SCALABLE UNBALANCED OPTIMAL TRANSPORT USING GENERATIVE ADVERSARIAL NETWORKS

Karren D. Yang & Caroline Uhler
Laboratory for Information & Decision Systems
Institute for Data, Systems and Society
Massachusetts Institute of Technology
Cambridge, MA, USA
{karren, cuhler}@mit.edu

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. We provide theoretical justification for this formulation, showing that it is closely related to an existing static formulation by Liero et al. (2018). We then propose an algorithm for solving this problem based on stochastic alternating gradient updates, similar in practice to GANs, and perform numerical experiments demonstrating how this methodology can 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).

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. Recently, Cuturi (2013) showed that regularizing the objective using an entropy term allows the dual problem to be solved more efficiently using the Sinkhorn algorithm. Stochastic methods based on the dual objective have been proposed for the continuous setting (Genevay et al., 2016; Seguy et al., 2017; Arjovsky et al., 2017). Optimal transport has been applied 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 &

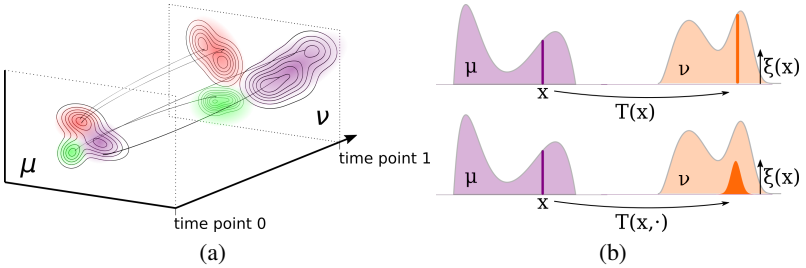


Figure 1: (a) Illustration of the problem of modeling the transformation of a source population μ to a target population ν. In this example, one sub-population is growing more rapidly than the others. (b) Schematic of Monge-like formulations of unbalanced optimal transport. The objective is to learn a transport map T (for transporting mass) and scaling factor ξ (for mass variation) to push the source μ to the target ν, using a deterministic transport map (top) (Chizat et al., 2015) or a stochastic transport map (bottom).

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.

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, while optimal entropy-transport allows mass variation, it cannot explicitly model it, and there are currently no methods that can perform unbalanced OT between *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 propose to solve a Monge-like formulation of unbalanced OT, in which the goal is to learn a stochastic transport map and scaling factor to push a source to a target measure in a cost-optimal manner. This generalizes the unbalanced Monge OT problem by Chizat et al. (2015).
- By relaxing this problem, we obtain an alternative form of the optimal entropy-transport problem by Liero et al. (2018), which confers desirable theoretical properties.
- We develop scalable methodology for solving the relaxed problem. Our derivation uses a convex conjugate representation of divergences, resulting in an alternating gradient descent method similar 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, the CelebA dataset, and a recent single-cell RNA-seq dataset from zebrafish embrogenesis.

In addition to these main contributions, for completeness we also propose a new scalable method (Algorithm 2) in the Appendix for solving the optimal-entropy transport problem by Liero et al. (2018) in the continuous

setting. The algorithm extends the work of Seguy et al. (2017) to unbalanced OT and is a scalable alternative to the algorithm of Chizat et al. (2016) for very large or continuous datasets.

2 PRELIMINARIES

Notation. Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ 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 measurable function T, let $T_\#$ denote its pushforward operator: if μ is a measure, then $T_\#\mu$ is the pushforward measure of μ under T. Finally, let $\pi^{\mathcal{X}}$, $\pi^{\mathcal{Y}}$ be functions that project onto $\mathcal{X}$ and $\mathcal{Y}$; for a joint measure $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$, $\pi^{\mathcal{X}}_\#\gamma$ and $\pi^{\mathcal{Y}}_\#\gamma$ are 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. Monge (1781) formulated this problem as a search over deterministic transport maps. Specifically, given $\mu \in \mathscr{M}_+^1(\mathcal{X}), \nu \in \mathscr{M}_+^1(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, Monge OT seeks a measurable function $T : \mathcal{X} \to \mathcal{Y}$ minimizing

$$\inf_T \int_{\mathcal{X}} c(x, T(x))\, d\mu(x) \tag{1}$$

subject to the constraint $T_\#\mu = \nu$. While the optimal T has an intuitive interpretation as an optimal transport map, the Monge problem is non-convex and not always feasible depending on the choices of μ and ν. The *Kantorovich OT* problem is a convex relaxation of the Monge problem that formulates OT as a search over probabilistic transport plans. 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

$$W(\mu, \nu) := \inf_\gamma \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y). \tag{2}$$

Note that the conditional probability distributions $\gamma_{y|x}$ specify stochastic maps from $\mathcal{X}$ to $\mathcal{Y}$ and can be considered a "one-to-many" version of the deterministic map from the Monge problem. In terms of numerical methods, the relaxed problem is a linear program that is always feasible and can be solved in $O(n^3)$ time for discrete μ, ν. Cuturi (2013) recently showed that introducing entropic regularization results in a simpler dual optimization problem that can be solved efficiently using the Sinkhorn algorithm. Based on the entropy-regularized dual problem, Genevay et al. (2016) and Seguy et al. (2017) proposed stochastic algorithms for computing transport plans that can handle continuous measures.

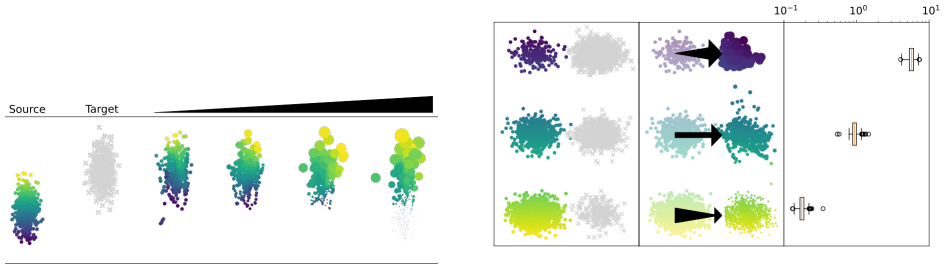
Unbalanced OT. Several formulations that extend classical OT to handle mass variation have been proposed (Chizat et al., 2015; 2018; Kondratyev et al., 2016). Existing numerical methods are based on a Kantorovich-like formulation known as *optimal-entropy transport* (Liero et al., 2018). This formulation is obtained by relaxing the marginal constraints of (2) 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

$$W_{ub}(\mu, \nu) := \inf_\gamma \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\gamma(x, y) + D_{\psi_1}(\pi^{\mathcal{X}}_\#\gamma|\mu) + D_{\psi_2}(\pi^{\mathcal{Y}}_\#\gamma|\nu), \tag{3}$$

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

$$D_\psi(P|Q) := \psi'_\infty P^\perp(\mathcal{T}) + \int_{\mathcal{T}} \psi\left(\frac{dP}{dQ}\right) dQ, \tag{4}$$

where $\psi'_\infty := \lim_{s\to\infty} \frac{\psi(s)}{s}$ 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



(a) (b)

Figure 2: **Motivating examples for Unbalanced Monge OT.**

methods, the state-of-the-art in the discrete setting is a class of iterative scaling algorithms (Chizat et al., 2016) that generalize the Sinkhorn algorithm for computing regularized OT plans (Cuturi, 2013). There are no practical algorithms for unbalanced OT between continuous measures, especially in high-dimensional spaces.

3 Scalable Unbalanced OT using GANs

In this section, we propose the first algorithm for unbalanced OT that directly models mass variation and can be applied towards transport between high-dimensional continuous measures. The starting point of our development is the following Monge-like formulation of unbalanced OT, in which the goal is to learn a stochastic transport map and scaling factor to push a source to a target measure in a cost-optimal manner.

Unbalanced Monge OT. Let $c_1 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be the cost of transport and $c_2 : \mathbb{R}^+ \to \mathbb{R}^+$ the cost of mass variation. Let the probability space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ be the source of randomness in the transport map T. Given two positive measures $\mu \in \mathscr{M}_+(\mathcal{X})$ and $\nu \in \mathscr{M}_+(\mathcal{Y})$, we seek a transport map $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and a scaling factor $\xi : \mathcal{X} \to \mathbb{R}^+$ minimizing

$$L(\mu, \nu) := \inf_{T,\xi} \int_{\mathcal{X}} \int_{\mathcal{Z}} c_1(x, T(x,z)) d\lambda(z) \xi(x) d\mu(x) + \int_{\mathcal{X}} c_2(\xi(x)) d\mu(x), \tag{5}$$

subject to the constraint $T_{\#}(\xi\mu \times \lambda) = \nu$. Concretely, the first and second terms of (5) penalize the cost of mass transport and variation respectively, and the equality constraint ensures that (T, ξ) pushes μ to ν exactly. A special case of (5) is the unbalanced Monge OT problem by Chizat et al. (2015), which employs a deterministic transport map (Figure 1b). We consider the more general case of stochastic (i.e. one-to-many) maps because it is a more suitable model for many practical problems. For example, in cell biology, it is natural to think of one cell in a source population as potentially giving rise to multiple cells in a target population. In practice, one can take $\mathcal{Z} = \mathbb{R}^n$ and λ to be the standard Gaussian measure if a stochastic map is desired; otherwise λ can be set to a deterministic distribution. The following are examples of problems that can be modeled using unbalanced Monge OT.

Example 3.1 (Figure 2a)**.** Suppose the objective is to model the transformation from a source measure (Column 1) to the target measure (Column 2), which represent a population of interest at two distinct time points. The transport map T models the transport/movement of points from the source to the target, while the scaling factor ξ models the growth (replication) or shrinkage (death) of these points. Different models of transformation are optimal depending on the relative costs of mass transport and variation (Columns 3-6).

Example 3.2 (Figure 2b)**.** Suppose the objective is to transport points from a source measure (1st panel, color) to a target measure (1st panel, grey) in the presence of class imbalances. A pure transport map would muddle together points from different classes, while an unbalanced transport map with a scaling factor is able to ameliorate the class imbalance (2nd panel). In this case, the scaling factor tells us explicitly how to downweigh or upweigh samples in the source distribution to balance the classes with the target distribution (3rd panel).

Relaxation. From an optimization standpoint, it is challenging to satisfy the constraint $T_{\#}(\xi\mu \times \lambda) = \nu$. We hence consider the following relaxation of (5) using a divergence penalty in place of the equality constraint:

$$L_\psi(\mu, \nu) := \inf_{T,\xi} \int_{\mathcal{X}} \int_{\mathcal{Z}} c_1(x, T(x, z)) d\lambda(z)\xi(x)d\mu(x) + \int_{\mathcal{X}} c_2(\xi(x))d\mu(x) + D_\psi(T_{\#}(\xi\mu \times \lambda)|\nu), \quad (6)$$

using an appropriate choice of ψ that satisfies the requirements of Lemma C.2 in the Appendix[1].This relaxation is the Monge-like version of the optimal-entropy transport problem (3) by Liero et al. (2018). Specifically, (T, ξ) specifies a joint measure $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$ given by

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

and by reformulating (6) in terms of γ instead of (T, ξ), one obtains the objective function for optimal-entropy transport. The main difference between the formulations is their search space, since not all joint measures $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$ can be specified by some choice of (T, ξ). For example, if T is a deterministic transport map, then γ is necessarily restricted to the set of deterministic couplings. Even if T is sufficiently random, it is generally not possible to specify all joint measures $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$: in the asymmetric Monge formulation (6), all the mass transported to $\mathcal{Y}$ must come from somewhere within the support of μ, since the scaling factor ξ allows mass to grow but not to materialize outside of its original support. Therefore equivalence can be established in general only when restricting the support of γ to $supp(\mu) \times \mathcal{Y}$ as described in the following lemma, whose proof is given in the Appendix.

Lemma 3.3. *Let $\mathcal{G}$ be the set of joint measures supported on $supp(\mu) \times \mathcal{Y}$, and define*

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

If $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ is atomless and c_2 is an entropy function, then $L_\psi(\mu, \nu) = \tilde{W}_{c_1,c_2,\psi}(\mu, \nu)$.

Based on the relation between (3) and (6), several theoretical results for (6) follow from the analysis of optimal entropy-transport by Liero et al. (2018). Importantly, one can show the following theorem, namely that for an appropriate and sufficiently large choice of divergence penalty, solutions of the relaxed problem (6) converge to solutions of the original problem (5). The proof is given in the Appendix.

Theorem 3.4. *Suppose c_1, c_2, ψ satisfy the existence assumptions of Proposition B.1 in the Appendix, and let $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ be an atomless probability space. 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 joint measure specified 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 joint measure specified by a minimizer of $L(\mu, \nu)$.*

Algorithm. Using the relaxation of unbalanced Monge OT in (6), we now show that the transport map and scaling factor can be learned by stochastic gradient methods. While the divergence term cannot easily be minimized using the definition in (4), we can write it as a penalty witnessed by an *adversary* function $f : \mathcal{Y} \to (-\infty, \psi'_\infty]$ using the convex conjugate representation (see Lemma B.2):

$$D_\psi(T_{\#}(\xi\mu \times \lambda)|\nu) = \sup_f \int_{\mathcal{X}} \int_{\mathcal{Z}} f(T(x, z))d\lambda(z)\xi(x)d\mu(x) - \int_{\mathcal{Y}} \psi^*(f(y))d\nu(y), \quad (7)$$

where ψ^* is the convex conjugate of ψ. The objective in (6) can now be optimized using alternating stochastic gradient updates after parameterizing T, ξ, and f with neural networks; see Algorithm 1 [2]. The optimization

[1]Note that $D_\psi(T_{\#}(\xi\mu \times \lambda)|\nu) = 0 \not\Rightarrow T_{\#}(\xi\mu \times \lambda) = \nu$ in general since the total mass of the transported measure is not constrained; for the relaxation to be valid, $\psi(s)$ should attain a unique minimum at $s = 1$ (see Lemma C.2).

[2]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.

procedure is similar to GAN training and can be interpreted as an adversarial game between (T, ξ) and f:

- T takes a point $x \sim \mu$ and transports it from $\mathcal{X}$ to $\mathcal{Y}$ by generating $T(x, z)$ where $z \sim \lambda$.
- ξ determines the importance weight of each transported point.
- Their shared objective is to minimize the divergence between transported samples and real samples from ν that is measured by the adversary f.
- Additionally, cost functions c_1 and c_2 encourage T, ξ to find the most cost-efficient strategy.

Algorithm 1 Generative-Adversarial Framework for Unbalanced Monge 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;

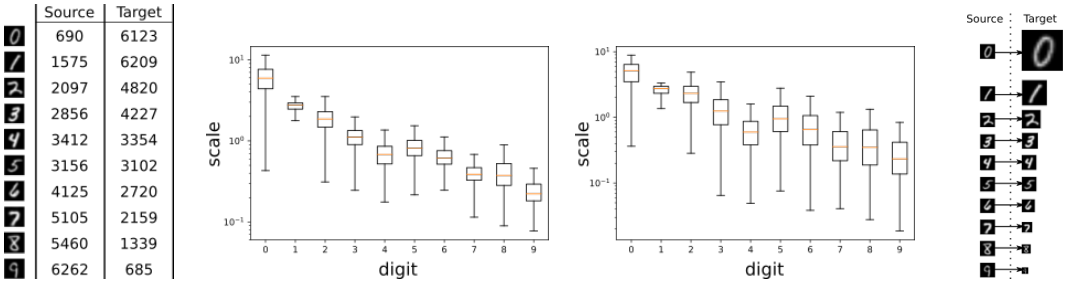
$$\ell(\theta, \phi, \omega) := \frac{1}{n} \sum_{i=1}^{n} [c_\mu c_1(x_i, T_\theta(x_i, z_i))\xi_\phi(x_i) + c_\mu c_2(\xi_\phi(x_i)) + c_\mu \xi_\phi(x_i) f_\omega(T_\theta(x_i, z_i)) - c_\nu \psi^*(f_\omega(y_i)).] \tag{8}$$

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

Table 1 in the Appendix provides some examples of divergences with corresponding entropy functions and convex conjugates that can be plugged into (7). Further practical considerations for implementation and training are discussed in Appendix C.

Relation to other approaches. The probabilistic Monge-like formulation (6) is similar to the Kantorovich-like entropy-transport problem (3) in theory, but they result in quite different numerical methods in practice. Algorithm 1 solves the non-convex formulation (6) and learns a transport map T and scaling factor ξ parameterized by neural networks, enabling scalable optimization using stochastic gradient descent. The networks are immediately useful for many practical applications; for instance, it only requires a single forward pass to compute the transport and scaling of a point from the source domain to the target. Furthermore, the neural architectures of T, ξ imbue their function classes with a particular structure, and when chosen appropriately, enable effective learning of these functions in high-dimensional settings. Due to the non-convexity of the optimization problem, however, Algorithm 1 is not guaranteed to find the global optimum. In contrast, the scaling algorithm of Chizat et al. (2016) based on (3) solves a convex optimization problem and is proven to converge, but is currently only practical for discrete problems and has limited scalability. For completeness, in Section A of the Appendix, we propose a *new stochastic method* based on the same dual objective as Chizat et al. (2016) that can handle transport between continuous measures (Algorithm 2 in the Appendix). This method generalizes the approach of Seguy et al. (2017) for handling transport between continuous measures and overcomes the scalability limitations of Chizat et al. (2016). However, the output is in the form of the dual solution, which is less interpretable for practical applications compared to the output of Algorithm 1. In particular, while one can compute a deterministic transport map known as a barycentric projection from the dual solution, it is unclear how best to obtain a scaling factor or a stochastic transport map that can generate samples outside of the target dataset. In the numerical experiments of Section 4, we show the advantage of directly learning a transport map and scaling factor using Algorithm 1.

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



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

Figure 3: Learning weights on MNIST data using unbalanced OT.

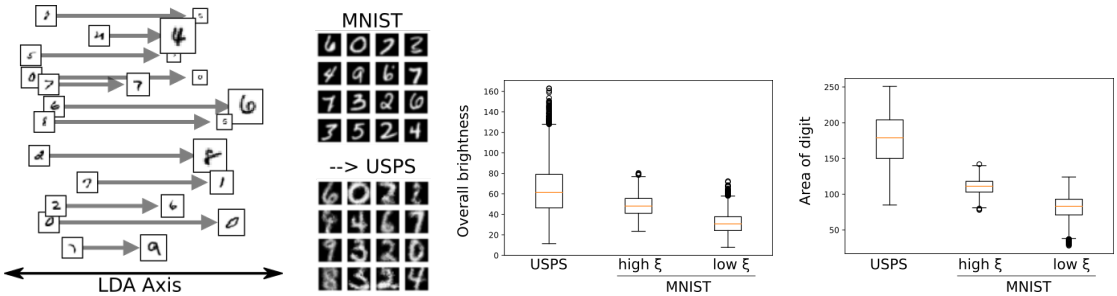
control population based on how likely they are to be present in the treated population, in order to eliminate selection biases in the inference of treatment effects. Kallus (2018) proposed a generative-adversarial method for learning the scaling factor, but they do not consider transport.

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 growth or decline of different classes in a population. Figure 3d is a schematic illustrating the reweighting that occurs during unbalanced OT.

MNIST-to-USPS. Next, we apply unbalanced OT from the MNIST dataset to the USPS dataset. 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 unbalanced 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 relatively increasing or decreasing in prominence in the USPS dataset compared to the MNIST dataset according to the unbalanced OT 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.

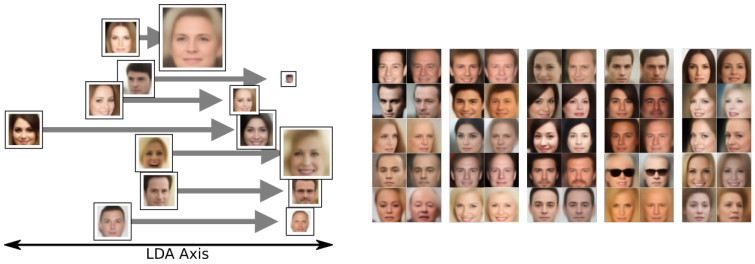


(a) (b) (c) (d)

Figure 4: Unbalanced Optimal Transport from MNIST to USPS digits.

CelebA-Young-to-CelebA-Aged. We applied Algorithm 1 on the CelebA dataset to perform unbalanced OT from the population of young faces to the population of aged faces. This synthetic problem 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 OT from the encoded young to the encoded aged faces, taking the transport cost to be the Euclidean distance in the latent space. A summary of the unbalanced 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; 9.6% (9,913/103,287) of young female faces had a high scaling factor as compared to 18.5% (8,029/53,447) for young male faces (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).

Zebrafish embroygenesis. A problem of great interest in biology is lineage tracing of cells between different developmental stages or during disease progression. This is a natural application of transport in which the source and target distributions are unbalanced: some cells in the earlier stage are more poised to develop into cells seen in the later stage. To showcase the relevance of learning the scaling factor, we apply Algorithm 1



(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.

to recent single-cell gene expression data from two stages of zebrafish embryogenesis (Farrell et al., 2018). The source population is from a late stage of blastulation and the target population from an early stage of gastrulation (Figure 6a). The results of the transport are plotted in Figure 6b-c after dimensionality reduction by PCA and T-SNE (Maaten & Hinton, 2008). To assess the scaling factor, we extracted the cells from the blastula stage with higher scaling factors (i.e. over 90th percentile) and compared them to the remainder of the cells using differential gene expression analysis, producing a ranked list of upregulated genes. Using the GOrilla tool (Eden et al., 2009), we found that the cells with higher scaling factors were significantly enriched for genes associated with differentiation and development of the mesoderm (Figure 6d). This experiment shows that analysis of the scaling factor can be applied towards interesting and meaningful biological discovery.

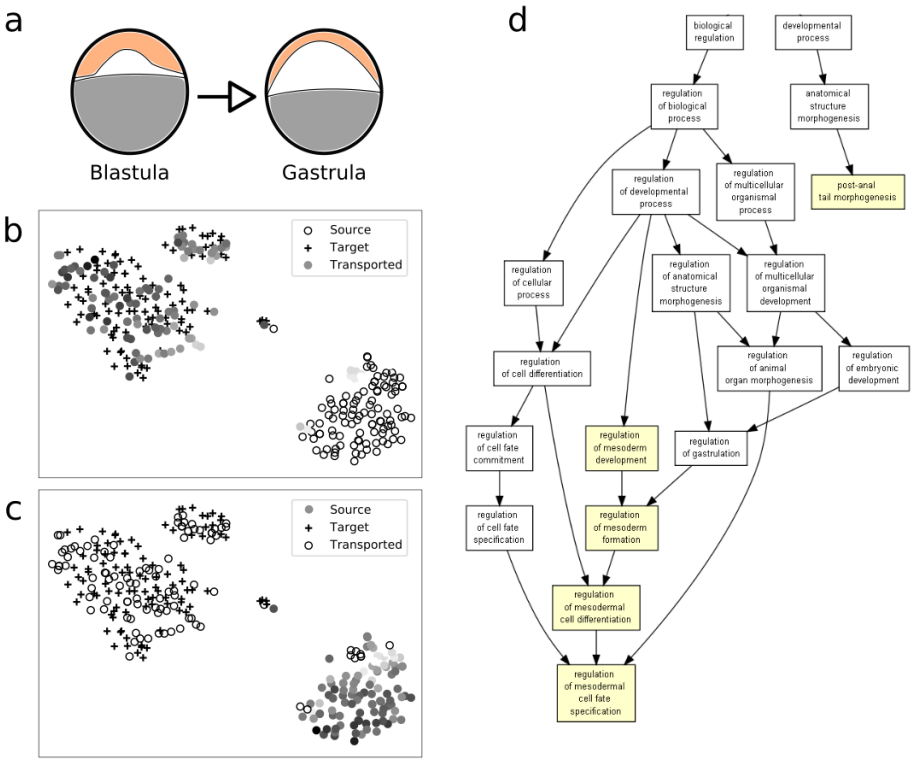


Figure 6: **Unbalanced OT on Zebrafish Single-Cell Gene Expression Data.** (a) Illustration of the blastula and gastrula stages of zebrafish embryogenesis. (b) T-SNE plot (Maaten & Hinton, 2008) of the unbalanced OT results for a subset of datapoints. The color of the transported points indicates the relative magnitude of the scaling factor (black = high, white = low). (c) Same plot as (b), where we have colored the source points instead of the transported points. (d) GOrilla output of significantly enriched processes (Eden et al., 2009) based on ranked list of enriched genes in cells with high scaling factors (black points from (c)) from a differential gene expression analysis. Processes in the graph are organized from more general (upstream) to more specific (downstream).

REFERENCES

Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv:1802.10151*, 2018.

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

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv:1701.04862*, 2017.

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: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:1712.06124*, 2017.

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

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

Richard M. Dudley. *Real Analysis and Probability: 0*. Chapman and Hall/CRC, 2018.

Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48, 2009.

Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360, 2018.

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.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANS. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

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: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: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.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, pp. 3478–3487, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *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: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:1703.10593*, 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 (3) 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 given by,

$$W_{ub}^{\epsilon}(\mu, \nu) := \sup_{u,v} -\int \psi_1^*(-u)d\mu - \int \psi_2^*(-v)d\nu - \epsilon \int e^{(u+v-c)/\epsilon} d(\mu \otimes \nu), \tag{9}$$

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

$$d\gamma^* = e^{(u^*+v^*-c)/\epsilon} d(\mu \otimes \nu). \tag{10}$$

Next, we rewrite (9) 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, then samples to the normalized measures $\tilde{\mu}, \tilde{\nu}$ as well as the normalization constants. Based on these assumptions, we have

$$W_{ub}^{\epsilon}(\mu, \nu) = \sup_{u,v} -c_\mu \mathbb{E}_{x\sim\tilde{\mu}} \psi_1^*(-u(x)) - c_\nu \mathbb{E}_{y\sim\tilde{\nu}} \psi_2^*(-v(y)) - \epsilon c_\mu c_\nu \mathbb{E}_{x\sim\tilde{\mu}, y\sim\tilde{\nu}} e^{(u(x)+v(y)-c(x,y))/\epsilon}. \tag{11}$$

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, (9) 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 (10). 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 Supplement to Section 3

B.1 Relaxation to optimal-entropy transport

The *objectives* in (6) and (3) are equivalent if one reformulates (6) in terms of γ instead of (T, ξ), where $\gamma \in \mathscr{M}_+(\mathcal{X} \times \mathcal{Y})$ is a joint measure given by

$$\gamma(C) := \int_{\mathcal{X}} \int_{\mathcal{Z}} \mathbb{1}_C(x, T(x, z)) d\lambda(z) \xi(x) d\mu(x), \quad \forall C \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y}). \tag{12}$$

Furthermore, the formulations are equivalent if one restricts the search space of (3) to contain only those joint measures that can be specified by some (T, ξ). This relation between the formulations is formalized by Lemma 3.3.

Proof of Lemma 3.3. 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 γ by (12). Note by this definition that

$$\pi^{\mathcal{X}}_{\#}\gamma(A) = \int_{\mathcal{X}} \mathbb{1}_A(x)\xi(x)d\mu(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

i.e. ξ is the Radon-Nikodym derivative of $\pi^{\mathcal{X}}_{\#}\gamma$ with respect to μ. Also

$$\pi^{\mathcal{Y}}_{\#}\gamma(B) = \int_{\mathcal{X}}\int_{\mathcal{Z}} \mathbb{1}_B(T(x,z))d\lambda(z)\xi(x)d\mu(x) = T_{\#}(\xi\mu \times \lambda)(B), \quad \forall B \in \mathcal{B}(\mathcal{Y}).$$

It follows that

$$\begin{aligned}
&\int_{\mathcal{X}} \left(\int_{\mathcal{Z}} c_1(x,T(x,z))d\lambda(z)\right)\xi(x)d\mu(x) + \int_{\mathcal{X}} c_2(\xi(x))d\mu(x) + D_\psi(T_{\#}(\xi\mu\times\lambda)|\nu) \\
&\quad = \int_{\mathcal{X}\times\mathcal{Y}} c_1(x,y)d\gamma(x,y) + \int_{\mathcal{X}} c_2(\xi(x))d\mu(x) + D_\psi(T_{\#}(\xi\mu\times\lambda)|\nu) \\
&\qquad \text{(by definition of } \gamma\text{, linearity and monotone convergence)} \\
&\quad = \int_{\mathcal{X}\times\mathcal{Y}} c_1(x,y)d\gamma(x,y) + \int_{\mathcal{X}} c_2(\frac{d\pi^{\mathcal{X}}_{\#}\gamma}{d\mu})d\mu(x) + D_\psi(T_{\#}(\xi\mu\times\lambda)|\nu) \\
&\qquad \text{(since } \xi \text{ is the Radon-Nikodym derivative)} \\
&\quad = \int_{\mathcal{X}\times\mathcal{Y}} c_1(x,y)d\gamma(x,y) + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_\psi(T_{\#}(\xi\mu\times\lambda)|\nu) \\
&\quad = \int_{\mathcal{X}\times\mathcal{Y}} c_1(x,y)d\gamma(x,y) + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_\psi(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&\quad \geq \tilde{W}_{c,\psi_1,\psi_2}(\mu,\nu).
\end{aligned}$$

Since this inequality holds for any (T,ξ), taking the infimum over the left-hand 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 any solution. By the disintegration theorem, there exists a family of probability measures $\{\gamma_{y|x}\}_{x\in\mathcal{X}}$ in $\mathscr{M}^1_+(\mathcal{Y})$ such that

$$\gamma(C) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_C(x,y)d\gamma_{y|x}(y)d\pi^{\mathcal{X}}_{\#}\gamma, \quad \forall C \in \mathcal{B}(\mathcal{X})\times\mathcal{B}(\mathcal{Y}).$$

Since $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ is atomless, it follows from Proposition 9.1.2 and Theorem 13.1.1 in (Dudley, 2018) that there exists a family of 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 for all $x \in \mathcal{X}$. Denoting $T(x,z) : (x,z) \mapsto T_x(z)$, then by a change of variables,

$$\gamma(C) = \int_{\mathcal{X}}\int_{\mathcal{Z}} \mathbb{1}_C(x,T(x,z))d\lambda(z)d\pi^{\mathcal{X}}_{\#}\gamma, \quad \forall C \in \mathcal{B}(\mathcal{X})\times\mathcal{B}(\mathcal{Y}).$$

By hypothesis, $\pi^{\mathcal{X}}_{\#}\gamma$ is restricted to the support of μ, i.e. $\pi^{\mathcal{X}}_{\#}\gamma \lll \mu$. Let ξ be the Radon-Nikodym derivative $\frac{d\pi^{\mathcal{X}}_{\#}\gamma}{d\mu}$. It follows from the Radon-Nikodym theorem that (T,ξ) satisfy

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

which is the same relation as in (12). Same as before,

$$\pi^{\mathcal{X}}_{\#}\gamma(A) = \int_{\mathcal{X}} \mathbb{1}_A(x)\xi(x)d\mu(x), \quad \forall A \in \mathcal{B}(\mathcal{X}),$$

and

$$\pi^{\mathcal{Y}}_{\#}\gamma(B) = \int_{\mathcal{X}}\int_{\mathcal{Z}} \mathbb{1}_B(T(x,z))d\lambda(z)\xi(x)d\mu(x) = T_{\#}(\xi\mu \times \lambda)(B), \quad \forall B \in \mathcal{B}(\mathcal{Y}).$$

It then follows that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} c_1 d\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\psi}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} c_1(x,y) d\gamma_{y|x}(y) \right) d\pi^{\mathcal{X}}_{\#}\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\psi}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&\text{(by Fubini's Theorem for Markov kernels)} \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Z}} c(x, T(x,z)) d\lambda(z) \right) d\pi^{\mathcal{X}}_{\#}\gamma + D_{c_2}(\pi^{\mathcal{X}}_{\#}\gamma|\mu) + D_{\psi}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&\text{(by change of variables)} \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Z}} c(x, T(x,z)) d\lambda(z) \right) d\pi^{\mathcal{X}}_{\#}\gamma + \int c_2\left(\frac{d\pi^{\mathcal{X}}_{\#}\gamma}{d\mu}\right) d\mu(x) + D_{\psi}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Z}} c(x, T(x,z)) d\lambda(z) \right) \xi(x) d\mu(x) + \int c_2(\xi(x)) d\mu(x) + D_{\psi}(\pi^{\mathcal{Y}}_{\#}\gamma|\nu) \\
&\text{(by definition of } \xi) \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Z}} c(x, T(x,z)) d\lambda(z) \right) d\mu(x) + \int c_2(\xi(x)) d\mu(x) + D_{\psi}(T_{\#}(\xi\mu \times \lambda)|\nu) \\
&\geq L_{\psi}(\mu,\nu).
\end{aligned}$$

Since this inequality holds for any γ, this implies that $\tilde{W}_{c_1,c_2,\psi}(\mu,\nu) \geq L_{\psi}(\mu,\nu)$, which completes the proof. □

Due to the near equivalence of the formulations, several theoretical results for (6) follow from the analysis of optimal entropy-transport by Liero et al. (2018), such as the following existence and uniqueness result:

Proposition B.1. *Suppose $L_{\psi}(\mu,\nu) < \infty$, c_2 is convex and lower semi-continuous, and $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \lambda)$ is atomless. 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. If c_2, ψ are strictly convex, $\psi'_{\infty} = \infty$ and c_1 satisfies Corollary 3.6 (Liero et al., 2018), then the joint measure γ specified by any minimizer of $L_{\psi}(\mu,\nu)$ is unique.*

Proof of Proposition B.1. 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 3.3 that a minimizer of $L_{\psi}(\mu,\nu)$ also exists. For uniqueness, if $\psi'_{\infty} = \infty$, then it follows from Lemma 3.5 of (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 in (Liero et al., 2018). It follows from the construction of the proof of Lemma 3.3 that the product measure generated by the minimizers of $L_{\psi}(\mu,\nu)$ is unique, which completes the proof. □

For certain cost functions and divergences, it can be shown that L_ψ defines a proper metric between positive measures μ and ν, i.e. 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.

Based on Lemma 3.3, the theoretical analysis of Liero et al. (2018), and standard results on constrained optimization, it can be shown that for an appropriate and sufficiently large choice of divergence penalty, solutions of the relaxed problem (6) converge to solutions of the original problem (5) (Theorem 3.4).

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 in (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 $\pi^{\mathcal{Y}}_{\#}\gamma = \nu$. Hence

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

for all k. Therefore, $\lim_{k\to\infty}\tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) = \tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu)$, which then by Lemma 3.3 implies the first part of the proposition.

For the second part, by the hypothesis we have that $\tilde{W}_{c_1,c_2,\iota_=}(\mu,\nu) = C < \infty$ and as a consequence $\tilde{W}_{c_1,c_2,\zeta^k\psi}(\mu,\nu) \leq C$ for all k. Hence, by Proposition 2.10 in (Liero et al., 2018), the sequence of minimizers γ^k is bounded. If the assumptions of Proposition B.1 are satisfied, then the sequence γ^k is equally tight. For assumption (ii) this follows by Proposition 2.10 in (Liero et al., 2018) and for assumption (i) this follows 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 $\bar{\gamma}$. Then by lower semicontinuity, we obtain that

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

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 3.3, γ^k is equivalent to the product measure induced by minimizers of $L_{\zeta^k\psi}(\mu,\nu)$, which implies the second part of the proposition. □

B.2 Convex conjugate form of divergences

In this section, we present the convex conjugate form of ψ-divergence used to rewrite the main objective as a min-max problem.

Lemma B.2. *For non-negative finite measures P, Q over $\mathcal{T} \subset \mathbb{R}^d$, it holds that*

$$D_\psi(P|Q) \geq \sup_{f\in\mathcal{F}} \int f dP - \psi^*(f) dQ \tag{13}$$

where $\mathcal{F}$ is a subset of measurable functions $\{f : \mathcal{T} \to (-\infty, \psi'_\infty]\}$. 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 $P^\perp$.

We provide a simple proof of this result. A similar result under stronger assumptions was shown in Nguyen et al. (2008) and used by Nowozin et al. (2016) for generative modeling. A rigorous proof can be found in Liero et al. (2018).

Proof of Lemma B.2. Note that

$$\begin{aligned}
D_\psi(P|Q) &= \psi'_\infty P_\perp(\mathcal{T}) + \int_{\mathcal{T}} \psi\left(\frac{dP}{dQ}\right) dQ \\
&= \psi'_\infty P_\perp(\mathcal{T}) + \int_{\mathcal{T}} \sup_{\xi\in\mathbb{R}} \{\xi \frac{dP}{dQ} - \psi^*(\xi)\} dQ \\
&\text{(by defintion of convex conjugate)} \\
&= \psi'_\infty P_\perp(\mathcal{T}) + \int_{\mathcal{T}} \sup_{\xi\in(-\infty,\psi'_\infty]} \{\xi \frac{dP}{dQ} - \psi^*(\xi)\} dQ \\
&\text{(by Lemma B.3) below} \\
&= \int_{\mathcal{T}} \sup_{\xi\in(-\infty,\psi'_\infty]} \{\xi dP_\perp + \xi \frac{dP}{dQ} dQ - \psi^*(\xi) dQ\} \\
&= \sup_{f\in\mathcal{F}} \int_{\mathcal{T}} f dP - \psi^*(f) dQ.
\end{aligned}$$

By first-order optimality conditions, the optimal f over the support of Q is obtained when $\frac{dP}{dQ}$ belongs to the subdifferential of $\psi^*(f)$, or equivalently 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'_∞, which completes the proof. □

Lemma B.3. *If* $\xi > \psi'_\infty := \lim_{s\to\infty} \frac{\psi(s)}{s}$*, then* $\psi^*(\xi) = \infty$.

Proof.

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

□

C Practical Considerations for Numerical Experiments

Choice of cost functions. Proposition B.1 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 $\lim_{x\to 0} c_2(x) = \lim_{x\to\infty} c_2(x) = \infty$ to prevent ξ from becoming too small or too large. Any of the entropy functions shown in Table 1 are reasonable choices.

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

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).

Choice of ψ. 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 does not generally hold when 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 $\psi(s) = s\log s - (s+1)\log(s+1)$ (Nowozin et al., 2016). 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 non-negative measures with unconstrained total mass, the divergence is minimized when $P = \infty$ and $Q = 0$.

When P, Q are not probability measures, we require an additional constraint on ψ to ensure that divergence minimization matchces P to Q:

Lemma C.2. *Suppose P, Q are non-negative finite measures over $T \subseteq \mathbb{R}^n$. If $\psi(s)$ attains a unique minimum at $s = 1$ with $\psi(1) = 0$ and $\psi'_\infty > 0$, then $D_\psi(P|Q) = 0 \Rightarrow P = Q$. Otherwise, then $P \neq Q$ in general when $D_\psi(P|Q)$ is minimized.*

Proof. Suppose $\psi(s)$ attains a unique minimum at $s = 1$ with $\psi(1) = 0$, $\psi'_\infty > 0$, and $P \neq Q$ over a region with positive measure. It is straightforward to see by the definition in (4) that $D_\psi(P|Q) > 0$, since at least one of the two terms will be strictly positive. Therefore, the first statement holds. For the second statement, suppose either $\psi(s)$ does not attain a unique minimum at $s = 1$ or $\psi'_\infty \leq 0$. If $\psi(s)$ attains a minimum at some $s' \neq 1$, then taking $P = s'Q$ results in a divergence that is equal to or less than $P = Q$. If $\psi'_\infty \leq 0$, then letting $P = Q + P_\perp$ where $P_\perp$ is a positive measure orthogonal to Q results in a divergence that is equal to or less than $P = Q$. □

Table 1 provides some examples of ψ corresponding to common divergences that can be used for unbalanced OT.

Choice of f. According to Lemma B.2, 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, also known as an output activation layer (Nowozin et al., 2016). Table 1 provides some examples of activation layers that can be used.

Choice of neural architectures. For our experiments in Section 4, we used fully-connected feedforward networks with 3 hidden layers and ReLU activations. For T, the output activation layer was a sigmoid function to map the final pixel brightness to the range $(0, 1)$. For ξ, the output activation layer was a softplus function to map the scaling factor weight to the range $(0, \infty)$.

Gradient penalties. The training of GANs using alternating stochastic gradient descent is not guaranteed to converge locally (Mescheder et al., 2018). Stability can be improved using regularization in the form of added instance noise (Arjovsky & Bottou, 2017) or gradient penalties (Gulrajani et al., 2017). We observed that enforcing gradient penalities on ξ and f, while not necessary, improved the stability of Algorithm 1. A gradient penalty on ξ effectively restricts the search to sufficiently smooth ξ, which is reasonable in practice considering that similar regions of the source measure often have similar rates of growth or contraction. A gradient penalty on f changes the nature of the relaxation from (5) to (6): the right-hand side of (7) is no longer equivalent to the ψ-divergence, but is rather a lower-bound with a relation to bounded Lipschitz metrics (Gulrajani et al., 2017). In this case, while the problem formulation is not equivalent to optimal entropy-transport, it is still a valid relaxation of unbalanced Monge OT in (5).

Choice of λ. One can take λ to be the standard Gaussian measure if a stochastic mapping is desired, similar to Almahairi et al. (2018). If a deterministic mapping is desired, then λ is set to a deterministic distribution.

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

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

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 may improve very slowly for some inputs x_i. One way to address this issue 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.