SCALABLE UNBALANCED OPTIMAL TRANSPORT USING GENERATIVE ADVERSARIAL NETWORKS

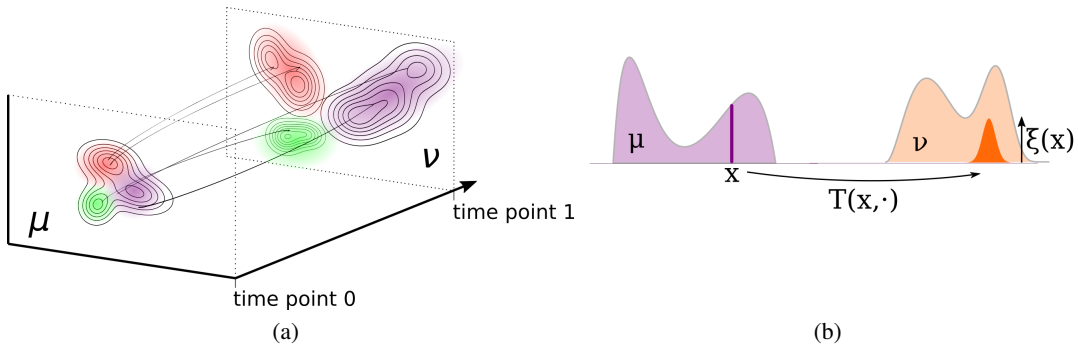
Anonymous authors
Paper under double-blind review

ABSTRACT

Generative adversarial networks (GANs) are an expressive class of neural generative models with tremendous success in modeling high-dimensional continuous measures. In this paper, we present a scalable method for unbalanced optimal transport (OT) based on the generative-adversarial framework. We formulate unbalanced OT as a problem of simultaneously learning a transport map and a scaling factor that push a source measure to a target measure in a cost-optimal manner, and propose a new algorithm based on stochastic alternating gradient updates, similar in practice to GANs. We also provide theoretical justification for this formulation, showing that it is closely related to an existing static formulation by Liero et al. (2018), and perform numerical experiments demonstrating how this methodology could be applied to population modeling.

1 INTRODUCTION

We consider the problem of unbalanced optimal transport: given two measures, find a cost-optimal way to transform one measure to the other using a combination of mass variation and transport. Such problems arise, for example, when modeling the transformation of a source population into a target population (Figure 1a). In this setting, one needs to model mass transport to account for the features that are evolving, as well as local mass variations to allow sub-populations to become more or less prominent in the target population (Schiebinger et al., 2017).



(a) (b)

Figure 1: (a) Illustration of the problem of modeling the transformation of a source population μ to a target population ν. In this example, one sub-population is growing more rapidly than the others. (b) Schematic of our formulation of unbalanced optimal transport. The objective is to learn a (probabilistic) transport map T (for transporting mass) and scaling factor ξ (for mass variation) to push the source μ to the target ν.

Classical optimal transport (OT) considers the problem of pushing a source to a target distribution in a way that is optimal with respect to some transport cost without allowing for mass variations. Modern approaches are based on the Kantorovich formulation (Kantorovich, 1942), which seeks the optimal probabilistic coupling between measures and can be solved using linear programming methods for discrete measures. Cuturi (2013) showed that regularizing the objective using an entropy term allows the dual problem to be solved more efficiently using the Sinkhorn algorithm. In the continuous setting, several stochastic methods based on this dual objective have been proposed (Genevay et al., 2016; Seguy et al., 2017). Optimal transport maps have seen applications to many areas, such as computer graphics (Ferradans et al., 2014; Solomon et al., 2015) and domain adaptation (Courty et al., 2014; 2017). In many applications where a transport cost is not available, transport maps can also be learned using generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014), which push a source distribution to a target distribution by training against an adversary. Numerous transport problems in image translation (Mirza & Osindero, 2014; Zhu et al., 2017; Yi et al., 2017), natural language translation (He et al., 2016), domain adaptation (Bousmalis et al., 2017) and biological data integration (Amodio & Krishnaswamy, 2018) have been tackled using variants of GANs, with strategies such as conditioning or cycle-consistency employed to enforce correspondence between original and transported samples. However, all these methods conserve mass between the source and target and therefore cannot handle mass variation. This limits applications to large-scale problems in population modeling, particularly in computational biology (Schiebinger et al., 2017).

Several formulations have been proposed for extending the theory of OT to the setting where the measures can have unbalanced masses (Chizat et al., 2015; 2018; Kondratyev et al., 2016; Liero et al., 2018; Frogner et al., 2015). In terms of numerical methods, a class of scaling algorithms (Chizat et al., 2016) that generalize the Sinkhorn algorithm for balanced OT have been developed for approximating the solution to *optimal entropy-transport* problems; this formulation of unbalanced OT by Liero et al. (2018) corresponds to the Kantorovich OT problem in which the hard marginal constraints are relaxed using divergences to allow for mass variation. In practice, these algorithms have been used to approximate unbalanced transport plans between discrete measures for applications such as computer graphics (Chizat et al., 2016), tumor growth modeling (Chizat & Di Marino, 2017) and computational biology (Schiebinger et al., 2017). However, optimal entropy-transport allows mass variation without explicitly modeling it, and there are currently no methods that can perform large-scale unbalanced transport between high-dimensional continuous measures.

Contributions. Inspired by the recent successes of GANs for high-dimensional transport problems, we present a novel framework for unbalanced optimal transport that directly models mass variation in addition to transport. Concretely, our contributions are the following:

- We formulate unbalanced OT as a problem of simultaneously learning a transport map and a scaling factor that push a measure μ to a measure ν in a cost-optimal manner (Figure 1b) and show how our formulation relates to the formulation by Liero et al. (2018), which has desirable theoretical properties.
- We propose scalable methodology for simultaneously learning the transport map and scaling factor. Our derivation is based on a variational inequality for divergences, which results in a stochastic alternating gradient descent method similar in practice to GANs (Goodfellow et al., 2014)
- We demonstrate in practice how our methodology can be applied towards population modeling using the MNIST and USPS handwritten digits datasets and the CelebA dataset.

2 Preliminaries and Related Work

Notation. Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$ be topological spaces and let $\mathcal{B}$ denote the Borel σ-algebra. Let $\mathscr{M}_+^1(\mathcal{X}), \mathscr{M}_+(\mathcal{X})$ denote respectively the space of probability measures and finite non-negative measures over $\mathcal{X}$. For a joint measure $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$, we let $\pi^{\mathcal{X}}_{\#}\gamma$ and $\pi^{\mathcal{Y}}_{\#}\gamma$ denote its marginals with respect to $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Optimal transport (OT) addresses the problem of transporting between measures in a cost-optimal manner. The original Monge formulation of OT searches over deterministic transport maps (Monge, 1781), while modern approaches are based on the Kantorovich relaxation (Kantorovich, 1942), which searches over probabilistic transport plans. Specifically, given $\mu \in \mathscr{M}_+^1(\mathcal{X}), \nu \in \mathscr{M}_+^1(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, Kantorovich OT seeks a joint measure $\gamma \in \mathscr{M}_+^1(\mathcal{X} \times \mathcal{Y})$ subject to $\pi^{\mathcal{X}}_{\#}\gamma = \mu$ and $\pi^{\mathcal{Y}}_{\#}\gamma = \nu$ minimizing

$$W(\mu, \nu) := \inf_{\gamma} \int c(x, y)\, d\gamma(x, y), \tag{1}$$

For discrete measures μ, ν, (1) can be solved as a linear program. Cuturi (2013) showed that introducing entropic regularization results in a simpler dual optimization problem that can be solved using the Sinkhorn algorithm. More recently, Genevay et al. (2016) and Seguy et al. (2017) proposed stochastic algorithms based on entropic regularization that extend to continuous measures. Arjovsky et al. (2017) proposed a stochastic algorithm based on a simpler dual formulation for computing the 1-Wasserstein distance.

Unbalanced OT. Several formulations that extend classical OT to handle mass variation have been proposed (Chizat et al., 2015; 2018; Kondratyev et al., 2016), the one most related to our work being *optimal-entropy transport* (Liero et al., 2018). This formulation is obtained by relaxing the marginal constraints of (1) using divergences as follows: Given two positive measures $\mu \in \mathscr{M}_+(\mathcal{X})$ and $\nu \in \mathscr{M}_+(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, optimal entropy-transport finds a measure $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$ that minimizes

$$W_{ub}(\mu, \nu) := \inf_{\gamma} \int c(x, y)\, d\gamma(x, y) + D_{\psi_1}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\psi_2}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu), \tag{2}$$

where D_{ψ_1} and D_{ψ_2} are ψ-divergences induced by ψ_1, ψ_2. The ψ-divergence between non-negative finite measures P, Q over $T \subset \mathbb{R}^d$ induced by a lower semi-continuous, convex entropy functions $\psi : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is defined as

$$D_{\psi}(P|Q) := \psi'_{\infty} P^{\perp}(T) + \int_T \psi\left(\frac{dP}{dQ}\right) dQ \tag{3}$$

where $\psi'_{\infty} := \lim_{r\to\infty} \frac{\psi(r)}{r}$ and $\frac{dP}{dQ}Q + P^{\perp}$ is the Lebesgue decomposition of P with respect to Q. Note that mass variation is allowed since the marginals of γ are not constrained to be μ and ν. In terms of numerical methods, while practical algorithms for the continuous setting are completely missing, the state-of-the-art in the discrete setting is a class of iterative scaling algorithms (Chizat et al., 2016) that generalizes the Sinkhorn algorithm for computing regularized OT plans (Cuturi, 2013). The key idea is to introduce an entropic regularization term into the objective so that the dual problem can be solved using Dykstra's algorithm (Chizat et al., 2016). In practice, this algorithm can be used to approximate unbalanced OT plans between discrete measures or low-dimensional continuous measures that can be discretized.

3 Proposed Formulation and Method

We propose an alternative formulation of unbalanced OT, in which the objective is to learn a (probabilistic) *transport map* T and *scaling factor* ξ to push μ to ν in a cost-optimal manner (Figure 1b). Let $c_1 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ denote the transport cost and $c_2 : \mathbb{R}^+ \to \mathbb{R}^+$ denote the cost of scaling mass. Let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ be a probability space[1]. Formally, we seek measurable functions $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and $\xi : \mathcal{X} \to \mathbb{R}^+$ minimizing

$$L(\mu, \nu) := \inf_{T,\xi} \int \int c_1(x, T(x, z)) d\lambda(z)\xi(x) d\mu(x) + \int c_2(\xi(x)) d\mu(x), \tag{4}$$

[1] If the probability space only contains one event, then the problem simplifies to a deterministic map T

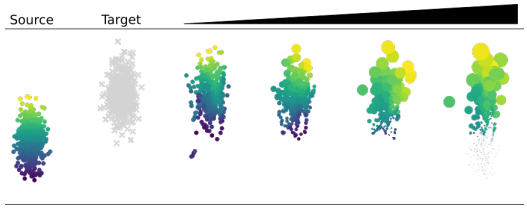


Figure 2: Illustration of unbalanced OT on synthetic dataset.

subject to

$$\nu(B) = \nu'_{T,\xi}(B) := \int \lambda(T_x^{-1}(B))\xi d\mu, \quad \forall B \in \mathcal{B}(\mathcal{Y}), \tag{5}$$

where $T_x(z) : z \mapsto T(x, z)$. Concretely, the first and second terms of (4) penalize the cost of mass transport and variation respectively, and the constraint (5) ensures that (T, ξ) pushes μ to ν exactly.

Intuitively, (4) describes a spectrum of unbalanced transport problems that interpolate between pure transport problems and pure mass variation problems. As an example, consider the unbalanced transport scenarios illustrated in Figure 2. The source and target measures are illustrated in Columns 1 and 2 respectively. When the cost of mass transport, c_1, is low compared to the cost of mass variation, c_2, the solution of (4) corresponds to a pure transport scheme, i.e. the transport map T translates the source distribution upward to the target distribution, and the scaling factor ξ is fixed to 1 and has no effect (Column 3). As the cost of mass transport is *increased* relative to the cost of mass variation, it becomes more costly to transport mass and less costly to change the amount of mass. Consequently, the solution of (4) tends more and more towards transporting only a small region of mass from the source to the target and growing this mass ($\xi > 1$), while mass that is farther from the target is allowed to shrink ($\xi < 1$) (Columns 4-6).

Relaxation. The form of (4) suggests that stochastic gradient methods could be used for optimization by sampling from $\mu \times \lambda$, but it would be challenging to satisfy the constraint (5) during the updates. In what follows, we consider a relaxation of (4) using a divergence penalty D_ψ in place of the hard constraint (5):

$$L_\psi(\mu, \nu) := \inf_{T,\xi} \int \int c_1(x, T(x, z))d\lambda(z)\xi(x)d\mu(x) + \int c_2(\xi(x))d\mu(x) + D_\psi(\nu'_{T,\xi}|\nu). \tag{6}$$

The divergence can be rewritten using the following variational inequality:

Lemma 3.1. *For non-negative finite measures P, Q over $T \subset \mathbb{R}^d$, it holds that*

$$D_\psi(P|Q) \geq \sup_{f \in \mathcal{F}} \int f dP - \psi^*(f)dQ \tag{7}$$

where $\mathcal{F}$ is a subset of measurable functions $\{f : T \to (-\infty, \psi'_\infty]\}$. Equality holds if and only if $\exists f \in \mathcal{F}$ such that (i) the restriction of f to the support of Q belongs to the subdifferential of $\psi(\frac{dP}{dQ})$, i.e. the Radon-Nikodym derivative of P with respect to Q, and (ii) $f = \psi'_\infty$ over the support of $P^\perp$.

Substituting (7) into (6), we obtain

$$\begin{aligned} L_\psi(\mu, \nu) = \inf_{T,\xi} \sup_{f \in \mathcal{F}} & \int \int c_1(x, T(x, z))d\lambda(z)\xi(x)d\mu(x) + \int c_2(\xi(x))d\mu(x) \\ & + \int \int f(T(x, z))d\lambda(z)\xi(x)d\mu(x) - \int \psi^*(f(y))d\nu(y) \end{aligned} \tag{8}$$

where $\mathcal{F}$ is a function class satisfying the equality conditions of Lemma 3.1. The objective in (8) can now be optimized using alternating gradient updates after parameterizing T, ξ, and f with neural networks, similar in practice to GANs as shown in Algorithm 1. We assume that one has access to samples from μ, ν, and in the setting where μ, ν are not normalized, then samples to the normalized measures $\tilde{\mu}, \tilde{\nu}$ as well as the normalization constants c_μ, c_ν. These are reasonable assumptions for practical applications: for example, in a biological assay, one might collect c_μ cells from time point 1 and c_ν cells from time point 2. In this case, the samples are the measurements of each cell and the normalization constants are c_μ, c_ν.

Algorithm 1 Generative adversarial method for unbalanced OT

Input: Initial parameters θ, ϕ, ω; step size η; normalized measures $\tilde{\mu}$, $\tilde{\nu}$, constants c_μ, c_ν
Output: Updated parameters θ, ϕ, ω
while (θ, ϕ, ω) not converged **do**
Sample $x_1, \cdots, x_n$ from $\tilde{\mu}$, $y_1, \cdots, y_n$ from $\tilde{\nu}$, $z_1, \cdots, z_n$ from λ

$$\ell(\theta, \phi, \omega) := \frac{1}{n}\sum_{i=1}^{n}[c_\mu c_1(x_i, T_\theta(x_i, z_i))\xi_\phi(x_i) + c_\mu c_2(\xi_\phi(x_i)) + c_\mu \xi_\phi(x_i) f_\omega(T_\theta(x_i, z_i)) - c_\nu \psi^*(f_\omega(y_i))] \tag{9}$$

Update ω by gradient descent on $-\ell(\theta, \phi, \omega)$
Update θ, ϕ by gradient descent $\ell(\theta, \phi, \omega)$
end while

Theoretical properties. While our objective in (6) is partly motivated by the generative-adversarial framework, we now show that it has a sound theoretical basis. For all statements, we assume $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ is an atomless probability space. Our first result points to the close relation between (6) and the optimal entropy-transport problem (2) proposed by Liero et al. (2018).

Lemma 3.2. *Define*

$$\tilde{W}_{c,\psi_1,\psi_2}(\mu, \nu) := \inf_{\gamma \in \mathcal{G}} \int c d\gamma + D_{\psi_1}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\psi_2}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \tag{10}$$

where $\mathcal{G} = \{\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y}) | \gamma(\mathcal{X}_\perp \times \mathcal{Y}) = 0\}$, and $\mathcal{X}_\perp = \mathcal{X} \backslash supp(\mu)$. If c_2 is convex and lower semi-continuous, then $L_\psi(\mu, \nu) = \tilde{W}_{c_1,c_2,\psi}(\mu, \nu)$.

One implication of this result is that for certain cost functions and divergences, L_ψ defines a proper metric between positive measures μ and ν. For instance, taking c_2, ψ to be entropy functions corresponding to the KL-divergence and $c_1 = \log \cos^2_+(d(x, y))$, then $L_\psi(\mu, \nu)$ corresponds to the Hellinger-Kantorovich (Liero et al., 2018) or the Wasserstein-Fisher-Rao (Chizat et al., 2018) metric between positive measures μ and ν.

Next, we give sufficient conditions under which minimizers of (6) exist and are unique. For a given μ, ν, define the product measure generated by any (T, ξ) by

$$\gamma(C) := \int\int \mathbb{1}_C(x, T(x, z)) d\lambda(z)\xi(x)d\mu(x), \quad \forall C \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y}).$$

Proposition 3.3. *Suppose $L_\psi(\mu, \nu) < \infty$ and c_2 is convex and lower semi-continuous. If (i) c_1 has compact sublevel sets in $\mathcal{X} \times \mathcal{Y}$ and $c'_{2\infty} + \psi'_\infty > 0$, or (ii) $c'_{2\infty} = \psi'_\infty = \infty$ then a minimizer of $L_\psi(\mu, \nu)$ exists. Moreover, if c_2, ψ are strictly convex, $\psi'_\infty = \infty$ and c_1 satisfies Corollary 3.6 (Liero et al., 2018), then the product measure γ generated by any minimizer of $L_\psi(\mu, \nu)$ is unique.*

Under these assumptions, we can formalize the notion that for a sufficiently large and appropriate choice of divergence penalty, the solution to (6) approaches the solution of the original problem (4). Then we have the following convergence result:

Theorem 3.4. *Suppose c_1, c_2, ψ satisfy the existence assumptions of Proposition 3.3. Furthermore, let ψ be uniquely minimized at $\psi(1) = 0$. Then for a sequence $0 < \zeta^1 < \cdots < \zeta^k < \cdots$ diverging to ∞ indexed by k, $\lim_{k\to\infty} L_{\zeta^k\psi}(\mu, \nu) = L(\mu, \nu)$. Additionally, let γ^k be the product measure generated by a minimizer of $L_{\zeta^k\psi}(\mu, \nu)$. If $L(\mu, \nu) < \infty$, then up to extraction of a subsequence, γ^k converges weakly to γ, the product measure generated by a minimizer of $L(\mu, \nu)$.*

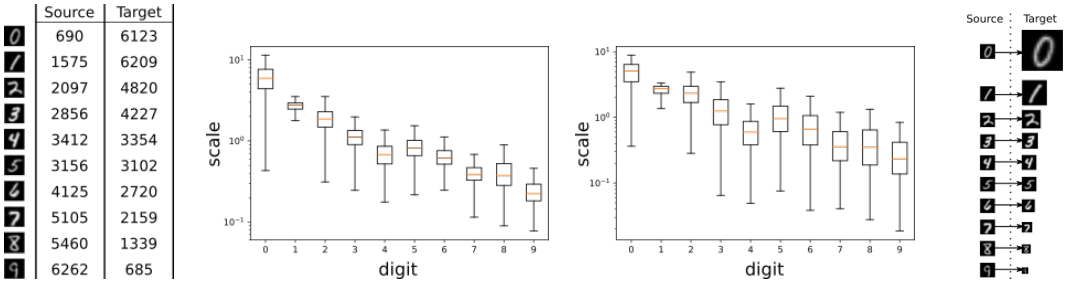
We emphasize that for this convergence result to hold, ψ must be uniquely minimized at $\psi(1) = 0$. In practice, this means that many ψ-divergences that can be used for matching probability measures, including the original GAN objective corresponding to $\psi(s) = s \log s - (s + 1) \log(s + 1)$ (Nowozin et al., 2016), cannot be used for unbalanced transport. Table 1 in the Appendix provides some examples of common ψ-divergences that satisfy Theorem 3.4, along with their corresponding entropy functions and convex conjugates.

By a similar argument as for Theorem 3.4, it can be shown that the solution to (6) converges to the solution for optimal transport (1) if we increase the cost of mass variation to ∞ along with the divergence penalty. Therefore, balanced OT is an extreme case of our formulation of unbalanced OT.

Relation to other GAN variants. Since the original GAN paper by Goodfellow et al. (2014), numerous variants of GANs have been proposed for learning transport maps between distributions (Zhu et al., 2017; Yi et al., 2017; Amodio & Krishnaswamy, 2018). These methods generally assume that the source and target measures are normalized probability measures and cannot handle mass variation. Algorithm 1 proposes a generalization of this strategy that can be used for unbalanced transport, based on the idea of learning both a transport map and a scaling factor. Our derivation of Algorithm 1 uses the variational inequality for divergences presented in Lemma 3.1, which enables us to minimize many divergences using alternating gradient updates similar to GAN training. A similar inequality with stronger assumptions proposed by Nguyen et al. (2008) was used by Nowozin et al. (2016) to propose a large class of generative models that generalize GANs. An important distinction of our work is that we consider divergence minimization between measures that are not probability measures, which requires additional assumptions on the choice of divergence (i.e. Theorem 3.4).

The problem of learning a scaling factor (or weighting factor) that "balances" measures μ and ν also arises in causal inference. Generally, μ is the distribution of covariates from a control population and ν is the distribution from a treated population. The goal is to scale the importance of different members from the control population based on how likely they are to be present in the treated population, in order to eliminate selection biases in the inference of treatment effects. Kallus (2018) proposed a generative-adversarial method for learning the scaling factor, but they do not consider transport.

Relation to Unbalanced OT. Lemma 3.2 points to the close relation between our problem formulation (6) and the optimal entropy-transport problem (2) by Liero et al. (2018). In practice, the formulations result in quite different numerical methods. The strategy proposed by Chizat et al. (2016) learns a joint measure whose marginals are approximately μ and ν (thus modeling mass variation implicitly), requires entropic regularization, and currently can only be used for discrete measures. On the other hand, our method learns a transport map and scaling factor for pushing μ to ν, does not require entropic regularization, and uses the generative-adversarial framework, which is scalable and can model high-dimensional measures with intricate geometry. For completeness, we present in Section A of the Appendix a *new stochastic method* for approximating the solution to the optimal-entropy problem based on the same entropy-regularized dual objective by Chizat et al. (2016), that can handle unbalanced transport between continuous measures. This method generalizes the approach of Seguy et al. (2017) for balanced OT and overcomes the scalability limitations of Chizat et al. (2016), but still does not directly model mass variation. In the following numerical experiments, we show the advantage of being able to model mass variation directly using Algorithm 1.



(a) (b) (c) (d)

Figure 3: Learning weights on MNIST data

4 Numerical Experiments

In this section, we illustrate in practice how Algorithm 1 performs unbalanced OT, with applications geared towards population modeling.

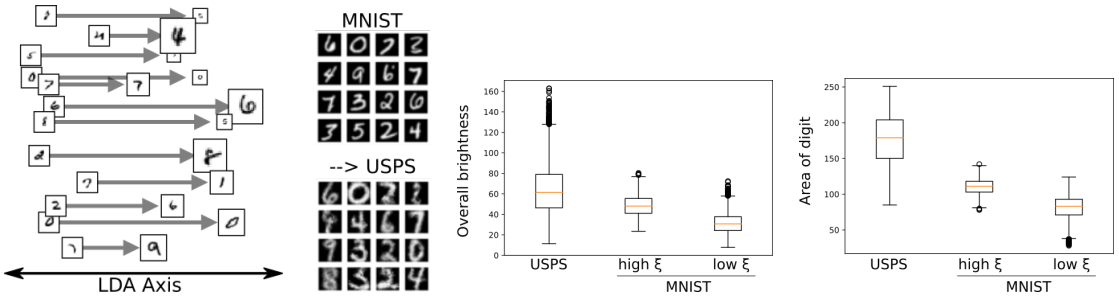
MNIST-to-MNIST. We first apply Algorithm 1 to perform unbalanced optimal transport between two modified MNIST datasets. The source dataset consists of regular MNIST digits with the class distribution shown in column 1 of Figure 3a. The target dataset consists of either regular (for the experiment in Figure 3b) or dimmed (for the experiment in Figure 3c) MNIST digits with the class distribution shown in column 2 of Figure 3a. The class imbalance between the source and target datasets imitates a scenerio in which certain classes (digits 0-3) become more popular and others (6-9) become less popular in the target population, while the change in brightness is meant to reflect population drift.

We evaluated Algorithm 1 on the problem of transporting the source distribution to the target distribution, enforcing a high cost of transport (w.r.t. Euclidean distance). In both cases, we found that the scaling factor over each of the digit classes roughly reflects its ratio of imbalance between the source and target distributions (Figure 3b-c). These experiments validate that the scaling factor learned by Algorithm 1 reflects the class imbalances and can be used to model the growth and decline of different classes in a population. Figure 3d is a schematic illustrating the reweighting that occurs during unbalanced transport.

MNIST-to-USPS. Next, we apply unbalanced OT from the MNIST dataset to the USPS datasets. As before, these two datasets are meant to imitate a population sampled at two different time points, this time with a large degree of evolution. We use Algorithm 1 to model the evolution of the MNIST distribution to the USPS distribution, taking as transport cost the Euclidean distance between the original and transported images.

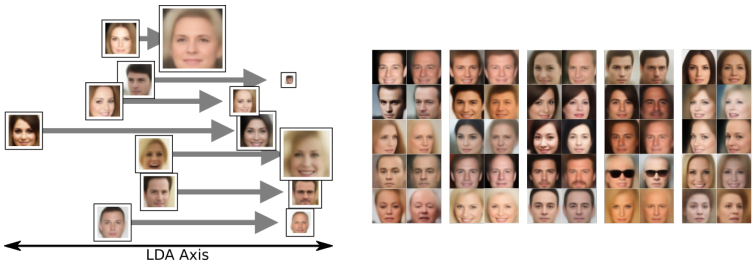
A summary of the transport is visualized in Figure 4a. Each arrow originates from a real MNIST image and points towards the predicted appearance of this image in the USPS dataset. The size of the image reflects the scaling factor of the original MNIST image, i.e. whether it is increasing or decreasing in prominence in the USPS dataset compared to the MNIST dataset according to the unbalanced transport model. Even though the Euclidean distance is not an ideal measure of correspondence between MNIST and USPS digits, many MNIST digits were able to preserve their likeness during the transport (Figure 4b). We analyzed which MNIST digits were considered as increasing or decreasing in prominence by the model. The MNIST digits with higher scaling factors were generally brighter (Figure 4c) and covered a larger area of pixels (Figure 4d) compared to the MNIST digits with lower scaling factors. These results are consistent with the observation that the target USPS digits are generally brighter and contain more pixels.

CelebA-Young-to-CelebA-Aged. We applied Algorithm 1 on the CelebA dataset to performed unbalanced transport from the population of young faces to the population of aged faces. This synthetic problem most



(a) (b) (c) (d)

Figure 4: Unbalanced Optimal Transport from MNIST to USPS digits



(a) (b)

	Female	Male
Low ξ	92386	44896
High ξ	9913	8029

	Female	Male
Young	103287	53447
Aged	14878	30987

(c)

Figure 5: Unbalanced Optimal Transport from Young to Aged CelebA Faces

closely imitates a real application of interest, which is modeling the transformation of a population based on samples taken from two timepoints. Since the Euclidean distance between two faces is a poor measure of semantic similarity, we first train a variational autoencoder (VAE) (Kingma & Welling, 2013) on the CelebA dataset and encode all samples into the latent space. We then apply Algorithm 1 to perform unbalanced optimal transport from the encoded young to the encoded aged faces, taking the transport cost to be the Euclidean distance in the latent space. The use of VAE latent encodings as a measure of semantic similarity between the decoded images was also considered by Engel et al. (2017).

A summary of the transport is visualized in Figure 5a. Each arrow originates from a real face from the young population and points towards the predicted appearance of this face in the aged population. Generally, the transported faces retain the most salient features of the original faces (Figure 5b), although there are exceptions (e.g. gender swaps) which reflects that some features are not prominent components of the VAE encodings. Interestingly, the young faces with higher scaling factors were significantly enriched for males compared to young faces with lower scaling factors (Figure 5c, top, $p = 0$). In other words, our model predicts growth in the prominence of male faces compared to female faces as the CelebA population evolves from young to aged. After observing this phenomenon, we confirmed based on checking the ground truth labels that there was indeed a strong gender imbalance between the young and aged populations: while the young population is predominantly female, the aged population is predominantly male (Figure 5c, bottom).

REFERENCES

Matthew Amodio and Smita Krishnaswamy. Magan: Aligning biological manifolds. *arXiv preprint arXiv:1803.00385*, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. *arXiv preprint arXiv:1710.06276*, 2017.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 7, 2017.

Lénaïc Chizat and Simone Di Marino. A tumor growth model of Hele-Shaw type as a gradient flow. *arXiv preprint arXiv:1712.06124*, 2017.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1): 1–44, 2018.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–289. Springer, 2014.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. *arXiv preprint arXiv:1711.05772*, 2017.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pp. 820–828, 2016.

Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*, 2018.

Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Stanislav Kondratyev, Léonard Monsaingeon, Dmitry Vorotnikov, et al. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164, 2016.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, pp. 1089–1096, 2008.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *BioRxiv*, pp. 191056, 2017.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pp. 2868–2876, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

APPENDIX

A DUAL STOCHASTIC METHOD

In this section, we present a stochastic method for unbalanced OT based on the regularized dual formulation of (Chizat et al., 2015), which can be considered a natural generalization of Seguy et al. (2017). The dual formulation of (2) is given by

$$\sup_{u,v} -\int \psi_1^*(-u)d\mu - \int \psi_2^*(-v)d\nu$$

subject to $u \oplus v \leq c$, where the supremum is taken over functions $u : \mathcal{X} \to [-\psi'_{1\infty}, \infty]$ and $v : \mathcal{Y} \to [-\psi'_{2\infty}, \infty]$. This is a constrained optimization problem that is challenging to solve. A standard technique for making the dual problem unconstrained is to add a strongly convex regularization term to the primal objective (Blondel et al., 2017), such as an entropic regularization term (Cuturi, 2013):

$$R_e(\gamma) = \epsilon D_{\psi_{KL}}(\gamma|\mu \otimes \nu)$$

where $\epsilon > 0$. Concretely, this term has a "smoothing" effect on the transport plan, in the sense that it encourages plans with high entropy. By the Fenchel-Rockafellar theorem, the dual of the regularized problem is,

$$W_{ub}^{\epsilon}(\mu, \nu) := \sup_{u,v} -\int \psi_1^*(-u)d\mu - \int \psi_2^*(-v)d\nu - \epsilon \int e^{(u+v-c)/\epsilon} d(\mu \otimes \nu), \tag{11}$$

where the supremum is taken over functions $u : \mathcal{X} \to [-\psi'_{1\infty}, \infty]$ and $v : \mathcal{Y} \to [-\psi'_{2\infty}, \infty]$, and the relationship between the primal optimizer γ^* and dual optimizer (u^*, v^*) is given by:

$$d\gamma^* = e^{(u^*+v^*-c)/\epsilon} d(\mu \otimes \nu). \tag{12}$$

Next, we rewrite (11) in terms of expectations. We assume that one has access to samples from μ, ν, and in the setting where μ, ν are not normalized, also samples to the normalized measures $\tilde{\mu}, \tilde{\nu}$ as well as the normalization constants. Based on these assumptions, we have

$$W_{ub}^{\epsilon}(\mu, \nu) = \sup_{u,v} -c_\mu \mathbb{E}_{x\sim\tilde{\mu}} \psi_1^*(-u(x)) - c_\nu \mathbb{E}_{y\sim\tilde{\nu}} \psi_2^*(-v(y)) - \epsilon c_\mu c_\nu \mathbb{E}_{x\sim\tilde{\mu}, y\sim\tilde{\nu}} e^{(u(x)+v(y)-c(x,y))/\epsilon}. \tag{13}$$

If ψ_1^*, ψ_2^* are differentiable, we can parameterize u, v with neural networks u_θ, v_ϕ and optimize θ, ϕ using stochastic gradient descent. This is described in Algorithm 2. Note that this algorithm is a generalization of the algorithm of (Seguy et al., 2017) from classical OT to unbalanced OT. Indeed, taking ψ_1, ψ_2 to be equality constraints, (11) becomes

$$\sup_{u,v} \int u \, d\mu + \int v \, d\nu - \epsilon \int e^{(u+v-c)/\epsilon} d(\mu \otimes \nu),$$

which is the dual of the entropy-regularized classical OT problem.

Algorithm 2 SGD for Unbalanced OT

Input: Initial parameters θ, ϕ; step size η; regularization parameter ϵ; constants c_μ, c_ν and normalized measures $\tilde{\mu}$, $\tilde{\nu}$
Output: Updated parameters θ, ϕ
while (θ, ϕ) not converged **do**
Sample $(x_1, y_1), \cdots, (x_n, y_n)$ from $\tilde{\mu} \otimes \tilde{\nu}$

$$\ell(\theta, \phi) := \frac{1}{n} \sum_{i=1}^{n} [c_\mu \psi_1^*(-u(x_i)) + c_\nu \psi_2^*(-v(y_i)) + \epsilon c_\mu c_\nu e^{(u(x_i)+v(y_i)-c(x_i,y_i))/\epsilon}]$$

Update θ, ϕ by gradient descent on $\ell(\theta, \phi)$
end while

The dual solution (u^*, v^*) learned from Algorithm 2 can be used to reconstruct the primal solution γ^* based on the relation in (12). Concretely, γ^* is a transport map that indicates the amount of mass transported between every pair of points in $\mathcal{X}$ and $\mathcal{Y}$. Note that the marginals of γ^* with respect to $\mathcal{X}$ and $\mathcal{Y}$ are not necessarily μ and ν, which is where mass variation is implicitly built into the problem. Given γ^*, it is possible to also learn an "averaged" deterministic mapping from $\mathcal{X}$ to $\mathcal{Y}$. A standard approach is to take the *barycentric projection* $T : \mathcal{X} \to \mathcal{Y}$, defined as,

$$T(x) = \min_{z \in \mathcal{Y}} \mathbb{E}_{y \sim \gamma^*(\cdot|x)} d(z, y),$$

with respect to some distance $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Seguy et al. (2017) proposed a stochastic algorithm for learning such a map from the dual solution, which we reproduce in Algorithm 3.

Algorithm 3 Learning Barycentric Projection

Input: Learned functions u, v; initial T_θ; distance function d
Output: Updated T_θ
while T_θ not converged **do**
Sample $(x_1, y_1), \cdots, (x_n, y_n)$ from $\tilde{\mu} \otimes \tilde{\nu}$

$$\ell(\theta) := \frac{1}{n} \sum_{i=1}^{n} d(x_i, T_\theta(y_i)) e^{(u(x_i)+v(y_i)-c(x_i,y_i))/\epsilon}$$

Update T_θ by gradient descent on $\ell(\theta)$
end while

B PROOFS FROM SECTION 3

Proof of Lemma 3.1.

$$\begin{aligned} D_\psi(P|Q) &= \psi'_\infty P_\perp(T) + \int_T \psi\left(\frac{dP}{dQ}\right) dQ \\ &= \psi'_\infty P_\perp(T) + \int_T \sup_{\xi \in \mathbb{R}} \{\xi \frac{dP}{dQ} - \psi^*(\xi)\} dQ \\ &\quad \text{(by defintion of convex conjugate)} \end{aligned}$$

$$
\begin{aligned}
&= \psi'_\infty P_\perp(T) + \int_T \sup_{\xi \in (-\infty, \psi'_\infty]} \{\xi \frac{dP}{dQ} - \psi^*(\xi)\} dQ \\
&\text{(by Lemma B.1)} \\
&= \int_T \sup_{\xi \in (-\infty, \psi'_\infty]} \{\xi dP_\perp + \xi \frac{dP}{dQ} dQ - \psi^*(\xi) dQ\} \\
&= \sup_{f \in \mathcal{F}} \int_T f dP - \psi^*(f) dQ
\end{aligned}
$$

The optimal f over the support of Q is obtained when $\frac{dP}{dQ}$ belongs to the subdifferential of $\psi^*(f)$, or equivalently by duality, when f belongs to the subdifferential of $\psi(\frac{dP}{dQ})$. It is straightforward to see that the optimal f over the support of $P_\perp$ is equal to ψ'_∞.

□

Lemma B.1. *If* $\xi > \psi'_\infty := \lim_{s \to \infty} \frac{\psi(s)}{s}$*, then* $\psi^*(\xi) = \infty$.

Proof.

$$
\begin{aligned}
\psi^*(\xi) &= \sup_{s \in \mathbb{R}} s\xi - \psi(s) \\
&\geq \lim_{s \to \infty} s(\xi - \frac{\psi(s)}{s}) \\
&= \infty \text{ if } \xi > \psi'_\infty
\end{aligned}
$$

□

Proof of Lemma 3.2. First we show $L_\psi(\mu, \nu) \geq \tilde{W}_{c_1, c_2, \psi}(\mu, \nu)$. If $L_\psi(\mu, \nu) = \infty$, this is trivial, so assume $L_\psi(\mu, \nu) < \infty$. Let (T, ξ) be any solution and define $T_x : z \mapsto T(x, z)$. Let $\gamma_{y|x} \in \mathscr{M}_+^1(\mathcal{Y})$ denote the pushforward measure of λ under T_x for all $x \in \mathcal{X}$, and define $\gamma_\mathcal{X} \in \mathscr{M}_+(\mathcal{X})$ as

$$
\gamma_\mathcal{X}(A) := \int_\mathcal{X} \mathbb{1}_A(x) \xi(x) d\mu(x)
$$

for all $A \in \mathcal{B}(\mathcal{X})$, i.e. so ξ is the Radon-Nikodym derivative of $\gamma_\mathcal{X}$ with respect to μ. Let γ be the product measure of $\gamma_{y|x}$ and $\gamma_\mathcal{X}$, defined as

$$
\gamma(C) := \int \int \mathbb{1}_C(y, x) d\gamma_{y|x}(y) d\gamma_\mathcal{X}(x)
$$

for all $C \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y})$. Finally, by these definitions, note that $\pi^\mathcal{X}_\# \gamma = \gamma_\mathcal{X}$ and

$$
\pi^\mathcal{Y}_\# \gamma(B) = \int \gamma_{y|x}(B) \gamma_\mathcal{X}(dx) = \int \lambda(T_x^{-1}(B)) d\gamma_\mathcal{X}(x) = \nu'_{T,\xi}(B)
$$

for all $B \in \mathcal{B}(\mathcal{Y})$. It follows that

$$\begin{aligned}
&\int_X (\int_Z c_1(x, T_x(z))d\lambda(z))\xi(x)d\mu(x) + \int c_2(\xi(x))d\mu(x) + D_\psi(\nu'_{T,\xi}|\nu) \\
&= \int_X (\int_Y c_1(x, y)d\gamma_{y|x}(y))d\gamma_{\mathcal{X}} + \int c_2(\xi(x))d\mu(x) + D_\psi(\nu'_{T,\xi}|\nu) \\
&= \int_X (\int_Y c_1(x, y)d\gamma_{y|x}(y))d\gamma_{\mathcal{X}} + \int c_2(\frac{d\gamma_{\mathcal{X}}}{d\mu})d\mu(x) + D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&= \int c_1 d\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&\geq \tilde{W}_{c,\psi_1,\psi_2}(\mu, \nu)
\end{aligned}$$

Since this inequality holds for any (T, ξ), taking the infimum over the left side yields $L_\psi(\mu, \nu) \geq \tilde{W}_{c,\psi_1,\psi_2}(\mu, \nu)$. To show $L_\psi(\mu, \nu) \leq \tilde{W}_{c,\psi_1,\psi_2}(\mu, \nu)$, assume $\tilde{W}_{c,\psi_1,\psi_2}(\mu, \nu) < \infty$ and let γ be a solution. Let $\gamma_{\mathcal{X}} = \pi^{\mathcal{X}}_{\#}\gamma, \gamma_{\mathcal{Y}} = \pi^{\mathcal{Y}}_{\#}\gamma$ denote its marginals. Then by the disintegration theorem, there exist probability measures $\{\gamma_{y|x}\}_{x\in\mathcal{X}}$ in $\mathscr{M}^1_+(\mathcal{Y})$ such that

$$\gamma(C) := \int\int \mathbb{1}_C(y, x)d\gamma_{y|x}(y)d\gamma_{\mathcal{X}}(x)$$

for all $C \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y})$. Since $\mathcal{X}, \mathcal{Y}$ are in $\mathbb{R}^d$, this family of measures is $\gamma_{\mathcal{X}}$-a.e. unique. Since $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ is an atomless probability space, there exists measurable functions $\{T_x : \mathcal{Z} \to \mathcal{Y}\}_{x\in\mathcal{X}}$ such that $\gamma_{y|x}$ is the pushforward measure of λ under T_x. Note that

$$\gamma_{\mathcal{Y}}(B) = \int \gamma_{y|x}(B)\gamma_{\mathcal{X}}(dx) = \int \lambda(T_x^{-1}(B))d\gamma_{\mathcal{X}}(x) = \nu'_{T,\xi}(B)$$

for all $B \in \mathcal{B}(\mathcal{Y})$. It follows that

$$\begin{aligned}
&\int c_1 d\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&= \int_X \left(\int_Y c_1(x, y)d\gamma_{y|x}(y)\right) d\gamma_{\mathcal{X}} + D_{c_2}(\gamma_{\mathcal{X}}|\mu) + D_\psi(\gamma_{\mathcal{Y}}|\nu) \\
&= \int_X \left(\int_Z c(x, T_x(z))d\lambda(z)\right) \xi(x)d\mu(x) + D_{c_2}(\gamma_{\mathcal{X}}|\mu) + D_\psi(\gamma_{\mathcal{Y}}|\nu) \\
&= \int_X \left(\int_Z c(x, T_x(z))d\lambda(z)\right) \xi(x)d\mu(x) + \int c_2(\frac{d\gamma_{\mathcal{X}}}{d\mu})d\mu(x) + D_\psi(\nu'_{T,\xi}|\nu) \\
&= \int_X \left(\int_Z c(x, T_x(z))d\lambda(z)\right) \xi(x)d\mu(x) + \int c_2(\xi(x))d\mu(x) + D_\psi(\nu'_{T,\xi}|\nu) \\
&\geq L_\psi(\mu, \nu),
\end{aligned}$$

where ξ is taken to be the Radon-Nikodym derivative $\frac{d\gamma_{\mathcal{X}}}{d\mu}$. Since the inequality holds for any γ, this implies $\tilde{W}_{c_1,c_2,\psi}(\mu, \nu) \geq L_\psi(\mu, \nu)$.

□

Proof of Proposition 3.3. Note that $\tilde{W}_{c_1,c_2,\psi}(\mu,\nu)$ is equivalent to $W_{c_1,c_2,\psi}(\mu,\nu)$ when $\mathcal{X}$ is restricted to the support of μ. If c_1, c_2, ψ satisfy (i) or (ii), they also satisfy (i) and (ii) when $\mathcal{X}$ is restricted to the support of μ. By Theorem 3.3 of (Liero et al., 2018), $\tilde{W}_{c_1,c_2,\psi}(\mu,\nu)$ has a minimizer. It follows from the construction of the proof of Lemma 3.2 that a minimizer of $L_\psi(\mu,\nu)$ also exists. For uniqueness, if $\psi'_\infty = \infty$, then it follows from Lemma 3.5 (Liero et al., 2018) and the fact that minimizers are restricted to $\mathcal{G}$ that the marginals $\pi^{\mathcal{X}}_{\#}\gamma$, $\pi^{\mathcal{Y}}_{\#}\gamma$ are uniquely determined for any solution γ of $\tilde{W}_{c_1,c_2,\psi}(\mu,\nu)$. The uniqueness of γ then follows from the proof of Corollary 3.6 (Liero et al., 2018). It follows from the construction of the proof of Lemma 3.2 that the product measure generated by minimizers of $L_\psi(\mu,\nu)$ is unique. □

Proof of Theorem 3.4. Since $\zeta^k\psi(s)$ converges pointwise to the equality constraint $\iota_=(s)$, which is 0 for $s = 1$ and ∞ otherwise, by Lemma 3.9 (Liero et al., 2018), we have that $\liminf_{k\to\infty} \tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) \geq \tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$. Additionally, $\tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) \leq \tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$ for any value of k since for any minimizer γ of $\tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$, it holds that $\pi^{\mathcal{Y}}_{\#}\gamma = \nu$, so

$$\tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu) = \int c_1 d\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\zeta^k\psi_2}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \geq \tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu),$$

for all k. Therefore, $\lim_{k\to\infty} \tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) = \tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$, and by Lemma 3.2, the first part of the proposition holds.

For the second part, by hypothesis, we have that $\tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu) = C < \infty$. It follows that $\tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) \leq C$ for all k. By Proposition 2.10 (Liero et al., 2018), the sequence of minimizers γ^k is bounded. If the assumptions of Proposition 3.3 are satisfied, then the sequence γ^k is equally tight, under assumptions (ii) by Proposition 2.10 (Liero et al., 2018) again, and under assumptions (i), by the Markov inequality: for any $\lambda > 0$,

$$\gamma^k(\{(x,y) \in \mathcal{X}\times\mathcal{Y} | c_1(x,y) > \lambda\}) \leq \frac{1}{\lambda}\int c_1 d\gamma^k \leq \frac{C}{\lambda}.$$

Since γ^k are bounded and equally tight, by an extension of Prokhorov's theorem (Theorem 2.2 of (Liero et al., 2018)), there exists a subsequence of γ^k that is weakly convergent to some $\bar{\gamma}$. By lower semicontinuity, we have

$$\int c_1 d\bar{\gamma} + D_{c_2}(\pi^{\mathcal{X}}_{\#}\bar{\gamma}|\mu) + \limsup_{k\to\infty} D_{\zeta^k\psi_2}(\pi^{\mathcal{Y}}_{\#}\gamma^k|\nu) \leq \tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu) = C$$

Since $D_{\zeta^k\psi}(\pi^{\mathcal{Y}}_{\#}\gamma^k|\nu) = \zeta^k D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma^k|\nu) \geq 0$ and $\zeta^k \to \infty$, for the left side to be finite, $D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma^k|\nu)$ must converge to 0, so $D_\psi(\pi^{\mathcal{Y}}_{\#}\bar{\gamma}|\nu) = 0$ by lower semicontinuity. Therefore, $\bar{\gamma}$ is a minimizer of $\tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$. By construction of the proof of Lemma 3.2, γ^k is equivalent to the product measure induced by minimizers of $L_{\zeta^k\psi}(\mu,\nu)$, so the second part of the proposition holds. □

C Practical Considerations for Numerical Experiments

Choice of cost functions. Proposition 3.3 gives sufficient conditions on c_1, c_2 for the problem to be well-posed. In practice, it is often convenient the cost of transport, c_1, to be some measurement of correspondence between $\mathcal{X}$ and $\mathcal{Y}$. For example, we can take $c_1(x,y)$ to be the Euclidean distance between x and y after mapping them to some common feature space. For the cost of mass adjustment, c_2, it is generally sensible to choose some convex function that vanishes at 1 (i.e. no mass adjustment) and such that $\lim_{x\to 0} c_2(x) = \lim_{x\to\infty} c_2(x) =$

Name	$\psi(s)$	$D_\psi(P\|Q)$	$\psi^*(s)$	ψ'_∞
Kullback-Leibler (KL)	$s\log s - s + 1$	$\int \log \frac{dP}{dQ} dP - \int dP + \int dQ$	e^s	∞
Pearson χ^2	$(s-1)^2$	$\int (\frac{dP}{dQ} - 1)^2 dQ$	$\frac{s^2}{4} + s$	∞
Hellinger	$(\sqrt{s} - 1)^2$	$\int (\sqrt{\frac{dP}{dQ}} - 1)^2 dQ$	$\frac{s}{1-s}$	1
Jensen-Shannon	$s\log s - (s+1)\log\frac{s+1}{2}$	$\frac{1}{2}D_{KL}(P\|\frac{P+Q}{2}) + \frac{1}{2}D_{KL}(Q\|\frac{P+Q}{2})$	$-\log(2-e^s)$	$\log 2$

Table 1: Table of some common ψ-divergences, associated entropy functions ψ, and convex conjugates ψ^* for Algorithm 1, partly adapted from (Nowozin et al., 2016).

∞ to prevent ξ from becoming too small or too large. Any of the entropy functions shown in Table 1 are reasonable choices. In practice, we also found that employing a Lipschitz penalty on ξ stabilizes training.

Choice of ψ-divergence and f. Based on Theorem 3.4, ψ must be uniquely minimized at $\psi(1) = 0$. In Nowozin et al. (2016), it was shown that any ψ-divergence could be used to train generative models, i.e. to match a generated distribution P to a true data distribution Q. This is due to Jensen's inequality: for any convex lower semi-continous entropy function ψ, $D_\psi(P|Q)$ is uniquely minimized when $P = Q$, where P, Q are probability measures. However, this additional constraint is required in our case since P, Q are not probability measures, as illustrated by the following example.

Example C.1. In the original GAN paper, the discrminative objective,

$$\sup_f \int \log f(x) dP(x) - \int \log(1 - f(x)) dQ(x),$$

corresponds to $D_\psi(P|Q)$ with $\psi(s) = s\log s - (s+1)\log(s+1)$. If P, Q are probability measures, this divergence is equivalent to the Jensen-Shannon divergence and is minimized when $P = Q$. If P, Q are unnormalized positive measures, the divergence is minimized when $P = \infty$ and $Q = 0$.

Table 1 provides some common ψ-divergences that can be used for unbalanced transport.

According to Lemma 3.1, f should belong to a class of functions that maps from $\mathcal{Y}$ to $(-\infty, \psi'_\infty]$. In practice, this can be enforced by parameterizing f using a neural network with a final layer that maps to the correct range. In practice, we also found that employing a Lipschitz penalty on f stabilizes training.

Improved training dynamics. Recall that the objective function for our alternating gradient updates was

$$\ell(\theta, \phi, \omega) := \frac{1}{n}\sum_{i=1}^{n}[c_\mu c_1(x_i, T_\theta(x_i, z_i))\xi_\phi(x_i) + c_\mu c_2(\xi_\phi(x_i)) \\ + c_\mu \xi_\phi(x_i) f_\omega(T_\theta(x_i, z_i)) - c_\nu \psi^*(f_\omega(y_i))]$$

Early in training, ξ_ϕ can become very small for some x_i as none of the transported samples resemble samples from the target distribution. As a result, T_θ improves very slowly for some inputs x_i. One way to address this without changing the fixed point is to update ξ_ϕ, f_ω using the above objective and update T_θ using the following objective.

$$\ell(\theta, \phi, \omega) := \frac{1}{n}\sum_{i=1}^{n} c_\mu[c_1(x_i, T_\theta(x_i, z_i)) + f_\omega(T_\theta(x_i, z_i))]$$

Note that we have omitted terms in the original objective that do not include T_θ, and which therefore do not affect the gradient update. For the terms that remain, the difference is that we are rescaling the contribution of each sample x_i by $1/\xi(x_i)$. As long as $\xi(x_i) > 0$, this has the effect of rescaling the gradient update of the loss function with respect to each $T_\theta(x_i, z_i)$ *without* changing the direction of the update.