{\int \|x\|^2 d\tilde{\rho}^n} = W_2(\tilde{\rho}^n, \tilde{\delta}_0) &\leq W_2(\tilde{\rho}, \tilde{\delta}_0) + W_2(\tilde{\rho}, \tilde{\rho}^n) \\ &\leq \sqrt{\int \|x\|^2 d\tilde{\rho}} + W_2(\rho, \rho^n) \end{aligned} \tag{51}$$

Combining the inequalities (48), (49), (50), (51), we obtain

$$\int_{\mathbb{R}^d - B_R} \|x\|^2 d\mu^n(x) \leq \Big[\sup_{\tilde{\rho} \in M(R)} \sqrt{\int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho}} + W_2(\rho, \rho^n)\Big]^2 \tag{52}$$

for all R and n.

Now, we show that the $\sup_{M(R)}$ term in (52) vanishes as $R \to \infty$. Recall the definition of $M(R)$: any $\tilde{\rho} \in M(R)$ can be seen as a measure obtained from ρ by removing $(1 - C/R^2)$-amount of mass. Since the integrand $\|x\|^2$ is strictly increasing in the radial direction, in order to maximize $\int \|x\|^2 d\tilde{\rho}$, we should first remove the mass of ρ that is closest to 0. To formalize this intuition, we define

$$r(R) := \inf\{r \geq 0 \mid \rho(\mathbb{R}^d - B_r) \leq C/R^2\}$$

where B_r is the open ball. It follows that

$$\sup_{\tilde{\rho} \in M(R)} \int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho} \leq \int_{\mathbb{R}^d - B_{r(R)}} \|x\|^2 d\rho \tag{53}$$

where the upper bound is always finite, since (39) implies that $\rho(x)$ has finite second moment. Then, there are two possibilities: First, suppose that $\rho(x)$ is compactly-supported, say, inside some ball B_{R_0}. Then, we obtain the trivial bound

$$\sup_{\tilde{\rho} \in M(R)} \int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho} \leq R_0^2 \cdot \frac{C}{R^2}$$

which goes to 0 as $R \to \infty$. Second, $\rho(x)$ has unbounded support, and thus $\rho(\mathbb{R}^d - B_r) > 0$ for all r. So the function $r(R)$ must go to infinity as $R \to \infty$. It follows that the upper bound (53) goes to zero.

Hence, we always have

$$\lim_{R \to \infty} \sup_{\tilde{\rho} \in M(R)} \sqrt{\int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho}} = 0$$

Meanwhile, Lemma 14 indicates that $W_2(\rho, \rho^n) \to 0$ as $n \to \infty$. Then, condition (47) follows from (52):

$$\lim_{R \to \infty} \limsup_{n \to \infty} \int_{\mathbb{R}^d - B_R} \|x\|^2 d\mu^n(x) \leq \Big[\lim_{R \to \infty} \sup_{\tilde{\rho} \in M(R)} \sqrt{\int_{\mathbb{R}^d} \|x\|^2 d\tilde{\rho}} + \limsup_{n \to \infty} W_2(\rho, \rho^n) \Big]^2 = 0,$$

and we conclude that the subsequence μ^{n_i} converges to μ in W_2.

Finally, we show that μ is a barycenter of Ω. For any $\tilde{\mu} \in P_2(\mathbb{R}^d)$ and any n_i,

$$c^{n_i} = W_2^2(\Omega^{n_i}, \delta_{\mu^{n_i}}) \leq W_2^2(\Omega^{n_i}, \delta_{\tilde{\mu}}). \tag{54}$$

Since $W_2(\Omega, \Omega^n) \to 0$ and $W_2(\delta_\mu, \delta_{\mu^{n_i}}) = W_2(\mu, \mu^{n_i}) \to 0$, the cost $W_2^2(\Omega^{n_i}, \delta_{\mu^{n_i}})$ converges to $W_2^2(\Omega, \delta_\mu)$ by triangle inequality. Then, (54) implies that

$$W_2^2(\Omega, \delta_\mu) \leq \lim_{n_i \to \infty} W_2^2(\Omega^{n_i}, \delta_{\tilde{\mu}}) = W_2^2(\Omega, \delta_{\tilde{\mu}}).$$

Since the inequality holds for all $\tilde{\mu} \in P_2(\mathbb{R}^d)$, the limit μ is a barycenter of Ω. □

Now, we can take the limit in n in equation (43). Taking a subsequence if necessary, Lemma 14 and Lemma 15 imply that $\rho^n \to \rho$ and $\mu^n \to$ a barycenter μ in W_2, while the functions Var and W_2^2 are continuous over W_2, so

$$Var(\rho) = Var(\mu) + W_2^2(\Omega, \delta_\mu)$$

which finishes the proof of formula (8).

E Proof of Theorem 5

We apply the arguments of Appendix D. Since $\rho(x)$ is assumed to have finite second moment, $\rho(x,z)$ can be converted to a distribution $\Omega \in P_2(P_2(\mathbb{R}^d))$. We seek to construct a sequence of finitely-supported measures Ω^n that converge to Ω in W_2, such that each $\Omega^n = \sum_{k=1}^{K^n} P_k \delta_{\rho_k^n}$ and each ρ_k^n is a non-degenerate Gaussian (i.e. the covariance $S(z)$ is positive-definite). Fix some $\rho_0 \in \text{supp } \Omega$. For each n, let $C^n \subseteq P_2(\mathbb{R}^d)$ be a compact subset such that

$$\int_{P_2(\mathbb{R}^d)-C^n} W_2^2(\rho_0, \eta) d\Omega(\eta) < \frac{1}{n}.$$

Let $\{B(\rho_k^n, 1/2n)\}_{k=1}^{K^n}$ be a finite cover of C^n by open balls, where $\rho_k^n \in \text{supp } \Omega$ and thus are Gaussians. If ρ_k^n is degenerate, then replace it with some non-degenerate Gaussian $\tilde{\rho}_k^n \in B(\rho_k^n, 1/2n)$ and use the ball $B(\tilde{\rho}_k^n, 1/n)$. Define the disjoint cover $\{U_k^n\}_{k=1}^{K^n}$ by

$$U_k^n = B(\rho_k^n, 1/n) - \bigcup_{h=1}^{k-1} B(\rho_h^n, 1/n).$$

Define a map F^n on $P_2(\mathbb{R}^d)$ that sends each U_k^n to ρ_k^n and everything else to ρ_0. Define

$$\Omega^n = F^n \# \Omega = \sum_{k=1}^{K^n} \Omega(U_k^n) \delta_{\rho_k^n} + \Big(1 - \sum_{k=1}^{K^n} \Omega(U_k^n)\Big) \delta_{\rho_0}.$$

It follows that

$$W_2^2(\Omega, \Omega^n) \leq \int W_2^2(\eta, F^n(\eta)) d\Omega(\eta) < \frac{2}{n}$$

So $\Omega^n \to \Omega$ in W_2.

Denote ρ_k^n by $\mathcal{N}(\overline{x}_k^n, S_k^n)$, and denote the X-marginal of Ω^n by $\rho^n = \sum P_k \rho_k^n$. Now, Theorem 6.1 of [3] implies that Ω^n has a unique barycenter μ^n, which is a Gaussian whose covariance S^n satisfies

$$S^n = \sum_{k=1}^{K^n} P_k \sqrt{\sqrt{S^n} S_k^n \sqrt{S^n}} = \int \sqrt{\sqrt{S^n} S(\eta) \sqrt{S^n}} d\Omega^n(\eta), \tag{55}$$

where $S(\rho)$ denotes the covariance of ρ. Also, the argument (41) implies that the mean $\overline{x}^n$ of μ^n satisfies

$$\overline{x}^n = \sum P_k \overline{x}_k^n = \mathbb{E}_{\rho^n(x)}[x].$$

Lemma 14 implies that ρ^n converges to the marginal $\rho(x)$ in W_2, and Taking a subsequence if necessary, Lemma 15 implies that μ^n converges to a barycenter μ of Ω in W_2. Since the set of Gaussian distributions is closed in W_2, this barycenter μ must be some Gaussian $\mathcal{N}(\overline{x}, S)$. By Theorem 7.12 of [50], the covariance function $S(\eta)$ is continuous over $\eta \in (P_2(\mathbb{R}^d), W_2)$, so S^n converges to S. Then, we can take the limit on both sides of (55):

$$\begin{aligned} S = \lim_{n\to\infty} S^n &= \lim_{n\to\infty} \int \sqrt{\sqrt{S} S(\eta) \sqrt{S}} d\Omega^n(\eta) + O\big(\sqrt{\|S - S^n\|_{op}}\big) \cdot \int \sqrt{S(\eta)} d\Omega^n(\eta) \\ &= \int \sqrt{\sqrt{S} S(\eta) \sqrt{S}} d\Omega(\eta). \end{aligned}$$

Similarly,

$$\overline{x} = \lim_{n\to\infty} \int x d\rho^n(x) = \int x d\rho(x).$$

Finally, suppose that the set of $z \in Z$ such that $\rho(x|z)$ is a non-degenerate Gaussian (and thus, absolutely continuous) has positive measure. Lemma 3.2.1 of [38] implies that for each such $\rho(x|z)$, the optimal transport cost $\mu \mapsto W_2^2(\mu, \rho(x|z))$ is strictly convex. Then, the total transport cost $\mu \mapsto \int W_2^2(\mu, \rho(x|z)) dv$ is strictly convex and the barycenter is unique.

F Saddle point algorithms

Let $\inf_\tau \sup_\xi L(\tau, \xi)$ be a min-max problem. For convenience, denote

$$J = \begin{pmatrix} I_{dim\tau} & \\ & -I_{dim\xi} \end{pmatrix}, \; w = \begin{bmatrix} \tau \\ \xi \end{bmatrix}$$

Then, the OMD algorithm [32] (using Euclidean squared distance) becomes,

```
Parameters: Learning rates $\eta^n$.
for $n \leftarrow 1, 2, \ldots$ do
    Compute the waiting state $\tilde{w} \leftarrow w^n - \eta^n J \nabla L(w^n)$
    Actual update $w^{n+1} \leftarrow w^n - \eta^n J \nabla L(\tilde{w})$
end
return $w^\infty$
```

Algorithm 1: Optimistic mirror descent

We present the QITD algorithm [14] in its data-based setting, such that L is estimated from samples, just as in (13),

```
Parameters: Iteration number $T$. Batch size $M$. Initial learning rate $\eta^0$. Decay rate
  $\gamma \in (0,1)$. Stopping threshold $\epsilon \ll 1$. Increase factor $\beta > 0$. Maximum learning rate $\eta_{\max}$.
Data: Sample $X = \{x_i\}_{i=1}^N$.
Initialize quasi Newton matrix $B^1 \leftarrow J$.
for $n \leftarrow 1$ to $T$ do
    Randomly sample a minibatch $X^n$ of size $M$.
    Compute gradient $g^n \leftarrow \nabla L(w^n \; |X^n)$.
    Initialize learning rate $\eta^n \leftarrow \eta^{n-1}$.
    Compute update $w^{n+1} \leftarrow w^n - \eta^n B^n g^n$.
    while $\eta^n > \epsilon \eta^{n-1}$ and the anticipatory constraint

        $$L(\tau^{n+1}, \xi^n \; |X^n) \leq L(\tau^{n+1}, \xi^{n+1} \; |X^n) \leq L(\tau^n, \xi^{n+1} \; |X^n) \quad (56)$$

     is not satisfied do
        Line search $\eta^n \leftarrow \gamma \eta^n$.
        Update $w^{n+1} \leftarrow w^n - \eta^n B^n g^n$.
    end
    if constraint (56) has been satisfied then
        Increase learning rate $\eta^n \leftarrow \max((1+\beta)\eta^n, \eta_{\max})$
    end
    Compute new gradient $g^{n+1} \leftarrow \nabla L(w^{n+1} \; |X^n)$.
    Rank-one update of $B^n$:

        $$s \leftarrow J g^{n+1} - B^n g^n$$
        $$\alpha \leftarrow \frac{\|s\|^2}{\langle g^n, s \rangle}$$
        $$\alpha \leftarrow sign(\alpha) \min\big(|\alpha|, 1\big)$$
        $$B^{n+1} \leftarrow B^n + \alpha \frac{s \cdot s^T}{\|s\|^2}$$

end
return $w^T$
```

Algorithm 2: Stochastic quasi implicit twisted descent

The algorithm has time complexity $O(TD^2)$, where $D = dim\tau + \dim \xi$. With fixed decay rate γ and threshold ϵ, the anticipatory constraint (56) contributes a multiplicative factor to running time. The quasi Newton matrix B^n can be replaced by a list of its rank-one updates: $\{s^n, \alpha^n\}_{n=1}^T$, changing the time complexity to $O(T^2 D)$.

G Implementation details

All BaryNet models implemented in Section 4 are based on neural networks, and the network architectures are summarized below:

Experiment	$z(x)$	Transport residual	ψ^Y	ψ^Z
§4.1	-	$3 \to 7 \to 7 \to 2$	$2 \to 6 \to 6 \to 1$	$1 \to 5 \to 1$
§4.2 climate	$56 \to 5 \to 5 \to 1$	$57 \to 2 \to 56$	$56 \to 14 \to 1$	$1 \to 5 \to 1$
§4.2 seismic	$2 \to 6 \to 6 \to 1$	$3 \to 7 \to 7 \to 7 \to 2$	$2 \to 5 \to 5 \to 5 \to 1$	$1 \to 5 \to 5 \to 1$
§4.3	-	$5 \to 9 \to 9 \to 1$	$1 \to 5 \to 5 \to 1$	$4 \to 8 \to 1$
§4.4	-	$3 \to 25 \to 3 \to 25 \to 3$	$3 \to 25 \to 3 \to 25 \to 1$	$\mathbb{R}^3$

Table 1: BaryNet architectures for experiments in Section 4. "Transport residual" refers to the residual part of the transport map (See Section 3.1.1), and the inverse transport maps have the same architecture. Unless specified in the comments below, all activation functions are ReLU.

The training schemes are based on either OMD or QITD, and their parameters are summarized below:

Experiment	Algorithm	T	M	μ^0	γ	ϵ	β	$\eta_{\max}$
§4.1	QITD	10^4	full	4×10^{-3}	0.75	10^{-3}	0.1	2×10^{-2}
§4.2 climate	OMD	5×10^4	10^3	10^{-5}	-	-	-	-
§4.2 seismic	OMD	2×10^4	full	10^{-3}	-	-	-	-
§4.3	QITD	10^4	2400	2×10^{-3}	0.75	10^{-3}	0.1	10^{-2}
§4.4	OMD	10^6	10^3	10^{-3}	-	-	-	-

Table 2: Training parameters. The notations are based on Appendix F: For QITD, we have iteration number T, batch size M, initial learning rate η^0, decay rate γ, stopping threshold ϵ, increase factor β, and maximum learning rate $\eta_{\max}$. For OMD, the learning rate is fixed $\mu \equiv \mu^0$. If M =full, then gradient descent is based on the population loss over the entire sample, instead of the minibatch loss.

The training of the inverse transport map (19) are based on either SGD or Adam:

Experiment	Algorithm	T	M	μ
§4.1	SGD	2×10^4	full	5×10^{-2}
§4.3	SGD	10^4	24×365	5×10^{-3}
§4.4	Adam	10^4	3000	10^{-3}

Table 3: Training parameters: iteration number T, batch size M, and learning rate η. We did not use weight decay or momentum.

For table 1, if the latent variable z is continuous, then the transport map has the form $T(x, z) : \mathbb{R}^{d+k} \to \mathbb{R}^d$. If z is discrete, then we model each $T_z(x) : \mathbb{R}^d \to \mathbb{R}^d$ by networks with the same architecture. Regarding ψ^Z and $z(x)$, the affine map at the last layer is always bias-free, in order to reduce unnecessary parameters. The reason is that any additive constant in ψ^Z would be removed by (14), while the latent variables z are invariant under permutation so a constant has no effect on the assignment $z(x)$ (Section 3.1.2)

Here are the details of each experiment:

§4.2 climate The daily temperature data is taken from NOAA [35]. We chose the 56 stations with the fewest missing values, and any x_i with missing entry is discarded.

Training in high dimensions becomes more unstable, so we follow the technique introduced by the DCGAN paper [40]: we applied batch normalization to all hidden layers of all networks and also applied leaky ReLU (with negative slope 0.1) to the test functions ψ^Y, ψ^Z and labeling function $z(x)$.

We have enforced Lipschitz continuity in $z(x)$, by clamping its parameters to be within $[-0.1, 0.1]$ after each update step.

§4.2 seismic The earthquake data is taken from [49] and scaled by $\pi/180$ to yield spherical coordinates. We applied batch normalization to all hidden layers of $T(x, z), \psi^Y, \psi^Z, z(x)$ and also applied leaky ReLU (with negative slope 0.1) to ψ^Y, ψ^Z and $z(x)$.

§4.3 The hourly temperature data is taken from [36]. We chose Ithaca, NY and the time range Jan 1 2007 to Dec 31 2016 because there are few missing values, which are filled by linear interpolation.

§4.4 Each of the RGB color channels has range $[0, 1]$. Theoretically, the transported distributions $S_j \circ T_k \# \rho_k$ are always supported in $[0, 1]^3$. Nevertheless, the computed result is only an approximation to the true distribution, and sometimes a few points are mapped outside of $[0, 1]^3$, so we project them back.