Optimal transport mapping via input convex neural networks

Ashok Vardhan Makkuva [*1] **Amirhossein Taghvaei** [*2] **Jason D. Lee** [3] **Sewoong Oh** [4]

Abstract

In this paper, we present a novel and principled approach to learn the optimal transport between two distributions, from samples. Guided by the optimal transport theory, we learn the optimal Kantorovich potential which induces the optimal transport map. This involves learning two convex functions, by solving a novel minimax optimization. Building upon recent advances in the field of input convex neural networks, we propose a new framework to estimate the optimal transport mapping as the gradient of a convex function that is trained via minimax optimization. Numerical experiments confirm the accuracy of the learned transport map. Our approach can be readily used to train a deep generative model. When trained between a simple distribution in the latent space and a target distribution, the learned optimal transport map acts as a deep generative model. Although scaling this to a large dataset is challenging, we demonstrate two important strengths over standard adversarial training: robustness and discontinuity. As we seek the optimal transport, the learned generative model provides the same mapping regardless of how we initialize the neural networks. Further, a gradient of a neural network can easily represent discontinuous mappings, unlike standard neural networks that are constrained to be continuous. This allows the learned transport map to match any target distribution with many discontinuous supports and achieve sharp boundaries.

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. [2]Department of Mechanical and Aerospace Engineering, University of California, Irvine. [3]Department of Electrical Engineering, Princeton University, [4]Allen School of Computer Science & Engineering, University of Washington. Correspondence to: Ashok <makkuva2@illinois.edu>, Amir <amirhoseintghv@gmail.com>.

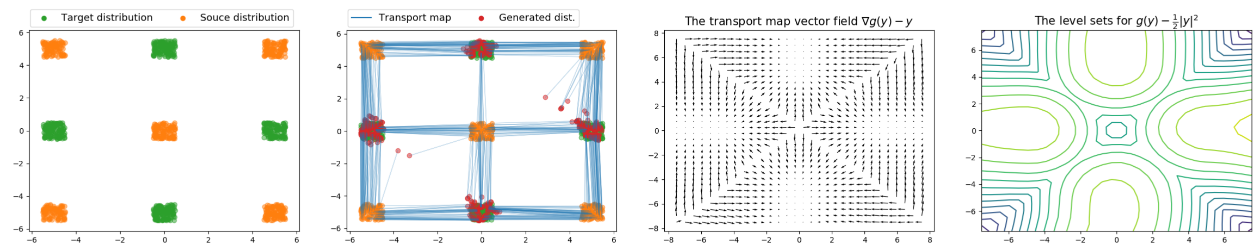
Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1. Introduction

Finding a mapping that transports mass from one distribution Q to another distribution P is an important task in various machine learning applications, such as deep generative models (Goodfellow et al., 2014; Kingma & Welling, 2013) and domain adaptation (Gopalan et al., 2011; Ben-David et al., 2010). Among infinitely many transport maps T that can map a random variable X from Q such that $T(X)$ is distributed as P, several recent advances focus on discovering some inductive bias to find a transport map with desirable properties. Research in optimal transport has been leading such efforts, in applications such as color transfer (Ferradans et al., 2014), shape matching (Su et al., 2015), data assimilation (Reich, 2013), and Bayesian inference (El Moselhy & Marzouk, 2012). Searching for an optimal transport encourages a mapping that minimizes the total cost of transporting mass from Q to P, as originally formulated in Monge (1781), and provides the inductive bias needed in many such applications. However, finding the optimal transport map in general is a challenging task, especially in high dimensions where efficient approaches are critical.

Algorithmic solutions are well-established for discrete variables; the optimal transport can be found as a solution to linear program. Building upon this mature area, typical approaches for general distributions use quantization of the space, which becomes computationally intensive for high-dimensional variables we encounter in modern applications (Evans & Gangbo, 1999; Benamou & Brenier, 2000; Papadakis et al., 2014).

There has been a vast amount of works to extend the computation of the optimal transport map to high-dimensional setting (Seguy et al., 2017; Genevay et al., 2016; Xie et al., 2019; Liu et al., 2018; Chen et al., 2019). In this paper, we propose a novel minimax optimization approach to search for the optimal transport under the quadratic distance (i.e. 2-Wassertstein metric). A major challenge in a minimax formulation of optimal transport is that the constraints in the Kantorovich dual formulation (3) are notoriously challenging. They require the evaluation of the functions at every point in the domain, which is not tractable. A common straightforward heuristics sample some points and add those sampled constraints as regularizers. Such regularizations create biases that hinder learning the true optimal transport.



(a) Data samples (b) Our transport map (c) Displacement vector field (d) Level sets

Figure 1. Results on Checkerboard dataset. (a) Samples from the source (orange) and target (green) distributions; (b) The learned transport map and the generated distribution, via Algorithm 1; (c) The learned displacement vector field generated by $\nabla g(y) - y$; (d) The level sets of the original dual variable $g(y) - \frac{1}{2}|y|^2$. The experimental details are included in Section 4.1.

Our key innovation is to depart from this common practice; we instead eliminate the constraints by restricting our search to the set of all convex functions, building upon the fundamental connection from Brenier's Theorem 3.1. This leads to a novel minimax formulation in (5). Leveraging on recent advances in input convex neural networks, we propose a new architecture and a training algorithm for solving this minimax optimization. We establish the consistency of our proposed minimax formulation in Theorem 3.3. In particular, we show that the solution to this optimization problems yields the exact optimal transport map. We provide stability analysis for the proposed estimator in Theorem 3.6.

Further, when used to train deep generative models, our approach can be viewed as a novel framework to train a generator that is modeled as a *gradient of a convex function*. We provide a principled training rule based on the optimal transport theory. This ensures that (i) the generator converges to the optimal transport, independent of how we initialize the neural network; and (ii) represent sharp boundaries when the target has multiple disconnected supports. Gradient of a neural network naturally represents discontinuous functions, which is critical in mapping from a single connected support to disconnected supports.

To model convex functions, we leverage Input Convex Neural Networks (ICNNs), a class of scalar-valued neural networks $f(x;\theta)$ such that the function $x \mapsto f(x;\theta) \in \mathbb{R}$ is convex. These neural networks were introduced by Amos et al. (2016) to provide efficient inference and optimization procedures for structured prediction, data imputation and reinforcement learning tasks. In this paper, we show that ICNNs can be efficiently trained to learn the optimal transport map between two distributions P and Q. To the best of our knowledge, this is the first such instance where ICNNs are leveraged for the well-known task of learning optimal transport maps in a *scalable* fashion. This framework opens up a new realm for understanding problems in optimal transport theory using parametric convex neural networks, both in theory and practice. Figure 1 provides an example where the optimal transport map has been learned via our proposed Algorithm 1 from the orange distribution to the green distribution.

Notation. $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on a Polish space $\mathcal{X}$, and $\mathcal{B}(\mathcal{X})$ denotes the Borel subsets of $\mathcal{X}$. For $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, $P \otimes Q$ denotes the product measure on $\mathcal{X} \times \mathcal{Y}$. For measurable map $T : \mathcal{X} \to \mathcal{Y}$, $T_{\#}P$ denotes the push-forward of P under T, i.e. $(T_{\#}P)(A) = P(T^{-1}(A))$, $\forall A \in \mathcal{B}(\mathcal{Y})$. $L^1(P) \triangleq \{f \text{ is measurable } \& \int f \, \mathrm{d}P < \infty\}$ denotes the set of integrable functions with respect to P. $\texttt{CVX}(P)$ denotes the set of all convex functions in $L^1(P)$. $\mathrm{Id} : x \mapsto x$ denotes the identity function. $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the inner-product and ℓ_2-Euclidean norm.

2. Background on optimal transport

Let P and Q be two probability distributions on $\mathbb{R}^d$ with finite second order moments. The *Monge's optimal transportation problem* is to transport the probability mass under Q to P with the least amount of cost[1], i.e.

$$\underset{T:T_{\#}Q=P}{\text{minimize}} \quad \frac{1}{2}\mathbb{E}_{X\sim Q}\|X - T(X)\|^2. \tag{1}$$

Any transport map T achieving the minimum in (1) is called *optimal transport map*. Optimal transport map may not exist. In fact, the feasible set in the above optimization problem may itself be empty, for example when Q is a Dirac distribution and P is any non-Dirac distribution.

The Monge problem (1) is highly nonlinear and difficult to analyze. Kantorovich introduced a relaxation of the problem,

$$W_2^2(P,Q) \triangleq \inf_{\pi \in \Pi(P,Q)} \frac{1}{2}\mathbb{E}_{(X,Y)\sim\pi}\|X - Y\|^2, \tag{2}$$

where $\Pi(P,Q)$ denotes the set of all joint probability distributions (or equivalently, couplings) whose first and second marginals are P and Q, respectively. The optimal value

[1] In general, Monge's problem is defined in terms of cost function $c(x,y)$. This paper is concerned with quadratic cost function $c(x,y) = \frac{1}{2}\|x-y\|^2$ because of its nice geometrical properties and connection to convex analysis (Villani, 2003, Ch. 2).

in (2) is the 2-Wasserstein distance $W_2(\cdot,\cdot)$ squared. Any coupling π achieving the infimum is called the *optimal coupling*. Optimization problem (2) is also referred to as the *primal formulation* for 2-Wasserstein distance.

Kantorovich also provided a dual formulation for (2), known as the Kantorovich duality (Villani, 2003, Theorem 1.3),

$$W_2^2(P,Q) = \sup_{(f,g)\in\Phi_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)], \qquad (3)$$

where Φ_c denotes the constrained space of functions, defined as $\Phi_c \triangleq \big\{(f,g) \in L^1(P)\times L^1(Q):\ f(x)+g(y) \le \frac{1}{2}\|x-y\|_2^2, \quad \forall (x,y)\ dP\otimes dQ \text{ a.e.}\big\}$.

The dual problem (3) can be recast as an stochastic optimization problem by approximating the expectations using independent samples from P and Q. However, there is no easy way to ensure the feasibility of the constraint $(f,g)\in\Phi_c$ along the gradient updates. Common approach is to translate the optimization into a tractable form, while sacrificing the original goal of finding the optimal transport map. Concretely, an entropic or a quadratic regularizer is added to the primal problem (2) (Cuturi, 2013; Essid & Solomon, 2018; Peyré et al., 2019; Blondel et al., 2017). Then, the dual to the regularized primal problem is an unconstrained version of (3) with additional penalty term. The unconstrained problem can be numerically solved using Sinkhorn algorithm in discrete setting (Cuturi, 2013) or stochastic gradient methods with suitable function representation in continuous setting (Genevay et al., 2016; Seguy et al., 2017). The optimal transport can then be obtained from f and g, using the first-order optimality conditions of the Fenchel-Rockafellar's duality theorem (Seguy et al., 2017), or by training a generator through an adversarial computational procedure (Leygonie et al., 2019).

In this paper, we take a different approach: solve the dual problem without introducing a regularization. This builds upon (Taghvaei & Jalali, 2019), where ICNN for the task of approximating the Wasserstein distance and optimal transport map is originally proposed. We bring the idea proposed (Taghvaei & Jalali, 2019) into practice by introducing a novel minimax optimization formulation. We describe our proposed method in Section 3 and provide a detailed comparison in Remark 3.5. Discussion about other related works (Lei et al., 2017; Guo et al., 2019; Xie et al., 2019; Muzellec & Cuturi, 2019; Rabin et al., 2011; Korotin et al., 2019; Liu et al., 2018; Chen et al., 2019) appears in Appendix **??**.

3. A novel minimax formulation to learn optimal transport

Our goal is to learn the optimal transport map T^* from Q to P, from samples drawn from P and Q, respectively. We use the fundamental connection between optimal transport and Kantorovich dual in Theorem 3.1, to formulate learning T^* as a problem of estimating $W_2^2(P,Q)$. However, $W_2^2(P,Q)$ is notoriously hard to estimate. The standard Kantorovich dual formulation in Eq. (3) involves a supremum over a set Φ_c with a pointwise constraints, which is challenging to even approximately project onto. To this end, we derive an alternative optimization formulation in Eq. (5), inspired by the convexification trick (Villani, 2003, Section 2.1.2). This allows us to eliminate the distance constraint of Φ_c, and instead constrain our search over all *convex functions*. This constrained optimization can now be seamlessly integrated with recent advances in designing deep neural architectures with convexity guarantees. This leads to a novel minimax optimization to learn the optimal transport.

We exploit the fundamental properties of $W_2^2(P,Q)$ and the corresponding optimal transport to reparametrize the optimization formulation. Note that for any $(f,g)\in\Phi_c$,

$$\begin{aligned} & f(x)+g(y) \le \frac{1}{2}\|x-y\|_2^2 \iff \\ & \left[\frac{1}{2}\|x\|_2^2 - f(x)\right] + \left[\frac{1}{2}\|y\|_2^2 - g(y)\right] \ge \langle x,y\rangle. \end{aligned}$$

Hence reparametrizing $\frac{1}{2}\|\cdot\|_2^2 - f(\cdot)$ and $\frac{1}{2}\|\cdot\|_2^2 - g(\cdot)$ by f and g respectively, and substituting them in (3) yields

$$W_2^2(P,Q) = C_{P,Q} - \inf_{(f,g)\in\widetilde{\Phi}_c} \Big\{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]\Big\},$$

where $C_{P,Q} = (1/2)\mathbb{E}[\|X\|_2^2 + \|Y\|_2^2]$ is a constant independent of (f,g) and $\widetilde{\Phi}_c \triangleq \{(f,g)\in L^1(P)\times L^1(Q):\ f(x)+g(y)\ge\langle x,y\rangle, \quad \forall(x,y)\ dP\otimes dQ \text{ a.e.}\}$. While the above constrained optimization problem involves a pair of functions (f,g), it can be transformed into the following form involving only a single convex function f, thanks to Villani (2003, Theorem 2.9):

$$W_2^2(P,Q) = C_{P,Q} - \inf_{f\in\mathtt{CVX}(P)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)], \quad (4)$$

where $f^*(y) = \sup_x \langle x,y\rangle - f(x)$ is the convex conjugate of $f(\cdot)$.

The crucial tools behind our formulation are the following celebrated results due to Knott-Smith and Brenier (Villani, 2003), which relate the optimal solutions for the dual form in (4) and the primal form in (2).

Theorem 3.1 ((Villani, 2003, Theorem 2.12))**.** *Let P,Q be two probability distributions on $\mathbb{R}^d$ with finite second order moments. Then,*

1. *(**Knott-Smith optimality criterion**) A coupling $\pi \in \Pi(P,Q)$ is optimal for the primal (2) if and only if there exists a convex function $f\in\mathtt{CVX}(\mathbb{R}^d)$ such that $\mathrm{Supp}(\pi)\subset\mathrm{Graph}(\partial f)$. Or equivalently, for all $d\pi$-almost (x,y), $y\in\partial f(x)$. Moreover, the pair (f,f^*) achieves the minimum in the dual form (4).*

2. *(**Brenier's theorem**) If Q admits a density with respect to the Lebesgue measure on $\mathbb{R}^d$, then there is a unique optimal coupling π for the primal problem. In particular, the optimal coupling satisfies that*

$$d\pi(x,y) = dQ(y)\delta_{x=\nabla f^*(y)},$$

where the convex pair (f, f^*) achieves the minimum in the dual problem (4). Equivalently, $\pi = (\nabla f^* \times \mathrm{Id})_{\#}Q$.

3. *Under the above assumptions of Brenier's theorem, ∇f^* in the unique solution to Monge transportation problem from Q to P, i.e.*

$$\mathbb{E}_Q\|\nabla f^*(Y) - Y\|^2 = \inf_{T:T_{\#}Q=P} \mathbb{E}_Q\|T(Y) - Y\|^2.$$

Remark 3.2. Whenever Q admits a density, we refer to ∇f^* as the optimal transport map.

Henceforth, throughout the paper we assume that the distribution Q admits a density in $\mathbb{R}^d$. Note that in view of Theorem 3.1, any optimal pair (f, f^*) from the dual formulation in (4) provides us an optimal transport map ∇f^* pushing forward Q onto P. However, optimizing the objective (4) is challenging because it requires to compute the conjugate function f^*. To this end, we propose a novel minimax formulation in the following theorem where we replace the conjugate with a new convex function.

Theorem 3.3. *Whenever Q admits a density in $\mathbb{R}^d$, we have*

$$W_2^2(P,Q) = \sup_{\substack{f\in\mathtt{CVX}(P),\\ f^*\in L^1(Q)}} \inf_{g\in\mathtt{CVX}(Q)} \mathcal{V}_{P,Q}(f,g) + C_{P,Q}, \quad (5)$$

where $\mathcal{V}_{P,Q}(f,g)$ is a functional of f, g defined as

$$\mathcal{V}_{P,Q}(f,g) = -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))].$$

In addition, there exists an optimal pair (f_0, g_0) achieving the infimum and supremum respectively, where ∇g_0 is the optimal transport map from Q to P.

Proof sketch. The proof follows from the inequality $\langle y, \nabla g(y)\rangle - f(\nabla g(y)) \leq f^*(y)$ for all functions g, and then taking the expectation over Q, and observing that the equality is achieved with $g = f^*$. The technical details appear in Appendix **??**. □

Remark 3.4. For any convex function f, the function $g \in L^1(Q)$ that achieves the infimum in (5) is convex and equals f^*. Therefore, the constraint $g \in \mathtt{CVX}(Q)$ can be relaxed to $g \in L^1(Q)$ without changing the optimal value and optimizing functions. We numerically observe that the optimization algorithm performs better under this relaxation.

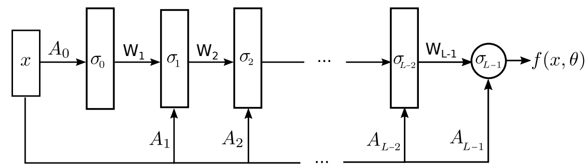


Figure 2. The input convex neural network (ICNN) architecture.

Formulation (5) now provides a principled approach to learn the optimal transport mapping $\nabla g(\cdot)$ as a solution of a minimax optimization. Since the optimization involves the search over the space of convex functions, we utilize the recent advances in input convex neural networks (ICNNs) to parametrize them as discussed in the following section.

3.1. Minimax optimization over ICNNs

We propose using parametric models based on deep neural networks to approximate the set of convex functions. This is known as input convex neural networks (Amos et al., 2016), denoted by $\mathtt{ICNN}(\mathbb{R}^d)$. We propose estimating the following approximate Wasserstein-2 distance, from samples:

$$\widetilde{W}_2^2(P,Q) = \sup_{f\in\mathtt{ICNN}(\mathbb{R}^d)} \inf_{g\in\mathtt{ICNN}(\mathbb{R}^d)} \mathcal{V}_{P,Q}(f,g) + C_{P,Q}. \quad (6)$$

ICNNs are a class of scalar-valued neural networks $f(x;\theta)$ such that the function $x \mapsto f(x;\theta) \in \mathbb{R}$ is convex.

The neural network architecture for an ICNN is as follows. Given an input $x \in \mathbb{R}^d$, the mapping $x \mapsto f(x;\theta)$ is given by a L-layer feed-forward NN using the following equations for $l = 0, 1, \ldots, L-1$:

$$z_{l+1} = \sigma_l(W_l z_l + A_l x + b_l), \quad f(x;\theta) = z_L,$$

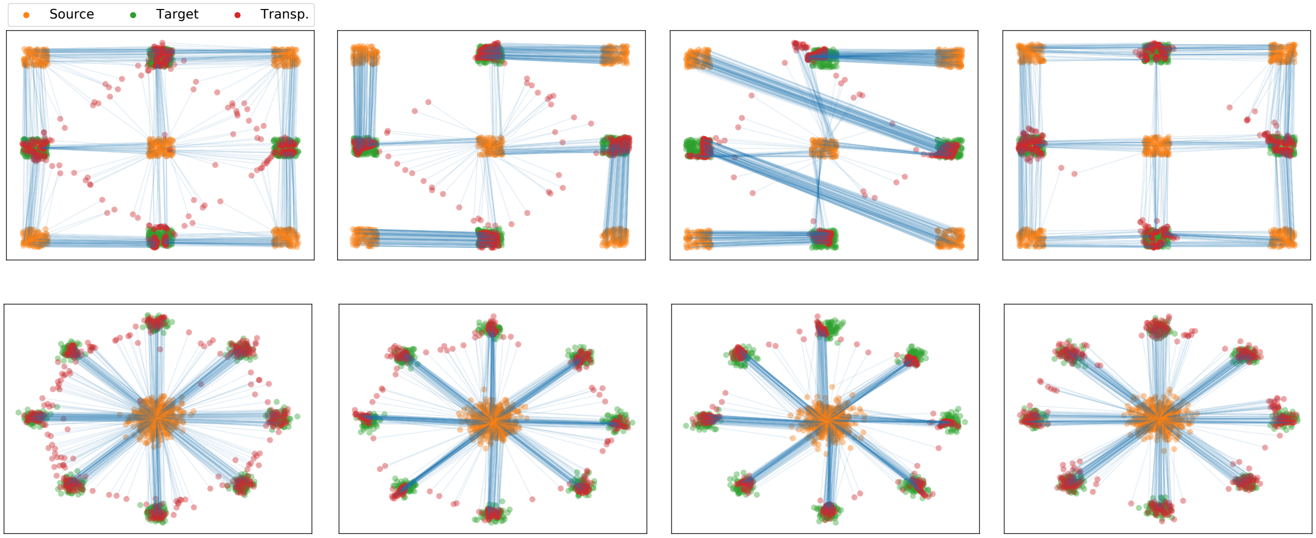
where $\{W_l\}$, $\{A_l\}$ are weight matrices (with the convention that $W_0 = 0$), and $\{b_l\}$ are the bias terms. σ_l denotes the entry-wise activation function at the layer l. This is illustrated in Figure 2. We denote the total set of parameters by $\theta = (\{W_l\}, \{A_l\}, \{b_l\})$. It follows from Amos et al. (2016, Proposition 1) that $f(x;\theta)$ is convex in x provided

(i) all entries of the weights W_l are non-negative;

(ii) activation function σ_0 is convex;

(iii) σ_l is convex and non-decreasing, for $l = 1, \ldots, L-1$.

While ICNNs are a specific parametric class of convex functions, it is important to understand if this class is rich enough representationally. This is answered positively by Chen et al. (2018, Theorem 1). In particular, they show that any convex function over a compact domain can be approximated in sup norm by a ICNN to the desired accuracy. This justifies the choice of ICNNs as a suitable approximating class for the convex functions.



(a) Barycentric-OT (b) W1-LP (c) W2GAN (d) Our approach

Figure 3. The transport maps learned by various approaches on 'Checker board' and 'mixture of eight Gaussians' datasets. (a) Barycentric-OT (Seguy et al., 2017); (b) W1-LP (Petzka et al., 2017); (c) W2-GAN (Leygonie et al., 2019); (d) Our approach (Algorithm 1). The source distribution Q is highlighted in orange, target distribution P in green, the transported distribution $T_{\#}Q$ in red, and the transport map with blue arrows.

The proposed framework for learning the optimal transport provides a novel training method for deep generative models, where (a) the generator is modeled as a gradient of a convex function and (b) the minimax optimization in (6) (and more concretely, Algorithm 1) provides the training methodology. On the surface, Eq. (6) resembles the minimax optimization of generative adversarial networks based on Wasserstein-1 distance (Arjovsky et al., 2017), called WGAN. However, there are several critical differences making our approach attractive.

First, because WGANs use optimal transportation distance only as a measure of distance, the learned generator map from the latent source to the target is arbitrary and sensitive to the initialization (see Figure 4) (Jacob et al., 2018). Sensitivity to the initialization is observed o lead to mode collapse in Stacked MNIST experiment (Lin et al., 2018). On the other hand, our proposed approach aims to find the *optimal* transport map and learns the same mapping regardless of the initialization (see Figure 1).

Secondly, in a WGAN architecture (Arjovsky et al., 2017; Petzka et al., 2017), the transport map (which is the generator) is represented with neural network that is a continuous mapping. Although, a discontinuous map can be approximated arbitrarily close with continuous neural networks, such a construction requires large weights making training unstable. On the other hand, through our proposed method, by representing the transport map with *gradient* of a neural network (equipped with ReLU type activation functions), we obtain a naturally *discontinuous map*. As a consequence we have sharp transition from one part of the support to the other, whereas GANs (including WGANs) suffer from spurious probability masses that are not present in the target. This is illustrated in Section 4.3. The same holds for regularization-based methods for learning optimal transport (Genevay et al., 2016; Seguy et al., 2017; Leygonie et al., 2019), where transport map is parametrized by continuous neural nets.

Remark 3.5. In a recent work, Taghvaei & Jalali (2019) proposed to solve the semi-dual optimization problem (4) by representing the function f with an ICNN and learning it using a stochastic optimization algorithm. However, each step of this algorithm requires computing the conjugate f^* for all samples in the batch via solving a inner convex optimization problem for each sample which makes it slow and challenging to scale to large datasets. Further it is memory intensive as each inner optimization step requires a copy of all the samples in the dataset. In contrast, we represent the convex conjugate f^* using ICNN and present a novel minimax formulation to learn it, in a scalable manner.

3.2. Stability analysis of the learned transport map

Theorem 3.3 establishes the consistency of our proposed optimization: if the objective (5) is solved exactly with a pair of functions (f_0, g_0), then ∇g_0 is the exact optimal transport map from Q to P. In this section, we study the error in approximating the optimal transport map ∇g_0, when the objective (5) is solved up to a small error. To this end, we build upon the recent results from Hütter & Rigollet

(2019, Prop. 8) regarding the stability of optimal transport maps.

Recall that the optimization objective (5) involves a minimization and a maximization. For any pair (f, g), let $\epsilon_1(f, g)$ denote the minimization gap and $\epsilon_2(g)$ denote the maximization gap, defined according to:

$$\begin{aligned}\epsilon_1(f,g) &= \mathcal{V}(f,g) - \inf_{\widetilde{g}\in \mathtt{CVX}(Q)} \mathcal{V}(f,\widetilde{g}), \\ \epsilon_2(f) &= \sup_{\widetilde{f}\in \mathtt{CVX}(P)} \inf_{\widetilde{g}\in \mathtt{CVX}(Q)} \mathcal{V}(\widetilde{f},\widetilde{g}) - \inf_{\widetilde{g}\in \mathtt{CVX}(Q)} \mathcal{V}(f,\widetilde{g})\end{aligned} \tag{7}$$

Then, the following theorem bounds the the error between ∇g and the optimal transport map ∇g_0 as a function ϵ_1 and ϵ_2. We defer its proof to Appendix **??**.

Theorem 3.6. *Consider the optimization problem (5). Assume Q admits a density and let $\nabla g_0(\cdot)$ denote the optimal transport map from Q to P. Then for any pair (f, g) such that f is strongly convex, we have*

$$\|\nabla g - \nabla g_0\|^2_{L^2(Q)} \le \frac{2}{\alpha}(\epsilon_1(f,g) + \epsilon_2(f)),$$

where ϵ_1 and ϵ_2 are defined in (7), and $\|\cdot\|_{L^2(Q)}$ denotes the L^2-norm with respect to measure Q.

Remark 3.7. *Dependency in Thm. 3.6 on α is fundamental. For small α, there are examples where the optimization gaps are small but the error in LHS is large. Further, Thm. 3.6 captures the difficulty of training if true distribution has supports that are (nearly) disjoint. In this case, the optimal function f is not strongly convex and hence the upper-bound is large. Numerically, in order to search over strong convex f, we can add a small quadratic term to the output of ICNN.*

4. Experiments

In this section, first we qualitatively illustrate our proposed approach (see Figure 3) on the following two-dimensional synthetic datasets: (a) Checkerboard, (b) Mixture of eight Gaussians. We compare our method with the following three baselines: (i) Barycentric-OT (Seguy et al., 2017), (ii) W1-LP, which is the state-of-the-art Wasserstein GAN introduced by (Petzka et al., 2017), (iii) W2GAN (Leygonie et al., 2019). Note that while the goal of W1-LP is not to learn the optimal transport map, the generator obtained at the end of its training can be viewed as a transport map. For all these baselines, we use the implementations (publicly available) of Leygonie et al. (2019) which has the best set of parameters for each of these methods. In Section 4.2 and Section 4.3, we highlight the respective robustness and the discontinuity of our transport maps as opposed to other approaches. Finally, in Section 4.4, we show the effectiveness of our approach on the challenging task of learning the optimal transport map on a variety of synthetic and real world high-dimensional data. Full experimental details are provided in Appendix **??**.

Training methodology. We utilize our minimax formulation in (6) to learn the optimal transport map. We parametrize the convex functions f and g using the same ICNN architecture (Figure 2). Recall that to ensure convexity, we need to restrict all weights W_ℓ's to be non-negative (Assumption (i) in ICNN). We enforce it strictly for f, as the maximization over g can be unbounded, making optimization unstable, whenever f is non-convex. However, we relax this constraint for g (as permitted according to Remark 3.4) and instead introduce a regularization term

$$R(\theta_g) = \lambda \sum_{W_l \in \theta_g} \|\max(-W_l, 0)\|^2_F, \tag{8}$$

where $\lambda > 0$ is a regularization constant and the maximum is taken entry-wise for all the weight parameters $\{W_l\} \subset \theta_g$. We empirically observe that this relaxation makes the optimization converge faster.

For both the maximization and minimization updates in (6), we use Adam (Kingma & Ba, 2014). At each iteration, we draw a batch of samples from P and Q denoted by $\{X_i\}_{i=1}^M$ and $\{Y_j\}_{j=1}^M$ respectively. Then, we use the following objective for optimization which is an empirical counterpart of (6):

$$\max_{\theta_f : W_\ell \ge 0, \forall \ell \in [L-1]} \min_{\theta_g} \; J(\theta_f, \theta_g) + R(\theta_g), \tag{9}$$

where θ_f, θ_g are the parameters of f and g, respectively, $W_\ell \ge 0$ is an entry-wise constraint, and

$$J(\theta_f,\theta_g) = \frac{1}{M}\sum_{i=1}^{M} f(\nabla g(Y_i)) - \langle Y_i, \nabla g(Y_i)\rangle - f(X_i).$$

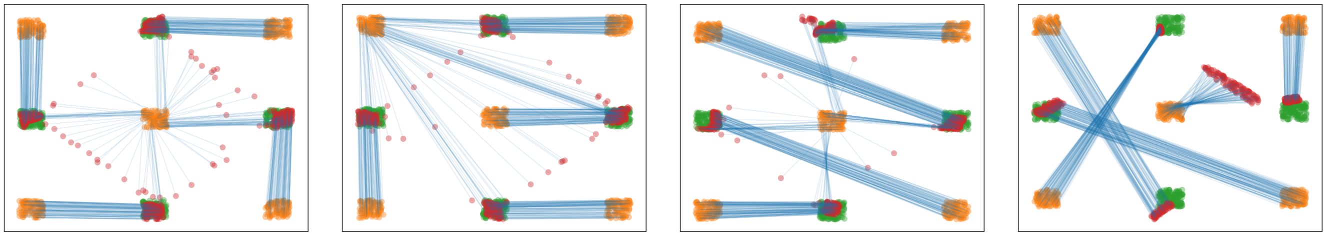
This is summarized in Algorithm 1. In the remainder of the paper, we interchangeably refer to Algorithm 1 as either 'Our approach' or 'Our algorithm'[2].

Remark 4.1. Note that the regularization term $R(\theta_g)$ is *data-independent* and does not introduce any bias to the optimization problem. For any convex function f, the minimizer of the problem (9) is still a convex function g as discussed in Remark 3.4. We use this regularization to guide the algorithm towards neural networks that are convex.

4.1. Learning the optimal transport map

As highlighted in Figure 1 and Figure 3d, qualitatively, we observe that our proposed procedure indeed learns the optimal transport map on both the Checkerboard and Mixture of eight Gaussians datasets. In particular, our transport map is

[2] Source code is available at https://github.com/AmirTag/OT-ICNN.



(a) W1-LP: Trial 1 (b) W1-LP: Trial 2 (c) W2GAN: Trial 1 (d) W2GAN: Trial 2

Figure 4. Transport maps learned by W1-LP (Petzka et al., 2017) and W2GAN (Leygonie et al., 2019) under different random initialization.

Algorithm 1 The numerical procedure to solve the optimization problem (9).

Input: Source dist. Q, Target dist. P, Batch size M, Generator iterations K, Total iterations T
for $t = 1, \ldots, T$ **do**
  Sample batch $\{X_i\}_{i=1}^{M} \sim P$
  for $k = 1, \ldots, K$ **do**
    Sample batch $\{Y_i\}_{i=1}^{M} \sim Q$
    Update θ_g to minimize (9) using Adam method
  end for
  Update θ_f to maximize (9) using Adam method
  Projection: $w \leftarrow \max(w, 0)$, for all $w \in \{W^l\} \in \theta_f$
end for

able to cut the continuous mass symmetrically and transport it to the nearest target support in both these examples. Also, Figure 3 illustrates the qualitative difference of our approach compared to other approaches, in terms of non-optimality and existence of trailing dots. The existence of trailing dots is due to representing the transport map with continuous neural networks, discussed in Section 4.3.

4.2. Robustness of learning transport maps

In this section we numerically illustrate that the generator in W1-LP and W2GAN finds arbitrary transport maps, and it is sensitive to initialization as discussed in Section 3. This is in stark contrast with our proposed approach which finds the *optimal* transport independent of the initialization. We consider the previous Checkerboard example (Figure 1a) and train W1-LP and W2GAN with different random initializations. The resulting transport maps for two different random trials are depicted in Figure 4a and Figure 4b for W1-LP, and Figure 4c and Figure 4d for W2-GAN. In addition to the fact that the learned transport map is very sensitive to initializations, the quality of the samples generated by those trained models are also sensitive. This is a major challenge in training GANs (Lin et al., 2018).

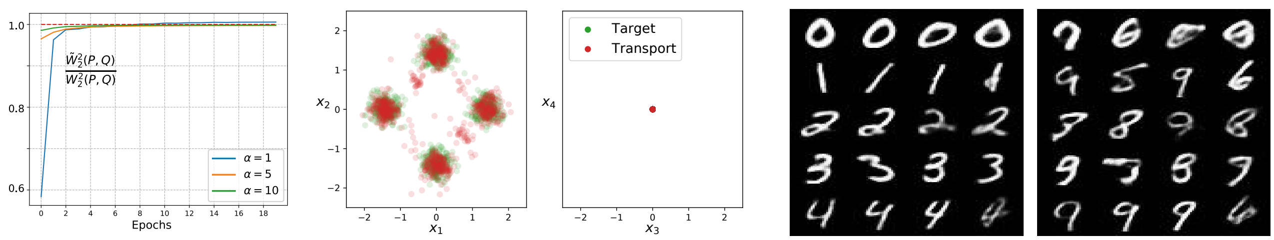
4.3. Learning discontinuous transport maps

The power to represent a discontinuous transport mapping is what fundamentally sets our proposed method apart from the existing approaches, as discussed in Section 3. Two prominent approaches for learning transport maps are generative adversarial networks (Arjovsky et al., 2017; Petzka et al., 2017) and regularized optimal transport (Genevay et al., 2016; Seguy et al., 2017). In both cases, the transport map is modeled by a standard neural network with finite depth and width, which is a continuous function. As a consequence, continuous transport maps suffer from unintended and undesired spurious probability mass that connects disjoint supports of the target probability distribution.

First, standard GANs including the original GAN (Goodfellow et al., 2014) and variants of WGAN (Arjovsky et al., 2017; Gulrajani et al., 2017; Wei et al., 2018) all suffer from spurious probability masses. Even those designed to tackle such spurious probability masses, like PacGAN (Lin et al., 2018), cannot overcome the barrier of continuous neural networks. This suggests that fundamental change in the architecture, like the one we propose, is necessary. Figure 3b illustrates the same scenario for the transport map learned through the WGAN framework. We can observe the trailing dots of spurious probability masses, resulting from undesired continuity of the learned transport maps.

Similarly, regularization methods to approximate optimal transport maps, explained in Section 2, suffer from the same phenomenon. Representing a transport map with an inherently continuous function class results in spurious probability masses connecting disjoint supports. Figure 3a, corresponding to Barycentric-OT, illustrates those trailing dots of spurious masses for the learned transport map from algorithm introduced in Seguy et al. (2017). We also observe a similar phenomenon with Leygonie et al. (2019) as illustrated in Figure 3c.

On the other hand, we represent the transport map with the *gradient* of a neural network (equipped with non-smooth ReLU type activation functions). The resulting transport map can naturally represent discontinuous transport maps,



(a) Estimated distance (b) High-dim Gaussian to 2-dim mixture (c) Source: First 5 digits (d) Transported samples

Figure 5. Numerical results on high-dimensional experiments of Section 4.4: (a) Convergence of our estimated W_2 distance to the actual value when transporting $\mathcal{N}(0, I_d)$ to $\mathcal{N}(\alpha 1, I_d)$ where $d = 784$; (b) Transporting a 784-dim Gaussian to a 2-dim Gaussian mixture embedded in 784-dim space; (c) Samples from the source distribution corresponding to first five MNIST digits, embedded into 16-dim. feature space. (d) Image of the samples under the learned optimal transport map, where the target distribution is the last five digits.

as illustrated in Figure 1b and Figure 3d. The vector field of the learned transport map in Figure 1c clearly shows the discontinuity of the learned optimal transport. The spurious points in Figure 3d are due to finite sample effects and are expected to decrease with more training samples.

4.4. High dimensional experiments

We consider the challenging task of learning optimal transport maps on high dimensional distributions. In particular, we consider both synthetic and real world high dimensional datasets and provide quantitative and qualitative illustration of the performance of our proposed approach.

Gaussian to Gaussian. Source distribution $Q = \mathcal{N}(0, I_d)$ and target distribution $P = \mathcal{N}(\mu, I_d)$, for some fixed $\mu \in \mathbb{R}^d$ and $d = 784$. The mean vector $\mu = \alpha(1, \dots, 1)^\top$ for some parameter $\alpha > 0$. Because both distributions are Gaussian, the optimal transport map is explicitly known: $T^*(x) = x + \mu$ and hence $W_2^2(P, Q) = \|\mu\|^2/2 = \alpha^2 d/2$. In Figure 5a, we compare our estimated distance $\widetilde{W}_2^2(P, Q)$, defined in (6), with the exact value $W_2^2(P, Q)$, as the training progresses for various values of $\alpha \in \{1, 5, 10\}$. Intuitively, learning is more challenging when α is larger. Further, error in learning the optimal transport map, quantified with the metric $\|\mu_{T(Q)} - \mu\|^2$, where $\mu_{T(Q)}$ is the mean of the transported distribution $T_\# Q$, is reported in Table 1. For comparison, the result using the W2GAN approach is included.

Table 1. The error between the mean of transported and that of the target distributions. The source and target are 728-dim. Gaussians.

Approach	$\alpha = 1$	$\alpha = 5$	$\alpha = 10$
our approach	0.19 ± 0.015	13.95 ± 1.45	29.05 ± 5.16
W2GAN	1.30	37.9	66.7

High-dim. Gaussian to low-dim. mixture. Source distribution Q is standard Gaussian $\mathcal{N}(0, I_d)$ with $d = 784$, and the target distribution P is a mixture of four Gaussians that lie in in the two-dimensional subspace of the high-dimensional space $\mathbb{R}^d$, i.e. the first two components of the random vector $X \sim P$ is mixture of four Gaussians, and the rest of the components are zero. The projection of the learned optimal transport map onto the first four components is depicted in Figure 5b. As illustrated in the left panel of 5b, our transport map correctly maps the source distribution to the mixture of four Gaussians in the first two components. And it maps the rest of the components to zero, as highlighted by a red blob at zero in the right panel.

MNIST $\{0, 1, 2, 3, 4\}$ to MNIST $\{5, 6, 7, 8, 9\}$. We consider the standard MNIST dataset (LeCun et al., 1998) with the goal of learning the optimal transport map from the set of images corresponding to first five digits $\{0, 1, 2, 3, 4\}$ to the last five digits $\{5, 6, 7, 8, 9\}$. To achieve this, we embed the images into the a space where the Euclidean norm $\| \cdot \|$ between the embedded images is meaningful. This is in alignment with the reported results in the literature for learning the L_2-optimal transport map (Yang & Karniadakis, 2019, Sec. 4.1). We consider the embeddings into a 16-dimensional latent feature space given by a pre-trained Variational Autoencoder (VAE). We simulate our algorithm on this feature space. The results of the learned transport map are depicted in Figure 5. Figure 5c presents samples from the source distribution and Figure 5d illustrates the source samples after transportation under the learned optimal transport map. We observe that the digits that look alike are coupled via the optimal transport map, e.g. $1 \to 9$, $2 \to 8$, and $4 \to 9$.

Gaussian to MNIST. The source is 16-dimensional standard Gaussian distribution, and the target is the 16-dimensional latent embeddings of all the MNIST digits. The MNIST like samples that are generated from the learned optimal transport map are depicted in Figure 6.

Figure 6. MNIST like samples generated by the learned optimal transport map from Gaussian source distribution in feature space.

These experiments serve as a proof of concept that the algorithm scales to high-dimensional setting and real-world dataset. We believe that further improvements on the performance of the proposed algorithm requires careful tuning of hyper-parameters which takes time to develop (similar to initial WGAN) and is a subject of ongoing work.

5. Conclusion

We presented a novel minimax framework to learn the optimal transport map under W_2-metric. Our framework is in contrast to regularization-based approaches, where the constraint of the dual Kantorovich problem is replaced with a penalty term. Instead, we represent the dual functions with ICNN, so that the constraint is automatically satisfied. Further, the transport map is expressed as gradient of a convex function, which is able to represent discontinuous maps. We believe that our framework paves way for bridging the optimal transport theory and practice.

References

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. *arXiv preprint arXiv:1609.07152*, 2016.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. *arXiv preprint arXiv:1710.06276*, 2017.

Chen, Y., Shi, Y., and Zhang, B. Optimal control via neural networks: A convex approach. *arXiv preprint arXiv:1805.11835*, 2018.

Chen, Y., Telgarsky, M., Zhang, C., Bailey, B., Hsu, D., and Peng, J. A gradual, semi-discrete approach to generative network training via explicit wasserstein minimization. *arXiv preprint arXiv:1906.03471*, 2019.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

El Moselhy, T. A. and Marzouk, Y. M. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

Essid, M. and Solomon, J. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Evans, L. C. and Gangbo, W. *Differential equations methods for the Monge-Kantorovich mass transfer problem*, volume 653. American Mathematical Soc., 1999.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

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

Gopalan, R., Li, R., and Chellappa, R. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pp. 999–1006. IEEE, 2011.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Guo, Y., An, D., Qi, X., Luo, Z., Yau, S.-T., Gu, X., et al. Mode collapse and regularity of optimal transportation maps. *arXiv preprint arXiv:1902.02934*, 2019.

Hütter, J.-C. and Rigollet, P. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Jacob, L., She, J., Almahairi, A., Rajeswar, S., and Courville, A. W2gan: Recovering an optimal transport map with a gan. 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Korotin, A., Egiazarian, V., Asadulaev, A., and Burnaev, E. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Lei, N., Su, K., Cui, L., Yau, S.-T., and Gu, D. X. A geometric view of optimal transportation and generative model. *arXiv preprint arXiv:1710.05488*, 2017.

Leygonie, J., She, J., Almahairi, A., Rajeswar, S., and Courville, A. Adversarial computation of optimal transport maps. *arXiv preprint arXiv:1906.09691*, 2019.

Lin, Z., Khetan, A., Fanti, G., and Oh, S. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 1498–1507, 2018.

Liu, H., Xianfeng, G., and Samaras, D. A two-step computation of the exact gan wasserstein distance. In *International Conference on Machine Learning*, pp. 3165–3174, 2018.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie royale des sciences de Paris*, 1781.

Muzellec, B. and Cuturi, M. Subspace detours: Building transport plans that are optimal on subspace projections. *arXiv preprint arXiv:1905.10099*, 2019.

Papadakis, N., Peyré, G., and Oudet, E. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

Petzka, H., Fischer, A., and Lukovnicov, D. On the regularization of wasserstein gans. *arXiv preprint arXiv:1709.08894*, 2017.

Peyré, G., Cuturi, M., et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

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

Reich, S. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Su, Z., Wang, Y., Shi, R., Zeng, W., Sun, J., Luo, F., and Gu, X. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015.

Taghvaei, A. and Jalali, A. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *arXiv preprint arXiv:1902.07197*, 2019.

Villani, C. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Wei, X., Gong, B., Liu, Z., Lu, W., and Wang, L. Improving the improved training of wasserstein gans: A consistency term and its dual effect. *arXiv preprint arXiv:1803.01541*, 2018.

Xie, Y., Chen, M., Jiang, H., Zhao, T., and Zha, H. On scalable and efficient computation of large scale optimal transport. *arXiv preprint arXiv:1905.00158*, 2019.

Yang, L. and Karniadakis, G. E. Potential flow generator with l_2 optimal transport regularity for generative models. *arXiv preprint arXiv:1908.11462*, 2019.