Minimax Optimization without Spurious Solutions in Optimal Transport Learning

Yejun Kim* **Kyungjae Lee*** **Donghwan Kim**
Department of Mathematical Sciences
KAIST
Daejeon, Republic of Korea
{yejun1677,niyanda01,donghwankim}@kaist.ac.kr

Abstract

Using neural networks to learn optimal transport (OT) maps via minimax optimization has gained increasing interest in generative modeling. However, the widely used semi-dual OT maximin formulation is known to suffer from spurious global solutions that do not correspond to valid transport maps, often addressed by regularizing the objective function. We provide an alternative explanation by showing that these spurious solutions are not stationary points of the associated loss, implying that their appearance in practice stems from the optimizer's failure to converge to stationary points in nonconvex-nonconcave settings. However, when the c-concavity constraint is imposed on the potential, commonly to stabilize training, we show that these spurious solutions of semi-dual problem become stationary. Based on these observations, we propose solving the semi-dual OT problem without c-concavity constraint, using optimizers that can reliably converge to stationary solutions corresponding to OT maps in nonconvex-nonconcave settings, such as a two-timescale extragradient (TTS-EG) method. At the formulation level, we further show that a Lagrangian-based OT minimax formulation provably eliminates spurious global solutions. Our empirical results demonstrate that TTS-EG reliably recovers OT maps under both semi-dual and Lagrangian-based OT formulations without any regularization.

1 Introduction

Optimal transport (OT) theory [Monge, 1781] provides a rigorous mathematical framework for determining a cost-minimal transport map that transforms a source distribution μ into a target distribution ν. Learning OT map has attracted growing interest across a wide range of machine learning tasks, such as generative modeling [Rout et al., 2022, Fan et al., 2022, Korotin et al., 2023], reinforcement learning [Fickinger et al., 2022, Luo et al., 2023], domain adaptation [Ganin and Lempitsky, 2015, Courty et al., 2017, Damodaran et al., 2018, Xie et al., 2019], barycenter estimation [Korotin et al., 2022, Kolesov et al., 2024], and biology [Bunne et al., 2023].

A prevalent approach to learning OT maps [Rout et al., 2022, Korotin et al., 2023] is leveraging neural networks through the semi-dual formulation of the Kantorovich problem [Villani, 2008], defined as:

$$\max_{V} \min_{T} \mathcal{L}(T, V) := \int_{\mathcal{X}} \Big\{ c(x, T(x)) - V(T(x)) \Big\} d\mu(x) + \int_{\mathcal{Y}} V(y) d\nu(y), \tag{1}$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ denotes the transportation cost function. Here, $T : \mathcal{X} \to \mathcal{Y}$ represents a transport map and $V : \mathcal{Y} \to \mathbb{R}$ represent a Kantorovich potential, which are both parameterized by

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

neural networks. Throughout this paper, we refer to this maximin formulation as the *Semi-Dual OT* problem.

If both the OT map $T^\star$ and optimal potential $V^\star$ exist, its pair $(T^\star, V^\star)$ is a global (maximin) solution of the Semi-Dual OT problem [Rout et al., 2022, Fan et al., 2022, 2023], that is,

$$V^\star \in \arg\max_V \left\{ \Phi(V) := \min_{T'} \mathcal{L}(T', V) \right\}, \qquad T^\star \in \arg\min_T \mathcal{L}(T, V^\star). \tag{2}$$

However, the converse does not necessarily hold: a global solution $(\tilde{T}, \tilde{V})$ does not guarantee that $\tilde{T}$ is an OT map. Recently, Choi et al. [2025] showed that such *spurious* global solutions (that do not correspond to the OT map) exist when the source distribution μ is not absolutely continuous with respect to the Lebesgue measure.

To address this issue, Korotin et al. [2023] and Choi et al. [2025] proposed regularizing the transport cost c and smoothing the source distribution μ, respectively. However, these approaches either alter the original objective or require carefully tuned smoothing schedules. This limitation leads to our central research question:

When does minimax optimization recover the optimal transport map without requiring regularization of the objective?

We address this question by providing a theoretical characterization of *spurious* and *non-spurious* global solutions in the Semi-Dual OT problem. Building on the observation that a global solution is not necessarily a stationary point [Jin et al., 2020], we prove that **a global solution** (2) **is non-spurious (*i.e.,* corresponds to the OT map) if and only if it is a stationary point**. This result implies that gradient-based methods that converge to stationary points can naturally bypass spurious global solutions. Moreover, this suggests that their appearance in practice stems from the optimizer's failure to converge to stationary points in nonconvex-nonconcave settings, rather than the mere existence of spurious solutions.

However, when the c-concavity constraint is imposed on the potential, we show that the spurious solutions become stationary. Enforcing c-concavity constraint (*e.g.,* via Input Convex Neural Networks [Amos et al., 2017]) has been found to improve the loss landscape and stabilize training [Makkuva et al., 2020, Taghvaei and Jalali, 2019]. These observations suggest that the c-concavity constraint can stabilize the training dynamics but make spurious solutions stationary, and thereby cause gradient-based methods to converge to undesirable solutions.

As another approach to learning OT, we show that the Lagrangian-based minimax formulation—obtained by simply interchanging the order of $\max$ and $\min$ in (1)—provably eliminates spurious global solutions at the formulation level. This formulation has not been used in practice, because the inner maximization is unbounded, making it seemingly intractable. However, it is trainable via gradient-based methods that do not require an explicit inner maximization oracle.

Motivated by these observations, we propose solving both Semi-Dual and Lagrangian-based OT problem without c-concavity constraint, using the recently proposed two-timescale extragradient (TTS-EG) method [Chae et al., 2024, Kim and Kim, 2024]. It is shown that TTS-EG almost surely escapes non-optimal stationary points and thereby can possibly converge to stationary global solutions in nonconvex-nonconcave settings. Our experiments demonstrate that TTS-EG reliably converges to non-spurious global solutions for both the Semi-Dual and Lagrangian-based formulations.

Our main contributions are summarized as follows:

- In Section 3, we establish a precise characterization of *spurious* global solutions in the Semi-Dual optimal transport (OT) formulation. In particular, we show that a global solution is *non-spurious*, *i.e.,* corresponds to the OT map, if and only if it is a stationary point of the underlying maximin loss.
- In Section 4, we show that spurious solutions become stationary, when the c-concavity constraint is enforced on the potential. While this constraint is commonly adopted for training stability in Semi-Dual OT, our result suggests that it may instead induce spurious solutions.
- In Section 5, we show that a Lagrangian-based OT formulation provably eliminates spurious global solutions at the level of problem formulation.

- In Section 6, we empirically demonstrate that the two-timescale extragradient method can reliably recover OT maps for both Semi-Dual and Lagrangian-based OT problems.

Notations Let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$, and let $\mathcal{P}_2(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ denote the subset of measures with finite second moments. For any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, let $\Pi(\mu, \nu)$ denote the set of all joint probability distributions (couplings) on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. The product measure of μ and ξ is denoted by $\mu \times \xi$. For a μ-measurable map $T : \mathcal{X} \to \mathcal{Y}$, the pushforward measure of μ by T is denoted by $T_\sharp \mu$. Finally, δ_x denotes the Dirac measure at x.

2 Background

Optimal Transport. Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable transportation cost function. The classical optimal transport problem, as originally proposed by Monge [1781], seeks a transport map $T : \mathcal{X} \to \mathcal{Y}$ that minimizes the total cost. For two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the Monge problem is defined as:

$$\inf_{T_\sharp \mu = \nu} \int_{\mathcal{X}} c(x, T(x))\, d\mu(x), \tag{3}$$

where the infimum is taken over all measurable maps T satisfying the pushforward constraint $T_\sharp \mu = \nu$. A minimizer $T^\star$, if it exists, is referred to as an *optimal transport (OT) map*. However, such deterministic OT maps may not exist when mass splitting is required, e.g., when transporting a discrete measure μ to a continuous measure ν.

To address the limitation, Kantorovich [2006] introduced a relaxed formulation that replaces the deterministic map with a coupling $\pi \in \Pi(\mu, \nu)$, which is a joint distribution with marginals μ and ν:

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\; d\pi(x, y). \tag{4}$$

Under mild conditions, a minimizer $\pi^\star$ is guaranteed to exist [Villani, 2008]; such a solution is referred to as an *optimal transport plan* between μ and ν.

Semi-Dual Problem of Optimal Transport. For a general cost function $c(\cdot, \cdot)$ that is lower semi-continuous and bounded below, the Kantorovich problem (4) has the following semi-dual form [Villani, 2008]:

$$\sup_{V \in \mathcal{V}} \left\{ \int_{\mathcal{X}} V^c(x)\, d\mu(x) + \int_{\mathcal{Y}} V(y)\, d\nu(y) \right\}, \tag{5}$$

where V^c denotes the *c-transform* of V, defined as $V^c(x) := \inf_{y \in \mathcal{Y}} c(x, y) - V(y)$. A maximizer $V^\star$ of (5) is referred to as a *Kantorovich optimal potential*.

Recent works [Rout et al., 2022, Korotin et al., 2023] learn OT maps by parameterizing the c-transform as an infimum over maps T. In the deterministic case, the first term in (5) is expressed using measurable maps $T : \mathcal{X} \to \mathcal{Y}$ [Rockafellar, 2006, Theorem 3A]:

$$\int V^c(x)\, d\mu(x) = \inf_{T:\mathcal{X}\to\mathcal{Y}} \int [c(x, T(x)) - V(T(x))] d\mu(x). \tag{6}$$

Substituting this into (5) yields the Semi-Dual objective defined in (1).

Weak Formulation of Monge's Problem. Korotin et al. [2023] reformulated the Kantorovich problem (4) by parameterizing the conditional distribution as $\pi(\cdot|x) = T(x, \cdot)_\sharp \xi$, where $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ is a stochastic map and $\xi \in \mathcal{P}(\mathcal{Z})$ is a noise distribution. This defines the joint measure as $\pi = (\mathrm{Id}_\mathcal{X}, T)_\sharp(\mu \times \xi)$, where $(\mathrm{Id}_\mathcal{X}, T)(x, z) = (x, T(x, z))$, leading to the following objective:

$$\min_{T_\sharp(\mu\times\xi)=\nu} \int_{\mathcal{X}} \int_{\mathcal{Z}} c(x, T(x, z))\, d\xi(z) d\mu(x). \tag{7}$$

When ξ is atomless[2], $(\mathrm{Id}_\mathcal{X}, T)_\sharp(\mu \times \xi)$ can represent any coupling in $\Pi(\mu, \nu)$ [Korotin et al., 2023]. The problem (7) is equivalent to the weak formulation of Monge problem [Gozlan et al., 2017].

[2]A probability measure ξ is atomless if $\xi(\{z\}) = 0$ for every $z \in \mathcal{Z}$; this condition is satisfied by standard continuous distributions, such as Gaussian measures.

Korotin et al. [2023] showed that the c-transform can also be expressed by a stochastic map:

$$\int_{\mathcal{X}} V^c(x)\, d\mu(x) = \inf_{T:\mathcal{X}\times\mathcal{Z}\to\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Z}} \left[c(x, T(x,z)) - V(T(x,z))\right]\, d(\mu\times\xi)(x,z). \tag{8}$$

Substituting (8) into the dual problem (5) yields the weak version of the Semi-Dual OT formulation. Since the weak OT problem reduces to the standard OT problem when the noise ξ is a Dirac measure (such that $T(x,z) \equiv T(x)$), we focus our subsequent analysis primarily on the weak OT framework.

Gradient-based Methods for Nonconvex-Nonconcave Minimax Optimization. A widely used method for the minimax problem is Gradient Descent Max-oracle (GDmax). Applied to our maximin formulation, it performs (sub)gradient ascent on the outer maximization $\max_V\{\Phi(V) := \min_{T'} \mathcal{L}(T', V)\}$. Its update rule is given by:

$$V_{t+1} = V_t + \gamma s_t, \quad \text{where } s_t \in \partial\Phi(V_t). \tag{9}$$

where γ denotes the step size. The subgradient s_t is computed via a minimization oracle on $\mathcal{L}(T, V_t)$ with respect to T, which in practice is approximated by a finite number of descent steps. Another common approach is Gradient Descent Ascent (GDA), which updates both variables simultaneously. Despite their popularity, both GDmax and GDA are known to show limit cycle behavior even in simple bilinear cases [Mescheder et al., 2018]. Extragradient (EG) methods [Korpelevich, 1976] address this issue by incorporating an additional predictive step. While this improves convergence behavior, it still does not guarantee convergence to solutions in general nonconvex-nonconcave problems.

Inspired by the success of gradient descent in non-convex minimization—often attributed to its ability to almost surely escape strict saddle points [Lee et al., 2016]—Chae et al. [2024] introduced the notion of *strict non-minimax* points as a minimax analogue of strict saddle points. They then proposed a two-timescale variant of the EG method defined by the following update rules:

$$\begin{aligned} \bar{V}_t &= V_t + \gamma_V \nabla_V \mathcal{L}(T_t, V_t), \qquad \bar{T}_t = T_t - \gamma_T \nabla_T \mathcal{L}(T_t, V_t) \\ V_{t+1} &= V_t + \gamma_V \nabla_V \mathcal{L}(\bar{T}_t, \bar{V}_t), \quad T_{t+1} = T_t - \gamma_T \nabla_T \mathcal{L}(\bar{T}_t, \bar{V}_t) \end{aligned} \tag{10}$$

where γ_V and γ_T denote the step sizes for variables V and T, respectively. Under the nonconvex-nonconcave setting, this two-timescale EG (TTS-EG) method can almost surely escape strict non-minimax points while remaining stable at global solutions, provided that a sufficiently smaller step size is assigned to the outer variable ($\gamma_V < \gamma_T$ in our case) [Chae et al., 2024, Kim and Kim, 2024]. In contrast, two-timescale GDA and other existing gradient-based methods do not enjoy this desirable property.

3 Spurious Global Solutions of Semi-Dual OT are Not Stationary

This section briefly reviews why spurious global solutions exist in the Semi-Dual OT formulation, and shows that they are not stationary points of the associated maximin objective.

3.1 Motivating Example : Perpendicular Distribution

The semi-dual problem (5) is designed to recover a Kantorovich optimal potential. Its relaxed formulation (1), introduced by Rout et al. [2022], Fan et al. [2022], augments the problem with an additional map parameter T and attains the same optimal value $\text{OT}_c(\mu,\nu)$ with the same Kantorovich optimal potential. However, this equivalence in optimal value does not guarantee that the optimal solution corresponds to an OT map. We call $(\tilde{T}, \tilde{V})$ a *spurious global solution* of Semi-Dual OT formulation if it satisfies the global optimality condition (2) but $\tilde{T}$ is not an OT map. Such spurious solutions can exist when the source distribution μ is not absolutely continuous with respect to Lebesgue measure [Choi et al., 2025]. To illustrate this issue, we revisit an example from Choi et al. [2025, Example 1] within the weak OT framework.

Example 3.1. Let $c(x,y) = \frac{1}{2}\|x-y\|^2$ and uniform measures μ, ν supported on perpendicular segments:

$$\mu = \mathcal{U}([-1,1]\times\{0\}), \quad \nu = \mathcal{U}(\{0\}\times[-1,1]). \tag{11}$$

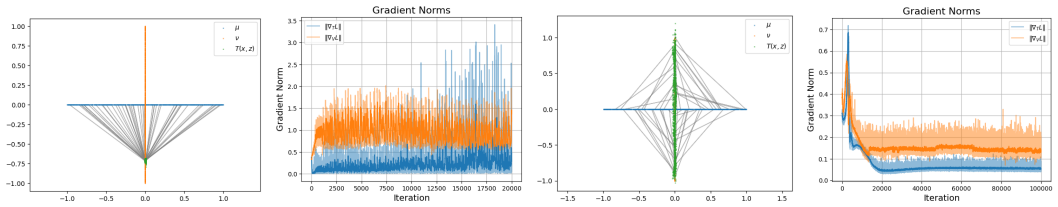


Figure 1: Comparison of GDmax (left) and EG (right) on Example 3.1 training stochastic neural network. GDmax does not converge, whereas EG converges to the non-spurious global solution.

In this case, any map T satisfying the pushforward constraint $T_\sharp(\mu \times \xi) = \nu$ is an OT map with $\text{OT}_c(\mu, \nu) = \frac{1}{3}$, and $V^\star(y_1, y_2) = -|y_1| + \frac{1}{2}y_2^2$ is an optimal potential. We will show that the pair $(\tilde{T}_0, V^\star)$ is a global solution where $\tilde{T}_0 \equiv 0$.

From optimality of $V^\star$, we get $V^\star \in \arg\max_V \{\Phi(V) := \min_{T'} \mathcal{L}(T', V)\}$. Therefore, it suffices to show that $\tilde{T}_0 \in \arg\min_T \mathcal{L}(T, V^\star)$. This is equivalent to showing that $w = (0, 0)$ minimizes the objective $c(x, w) - V^\star(w)$ for any $x = (x_1, 0) \in \text{spt}(\mu)$. It follows that:

$$c(x, w) - V^\star(w) = \tfrac{1}{2}(w_1 - x_1)^2 + \tfrac{1}{2}w_2^2 - \left(\tfrac{1}{2}w_2^2 - |w_1|\right) = \tfrac{1}{2}(w_1 - x_1)^2 + |w_1|. \tag{12}$$

Since $|x_1| \leq 1$, the expression above is indeed minimized at $w_1 = 0$ (and w_2 can be any value, including 0, since it cancels out). Consequently, $(\tilde{T}_0, V^\star)$ is a global solution to the Semi-Dual OT problem. However, because $(\tilde{T}_0)_\sharp(\mu \times \xi) \neq \nu$, this pair constitutes a spurious global solution. In fact, any measurable map of the form $\tilde{T}(x_1, 0, z) = (0, t(x_1, z))$ is a global solution to the Semi-Dual objective. See Appendix A for detail.

A crucial observation comes from the fact that a global solution to a minimax problem need not be a stationary point [Jin et al., 2020]. This applies to the spurious global solution $(\tilde{T}_0, V^\star)$ in Example 3.1, where the directional derivative along $\phi = 1_{\{y=\mathbf{0}\}}$ remains non-vanishing:

$$\begin{aligned}\frac{d}{dt}\mathcal{L}(\tilde{T}_0, V^\star + t\phi)\Big|_{t=0} &= \int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}\times\mathcal{Z}} \phi(\tilde{T}_0(x, z))\, d(\mu \times \xi)(x, z) \\ &= \nu(\{\mathbf{0}\}) - (\mu \times \xi)(\{(x, z) : \tilde{T}_0(x, z) = \mathbf{0}\}) = -1.\end{aligned}$$

Thus, such a *spurious* global solution $(\tilde{T}_0, V^\star)$ is not a stationary point of $\mathcal{L}$.

3.2 Spurious Global Solutions of Semi-Dual OT are Not Stationary

We now extend the previous observations to establish a general theoretical characterization of *spurious* and *non-spurious* global solutions in the Semi-Dual OT problem. Throughout the remainder of our paper, we assume $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and the cost function is the quadratic cost $c(x, y) = \alpha\|x - y\|^2$ and the measures μ, ν belong to $\mathcal{P}_2(\mathbb{R}^d)$.[3]

To discuss differentiability and stationarity, the domain of the loss $\mathcal{L}(T, V)$ must be defined on appropriate normed vector spaces. For transport maps, we let T to be in the Hilbert space $\mathcal{T} := L^2(\mu \times \xi; \mathbb{R}^d)$, which ensures that the pushforward measure $T_\sharp(\mu \times \xi)$ lies in $\mathcal{P}_2(\mathbb{R}^d)$. For the potential, we consider

$$\mathcal{V} := \{V \in C(\mathbb{R}^d) : |V(y)| \leq M(1 + \|y\|^2) \text{ for some } M > 0 \text{ and all } y \in \mathcal{Y}\} \tag{13}$$

from Gozlan et al. [2017], Backhoff-Veraguas et al. [2019], where $C(\mathbb{R}^d)$ denotes the space of all continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. The space $\mathcal{V}$ is a Banach space equipped with the norm $\|V\|_\mathcal{V} = \sup_y \frac{|V(y)|}{1+\|y\|^2}$, and ensures that $\max_{V\in\mathcal{V}} \min_{T\in\mathcal{T}} \mathcal{L}(T, V) = \text{OT}_c(\mu, \nu)$. Importantly, this space is sufficiently rich to approximate any continuous function arbitrarily well on compact subsets on $\mathbb{R}^d$. Therefore, restricting the potential V to the space $\mathcal{V}$ does not impose a meaningful constraint in practice, and does not act as a regularizer. Rather, this choice is made solely to ensure a well-defined functional setting for analyzing differentiability and stationarity.

[3]The analysis extends to nonnegative continuous costs c for which there exist $\alpha > 0$ and $r \geq 1$ such that $c(x, y) \leq \alpha(\|x\|^r + \|y\|^r)$ for all $x, y \in \mathbb{R}^d$, see Appendix B.

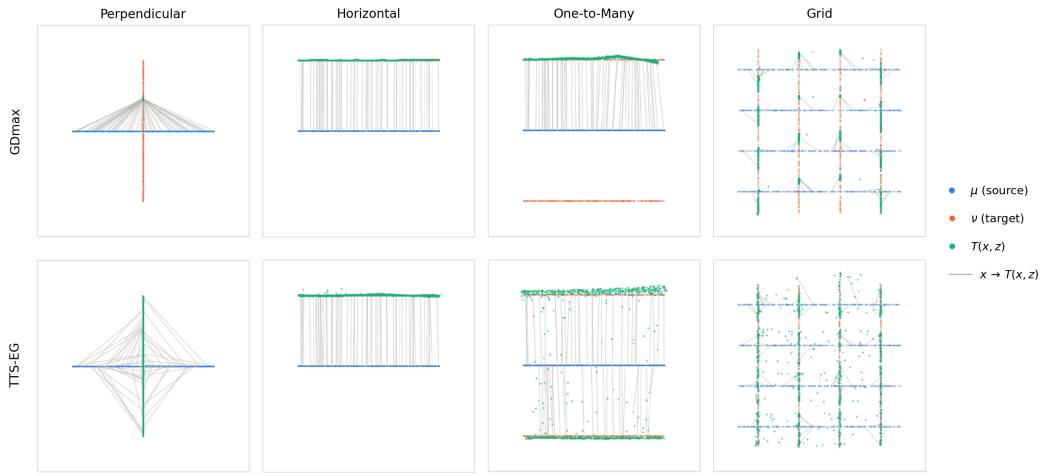


Figure 2: OT map learned by GDmax and TTS-EG in settings where spurious global solutions are known to exist [Choi et al., 2025]. TTS-EG outperforms GDmax except in the Horizontal setting, where both methods recover the OT map.

This choice, however, does not guarantee that an optimal solution exists in the domain $\mathcal{T} \times \mathcal{V}$, and we therefore impose the following assumption.

Assumption 3.2. There exists an optimal transport $T^\star \in \mathcal{T}$ and a Kantorovich optimal potential $V^\star \in \mathcal{V}$.

If ξ is atomless, for example, an optimal transport $T^\star$ is guaranteed to exist [Korotin et al., 2023, Lemma 1]. In contrast, a Kantorovich potential $V^\star$ does not necessarily belong to $\mathcal{V}$ in general. Nevertheless, Assumption 3.2 holds in many practical scenarios—for instance, when $\mathrm{spt}(\nu)$ is bounded. See Appendix B.1 for detailed explanations.

We now present the main theorem of this paper.

Theorem 3.3. *Let $c(x, y) = \alpha\|x - y\|^2$ be the quadratic cost, and let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\xi \in \mathcal{P}(\mathcal{Z})$ be the given measures. Suppose Assumption 3.2 holds. Then, a global maximin point $(T^\star, V^\star) \in \mathcal{T} \times \mathcal{V}$ of the Semi-Dual loss is non-spurious if and only if the Fréchet derivative with respect to V vanishes: $\nabla_V \mathcal{L}(T^\star, V^\star) = 0$ (see Definition B.1).*

Sketch of proof. Under Assumption 3.2, for any fixed $(T, V) \in \mathcal{T} \times \mathcal{V}$, the Fréchet derivative $\nabla_V \mathcal{L}(T, V)$ is a linear functional on $\mathcal{V}$ whose action on a test function $\phi \in \mathcal{V}$ is given by:

$$\langle \nabla_V \mathcal{L}(T, V), \phi \rangle = \int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}\times\mathcal{Z}} \phi(T(x, z))\, d(\mu \times \xi)(x, z). \tag{14}$$

This identity follows from the linearity of $\mathcal{L}(T, \cdot)$ with respect to V.

($\implies$) First, assume that $(T^\star, V^\star)$ is non-spurious. Then $T^\star$ is an OT map. By definition, this means the pushforward constraint is satisfied, i.e., $T^\star_\sharp(\mu \times \xi) = \nu$. Substituting this into (14), the two integrals cancel out for any measurable function $\phi \in \mathcal{V}$. Thus, $\langle \nabla_V \mathcal{L}(T^\star, V^\star), \phi \rangle = 0$ for all ϕ, which implies $\nabla_V \mathcal{L}(T^\star, V^\star) = 0$.

($\impliedby$) Conversely, assume the stationarity condition $\nabla_V \mathcal{L}(T^\star, V^\star) = 0$ holds. This implies that for all test functions $\phi \in \mathcal{V}$:

$$\int_{\mathcal{Y}} \phi(y)\, d\nu(y) = \int_{\mathcal{X}\times\mathcal{Z}} \phi(T^\star(x, z))\, d(\mu \times \xi)(x, z) = \int_{\mathcal{Y}} \phi(y)\, d(T^\star_\sharp(\mu \times \xi))(y). \tag{15}$$

Since $\mathcal{V}$ is rich enough to distinguish any two probability measures with finite second moments, the identity $\int \phi d\nu = \int \phi d(T^\star_\sharp(\mu \times \xi))$ for all $\phi \in \mathcal{V}$ yields $\nu = T^\star_\sharp(\mu \times \xi)$. Therefore, $(T^\star, V^\star)$ satisfies the pushforward constraint and is non-spurious. □

3.3 Convergence to Stationary Global Solutions: From GDmax to TTS-EG

Theorem 3.3 establishes that spurious global solutions in the Semi-Dual OT formulation are inherently non-stationary. This result yields a critical insight: the primary challenge in optimizing Semi-Dual OT is not the mere existence of spurious solutions themselves, but rather the use of algorithms that fail to converge to stationary points.

In this context, we compare the empirical behavior of GDmax and EG in Example 3.1. As illustrated in Figure 1, GDmax oscillates near the spurious solutions, while EG converges to a stationary global solution. This shows that GDmax is inappropriate for training Semi-Dual OT, particularly when spurious solutions exist.

More generally, however, nonconvex–nonconcave problems may also contain undesirable stationary points, including strict non-minimax points, which EG may fail to avoid. Hence, we use TTS-EG, which can avoid strict non-minimax points, as discussed in Section 2. Figure 2 shows stochastic OT maps recovered by TTS-EG in settings where spurious global solutions are known to exist [Choi et al., 2025]. In Section 6, we further evaluate TTS-EG by comparing it with GDmax and EG.

4 Spurious Global Solution Can be Stationary under c-concave Constraint

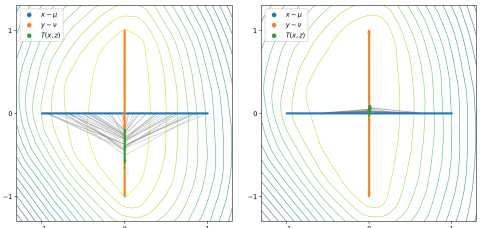
In the preceding section, we established that the OT map can be recovered through convergent optimizer. In this section, we examine that the c-concavity constraint, which is widely adopted in Semi-Dual OT for improving training stability. We show that this constraint makes spurious solutions stationary, causing gradient-based methods to converge to spurious global solutions.

4.1 Motivation for Using the c-concavity Constraint in Prior Works

In semi-dual Kantorovich problem (5), restricting the potential V to the space of c-concave functions[4] does not alter the optimal value [Santambrogio, 2015, Prop 1.11]. This allows us to reformulate the problem into the following maximin optimization:

$$\max_{V \in \mathcal{V} \cap \mathcal{S}_c} \min_{T \in \mathcal{T}} \mathcal{L}(T, V). \tag{16}$$

where $\mathcal{S}_c$ denotes a collection of c-concave functions. For the quadratic cost, the space of c-concave functions is characterized as follows:



(a) GDmax (b) TTS-EG

Figure 3: Learned map under $\mathcal{S}_c$ constraint. The potentials are represented by level curves.

$$\mathcal{S}_c = \left\{ V(y) = \alpha \|y\|^2 - u(y) \;\middle|\; u \text{ is a lower semicontinuous convex function} \right\}. \tag{17}$$

In practice, u is parameterized using Input Convex Neural Networks [Amos et al., 2017].

This form is known to alleviate the non-convexity of the inner minimization of Semi-Dual OT formulation, thereby improving the overall loss landscape [Makkuva et al., 2020, Taghvaei and Jalali, 2019]. We clarify this effect more formally by examining the problem within the functional space.

Proposition 4.1. *If $c(x, y) = \alpha \|x - y\|^2$, then the functional $\mathcal{L} : \mathcal{T} \times (\mathcal{V} \cap \mathcal{S}_c) \to \mathbb{R}$ is convex-concave. Moreover, the problem* (16) *is a constrained convex-concave maximin optimization.*

Here, the term "constrained" is used because the set $\mathcal{S}_c$ is not a vector space.

4.2 c-concavity Constraint Makes Spurious Global Solutions Stationary

Despite these seemingly desirable properties, we show that this constraint introduces a significant theoretical drawback: it can make spurious solutions stationary. In such constrained setting, a point $V^\star$ is stationary of $\mathcal{L}(T^\star, V)$ if it satisfies the following variational inequality for all $V \in \mathcal{V} \cap \mathcal{S}_c$:

$$\langle \nabla_V \mathcal{L}(T^\star, V^\star), V - V^\star \rangle \leq 0. \tag{18}$$

[4] A function $V : \mathcal{Y} \to \mathbb{R}$ is said to be c-concave if there exists a function $\phi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ such that $V(y) = \inf_{x \in \mathcal{X}} \{c(x, y) - \phi(x)\}$ for all $y \in \mathcal{Y}$.

Theorem 4.2. *Under the same assumptions as in Theorem 3.3 with the quadratic cost c, if there is a non-deterministic OT map $T^\star$, i.e., $\mu(\{x \mid T^\star(x,\cdot)_\sharp \xi$ is a non-Dirac measure almost surely$\}) > 0$, then the pair $(\tilde{T}, V^\star)$, where $\tilde{T}(x,z) := \mathbb{E}_{z' \sim \xi}[T^\star(x,z')]$ is an averaged map and $V^\star$ is an optimal potential, is a stationary spurious global solution of the constrained problem* (16)*.*

In Figure 3, both GDmax and TTS-EG achieve convergence under the c-concavity constraint; however, they fail to converge to the true OT map. Instead, they converge to spurious global solutions as the perpendicular distributions in Example 3.1 satisfy the conditions of Theorem 4.2, which admits an optimal stochastic map $T^\star(x,z) = (0,z)$ with $z \sim \mathcal{U}([-1,1])$.

Consequently, while c-concavity constraint may induce a better loss landscape and facilitate the convergence of gradient-based methods, they are inappropriate for recovering the true OT map.

5 Lagrangian-Based OT Formulation Does Not Have Spurious Solutions

While previous sections focused on identifying stationary global solutions through convergent optimizer, we now revisit an alternative Lagrangian-based OT formulation [Fan et al., 2022] and show that it eliminates spurious global solutions at the level of problem formulation.

We begin with revisiting the (weak) Monge problem (7), which is defined as a constrained minimization subject to $\nu - T_\sharp(\mu \times \xi) = 0$. A standard approach with constraints is the method of Lagrange multipliers, leading to the following term:

$$\int_{\mathcal{Y}} V(y)\,(\nu - T_\sharp(\mu \times \xi))(dy) = \int_{\mathcal{Y}} V(y)\,d\nu(y) - \int_{\mathcal{X}\times\mathcal{Z}} V(T(x,z))\,d\mu(x)d\xi(z), \tag{19}$$

where the V acts as the Lagrange multiplier. By incorporating this term, the constrained Monge problem is transformed into an unconstrained minimax problem:

$$\min_T \max_V \; \mathcal{L}(T,V), \tag{20}$$

which we refer to as the *Lagrangian-based OT problem*. While the objective function in (20) is identical to that of the Semi-Dual OT problem, the change of the order of the $\min$ and $\max$ operators fundamentally reshapes the structure of global solutions.

Proposition 5.1. *Under the same assumptions as in Theorem 3.3, the set of global solutions of Lagrangian-based OT* (20) *is characterized by $\{(T^\star, V) \in \mathcal{T} \times \mathcal{V} \mid T^\star$ is a (stochastic) OT map $\}$.*

This reveals a structural symmetry between the two formulations. While the Semi-Dual OT formulation identifies the optimal potential $V^\star$ but leaves the map T under-determined (yielding spurious solutions), the Lagrangian-based formulation identifies the exact OT map $T^\star$ while leaving the potential V under-determined.

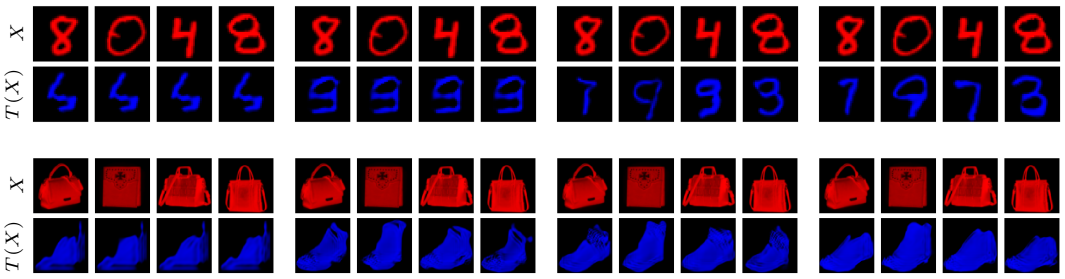
We would like to note that learning an OT map through this Lagrangian-based formulation has not been explored previously, since the objective $\max_{V'} \mathcal{L}(T,V')$ takes an infinite value unless the push-forward constraint is satisfied, making GDmax useless. In contrast, TTS-EG does not have this issue, as it requires no inner maximization. We also note that this Lagrangian-based formulation is closely related to the Zero-Infinity Generative Adversarial Networks (GAN) [Lee and Kim, 2025]; we discuss this connection and its relation to other GAN objectives in Appendix E.

In the next section, we empirically demonstrate that TTS-EG can reliably solve the Lagrangian-based formulation, as in the Semi-Dual formulation.

6 Experiments: Two-Timescale Extragradient Converges to OT Map

In this section, we empirically demonstrate that the TTS-EG method reliably recovers the OT map for both Semi-Dual and Lagrangian-based formulations without c-concavity constraint. These two formulations share the same objective function and only differ in the order of $\max$ and $\min$. Hence, when applying TTS-EG, only the step size ordering changes: $\gamma_T > \gamma_V$ for Semi-Dual OT and $\gamma_T < \gamma_V$ for Lagrangian OT. Detailed update rules are provided in Appendix F.

We evaluate the convergence of the optimizer using two metrics: cost error $D_{\text{cost}} = |\text{OT}_c(\mu,\nu) - \mathbb{E}[c(x,T(x,z))]|$, which measures the deviation of the learned transport cost from



(a) GDmax (b) Extragradient (c) TTS-EG (Semi-Dual) (d) TTS-EG (Lagrangian)

Figure 4: Comparison of image translation across different optimization methods. **Top:** Even to Odd digit translation. **Bottom:** Handbags to shoes translation. GDmax and standard EG often collapse to a single output.

the optimal value, and target error $D_{\text{target}} = \text{OT}_c(T_\sharp(\mu \times \xi), \nu)$, which quantifies how well the push-forward distribution matches the target distribution. Recovery of the OT map requires both errors to be small.

We consider stochastic OT-map learning on all four synthetic settings from Choi et al. [2025] where spurious global solutions are known to exist in Semi-Dual OT. We compare GDmax, EG and TTS-EG against GDmax with smoothing, which injects Gaussian noise into the source measure μ to eliminate spurious solutions and gradually reduces the noise to recover the OT plan. As shown in Table 1, TTS-EG achieves comparable or lower errors in both D_{cost} and D_{target}, without relying on careful noise annealing. Results for all four settings are provided in Appendix G.

Image Translation Tasks. Moving beyond toy examples, we evaluate the optimization methods on deterministic OT-map learning in settings where spurious solutions exist. We consider two image translation tasks: digit translation on the MNIST dataset and *handbags-to-shoes* translation. Due to the high complexity of these image translation tasks, we use Adam in all methods.

First, we examine the mapping from red even digits to blue odd digits, where the supports of two distributions are almost orthogonal. Similar to Example 3.1, such a setting gives rise to spurious solutions. In this setting, both GDmax and EG fail to capture the target distribution and instead collapse to a single digit, while TTS-EG methods converge to the correct solutions in both Semi-Dual and Lagrangian-based formulations.

Second, we evaluate a higher-dimensional task: translating red handbags to blue shoes using $3 \times 64 \times 64$ images. We use Amazon Handbags [Zhu et al., 2016] and UT Zappos50K [Yu and Grauman, 2014], which were previously used in neural OT experiments [Korotin et al., 2023]. Consistent with the MNIST results, TTS-EG performs favorably in both the semi-dual and Lagrangian formulations compared with GDmax and EG. Figure 4 and Table 2 present qualitative and quantitative results, respectively.

Method	$D_{\text{cost}} \downarrow$	$D_{\text{target}} \downarrow$
Data 1: Perpendicular		
GDmax	0.1303	0.5634
GDmax + Smoothing	0.1031	0.0114
EG	0.0118	0.0023
TTS-EG (Semi-Dual)	0.0025	0.0003
TTS-EG (Lagrangian)	**0.0018**	**0.0005**
Data 2: One-to-Many		
GDmax	0.6126	1.4156
GDmax + Smoothing	0.0062	0.0541
EG	0.0144	0.0196
TTS-EG (Semi-Dual)	0.0221	0.0332
TTS-EG (Lagrangian)	**0.0070**	**0.0218**

Table 1: Quantitative comparison on 2D synthetic tasks. Values are means over five independent runs. Both metrics are highlighted in bold for the method with the lowest $D_{\text{cost}} + D_{\text{target}}$ in each dataset.

Method	$D_{\text{cost}} \downarrow$	$D_{\text{target}} \downarrow$
Data 1: Red even to Blue odd		
GDmax	0.0084	0.1269
GDmax + Smoothing	0.0019	0.0578
EG	0.0366	0.1249
TTS-EG (Semi-Dual)	0.0045	0.0585
TTS-EG (Lagrangian)	**0.0019**	**0.0400**
Data 2: Red Handbags to Blue Shoes		
GDmax	0.0087	0.1209
GDmax + Smoothing	0.0795	0.0680
EG	0.0009	0.0607
TTS-EG (Semi-Dual)	0.0065	0.0533
TTS-EG (Lagrangian)	**0.0067**	**0.0495**

Table 2: Quantitative comparison on image translation tasks. Both metrics are highlighted in bold for the method with the lowest $D_{\text{cost}} + D_{\text{target}}$ in each dataset.

7 Conclusion

In this paper, we showed that spurious global solutions of Semi-Dual OT problem are not stationary points unless a c-concavity constraint is imposed on the potentials. Moreover, we established that the Lagrangian-based OT formulation does not have spurious solutions. Based on these observations, we propose solving both the Semi-Dual and Lagrangian-based OT problems using the Two-timescale Extragradient (TTS-EG) method, which can reliably converge to stationary solutions in nonconvex-nonconcave settings. Our empirical results demonstrate that TTS-EG reliably identifies optimal transport maps in both the Semi-Dual and Lagrangian-based frameworks.

Limitations and Future Work A primary limitation of our work is that the theoretical characterization of spurious global solutions as non-stationary points is derived within the function spaces. In practice, however, optimization is conducted in the neural network parameter spaces, which may contain numerous non-optimal stationary points. Extending our analysis to neural network parameterizations remains a promising avenue for future research. Furthermore, as our proposed framework facilitates the recovery of the exact OT map, establishing formal guarantees on the sample complexity and generalization of the learned maps will be an important next step.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (Nos. RS-2019-NR040050, RS-2025-00515159 and RS-2026-25613008), and by the KAIST Grand Challenge 30 Project (KC30) funded by MSIT and KAIST (No. N11260046).

References

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pages 146–155. PMLR, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Jiseok Chae, Kyuwon Kim, and Donghwan Kim. Two-timescale extragradient for finding local minimax points. In *ICLR*, 2024.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Analyzing and improving optimal-transport-based adversarial networks. In *ICLR*, 2024.

Jaemoo Choi, Jaewoong Choi, and Dohyun Kwon. Overcoming spurious solutions in semi-dual neural optimal transport: A smoothing approach for learning the optimal transport plan. In *Forty-second International Conference on Machine Learning*, 2025.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of Monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural Monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *International Conference on Learning Representations*, 2022.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Milena Gazdieva, Petr Mokrov, Litu Rout, Alexander Korotin, Andrey Kravchenko, Alexander Filippov, and Evgeny Burnaev. An optimal transport perspective on unpaired image super-resolution. *Journal of Optimization Theory and Applications*, 207(2):40, 2025.

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 (NeurIPS)*, 2014.

Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5967–5976. Ieee, 2017.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR, 2020.

LK Kantorovich. On a problem of Monge. *Journal of Mathematical Sciences*, 133(4), 2006.

Kyuwon Kim and Donghwan Kim. Double-step alternating extragradient with increasing timescale separation for finding local minimax points: Provable improvements. In *Forty-first International Conference on Machine Learning*, 2024.

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal transport. In *International Conference on Machine Learning*, pages 25016–25041. PMLR, 2024.

Alexander Korotin, Vage Egiazarian, Lingxiao Li, and Evgeny Burnaev. Wasserstein iterative networks for barycenter estimation. *Advances in Neural Information Processing Systems*, 35: 15672–15686, 2022.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

G. M. Korpelevich. An extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–56, 1976.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 2016.

Kyungjae Lee and Donghwan Kim. Zero-infinity GAN: Stable dynamics and implicit bias of extragradient. In *OPT 2025: Optimization for Machine Learning*, 2025.

Jacob Leygonie, Jennifer She, Amjad Almahairi, Sai Rajeswar, and Aaron Courville. Adversarial computation of optimal transport maps. *arXiv preprint arXiv:1906.09691*, 2019.

Shu Liu, Shaojun Ma, Yongxin Chen, Hongyuan Zha, and Haomin Zhou. Learning high dimensional wasserstein geodesics. *arXiv preprint arXiv:2102.02992*, 2021.

Guansong Lu, Zhiming Zhou, Jian Shen, Cheng Chen, Weinan Zhang, and Yong Yu. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

Yicheng Luo, Zhengyao Jiang, Samuel Cohen, Edward Grefenstette, and Marc Peter Deisenroth. Optimal transport for offline imitation learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

R Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators and the Calculus of Variations: Summer School Held in Bruxelles 8–19 September 1975*, pages 157–207. Springer, 2006.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

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

Villani. *Optimal transport: Old and new*, volume 338. Springer, 2008.

Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, and Hongyuan Zha. On scalable and efficient computation of large scale optimal transport. In *International Conference on Machine Learning*, pages 6882–6892. PMLR, 2019.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

A Additional Explanation for Toy Example in Section 3.1

Let $c(x, y) = \frac{1}{2}\|x - y\|^2$ and the source and target distributions be uniform measures supported on perpendicular line segments in $\mathbb{R}^2$:

$$\mu = \mathcal{U}([-1, 1] \times 0), \quad \nu = \mathcal{U}(0 \times [-1, 1]).$$

For any transport map T satisfying the pushforward constraint $T_\sharp(\mu \times \xi) = \nu$, the supports of μ and ν are orthogonal, implying $\langle x, T(x, z)\rangle = 0$ almost surely. Consequently, the transport cost is constant for all feasible maps:

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Z}} c(x, T(x, z))\mu(dx)\xi(dz) &= \frac{1}{2}\int_{\mathcal{X}\times\mathcal{Z}} \|x\|^2 d\mu(x) - \int_{\mathcal{X}} \langle x, T(x)\rangle d\mu(x) + \frac{1}{2}\int_{\mathcal{Y}} \|y\|^2 d\nu(y) \\ &= \frac{1}{2}\int_{-1}^{1} \frac{x_1^2}{2} dx_1 - 0 + \frac{1}{2}\int_{-1}^{1} \frac{y_2^2}{2} dy_2 \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.\end{aligned}$$

This implies that the optimal transport cost is $\mathrm{OT}_c(\mu, \nu) = \frac{1}{3}$.

Let $\tilde{T}(x_1, 0, z) = (0, t(x_1, z))$ with measurable function t, and let the potential be

$$V^\star(y_1, y_2) = \frac{1}{2}y_2^2 - |y_1|.$$

We now show that the pair $(\tilde{T}, V^\star)$ is a global maximin point of the Semi-Dual objective. First, we verify that $\tilde{T} \in \arg\min_T \mathcal{L}(T, V^\star)$. This condition is satisfied if:

$$\tilde{T}(x, z) \in \arg\min_w \left\{c(x, w) - V^\star(w)\right\}, \quad \mu\text{-a.e.}$$

For any $x = (x_1, 0) \in \mathrm{spt}(\mu)$, we examine the objective over $w = (w_1, w_2)$:

$$\begin{aligned}\min_{w_1,w_2} \left\{c(x, w) - V^\star(w)\right\} &= \min_{w_1,w_2} \left\{\frac{1}{2}\left((x_1 - w_1)^2 + w_2^2\right) - \left(\frac{1}{2}w_2^2 - |w_1|\right)\right\} \\ &= \min_{w_1,w_2} \left\{\frac{1}{2}(x_1 - w_1)^2 + |w_1|\right\}.\end{aligned}$$

The w_2^2 terms cancel out, meaning the objective is independent of w_2. For $x_1 \in [-1, 1]$, the function $f(w_1) = \frac{1}{2}(x_1 - w_1)^2 + |w_1|$ is minimized uniquely at $w_1 = 0$. Since w_2 is unconstrained in the minimization, any $\tilde{T}$ that maps to the support of ν (where $w_1 = 0$) is a valid minimizer. Finally, we evaluate the objective $\mathcal{L}$ at $(\tilde{T}, V^\star)$:

$$\begin{aligned}\mathcal{L}(\tilde{T}, V^\star) &= \int_{\mathcal{X}} \left[c(x, \tilde{T}(x, z)) - V^\star(\tilde{T}(x, z))\right] d\mu(x) + \int_{\mathcal{Y}} V^\star(y) d\nu(y) \\ &= \int_{-1}^{1} \frac{x_1^2}{4} dx_1 + \int_{-1}^{1} \frac{y_2^2}{4} dy_2 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.\end{aligned}$$

Since $\mathcal{L}(\tilde{T}, V^\star) = \mathrm{OT}_c(\mu, \nu)$, the pair $(\tilde{T}, V^\star)$ is a global maximin point.

B Proofs and Missing Details in Section 3

B.1 Theoretical justification of functional spaces choice

In this subsection, we provide a theoretical justification for our choice of functional spaces. Readers primarily interested in the main proofs may proceed directly to the next subsection.

We begin by recalling the Kantorovich formulation of optimal transport, which serves as a weak relaxation of the Monge problem. For two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the weak Monge problem is defined as

$$\inf_{T_\sharp(\mu\times\xi)=\nu} \int_{\mathcal{X}\times\mathcal{Z}} c(x, T(x, z))\, d(\mu \times \xi)(x, z).$$

We consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and a nonnegative lower-semicontinuous transportation cost satisfying

$$0 \leq c(x, y) \leq \alpha(\|x\|^r + \|y\|^r) \tag{21}$$

for some $\alpha > 0$ and $r \geq 1$. The quadratic cost $c(x, y) = \alpha\|x - y\|^2$ is a special case with $r = 2$. We assume $\mu, \nu \in \mathcal{P}_r(\mathbb{R}^d)$, which is sufficient to ensure finite transport cost.

For its dual problem,

$$\max_V \min_T \mathcal{L}(T, V) := \int_{\mathcal{X}\times\mathcal{Z}} \Big\{ c(x, T(x, z)) - V(T(x, z)) \Big\} d(\mu \times \xi) + \int_{\mathcal{Y}} V(y) d\nu(y),$$

the transport map T and the potential V are typically assumed to be measurable functions, and the infimum and supremum are taken over the space of all measurable functions.

However, the space of arbitrary measurable functions lacks a normed structure, making it unsuitable for defining the Fréchet derivative below.

Definition B.1. Let $\mathcal{E}$ be a normed vector space. A functional $\mathcal{L} : \mathcal{E} \to \mathbb{R}$ is called *Fréchet differentiable* at $u \in \mathcal{E}$ if there exists a bounded linear functional $A \in \mathcal{E}^*$ such that

$$\lim_{\|h\|_{\mathcal{E}} \to 0} \frac{|\mathcal{L}(u + h) - \mathcal{L}(u) - A(h)|}{\|h\|_{\mathcal{E}}} = 0,$$

where $\mathcal{E}^*$ denotes the continuous dual space of $\mathcal{E}$. The operator A is called the *Fréchet derivative* of $\mathcal{L}$ at u, and is denoted by $D\mathcal{L}(u)$ (or $\nabla\mathcal{L}(u)$).

To analyze the stationarity of the Semi-Dual loss $\mathcal{L}$, we must equip its domain with an appropriate norm. A natural choice for the transport map space is

$$\mathcal{T} = L^r(\mu \times \xi; \mathbb{R}^d),$$

as any valid map satisfying $T_\sharp(\mu \times \xi) = \nu$ belongs to this space. Defining a normed structure for the potential space requires careful consideration.

While $L^1(\nu)$ is a standard choice in classical optimal transport [Villani, 2008]—it is insufficient for defining the Fréchet derivative of the semi-dual loss. The primary issue is that elements of $L^1(\nu)$ are equivalence classes defined only up to ν-almost everywhere. Consequently, the functional $\mathcal{L}$ is ill-defined on $\mathcal{T} \times L^1(\nu)$ whenever $T_\sharp(\mu \times \xi) \not\ll \nu$.

For example, if $T_\sharp(\mu \times \xi) = \delta_0$ and $\nu = \delta_1$ and let $V_1(y) = 0$ and $V_2(y) = \mathbf{1}_{y=0}$. Since $V_1(1) = 0 = V_2(1)$, we have $V_1 = V_2$ ν-a.s. However, evaluating the semi-dual loss yields different results:

$$\mathcal{L}(T, V_1) - \mathcal{L}(T, V_2) = -V_1(0) + V_2(0) = 1.$$

An appropriate potential space $\mathcal{V}$ must satisfy the following criteria:

1. The functional $\mathcal{L}$ must be well-defined and Fréchet differentiable on $\mathcal{T} \times \mathcal{V}$.
2. The choice of space must not alter the global maximin value (Consistency), i.e.

$$\mathrm{OT}_c(\mu, \nu) = \max_{V \in L^1(\nu)} \min_{T \in \mathcal{T}} \mathcal{L}(T, V) = \max_{V \in \mathcal{V}} \min_{T \in \mathcal{T}} \mathcal{L}(T, V).$$

Based on these requirements, we adopt the potential space $\mathcal{V}$, equipped with the norm $\|V\|_{\mathcal{V}} := \sup_{y \in \mathbb{R}^d} \frac{|V(y)|}{1 + \|y\|^r}$, defined as:

$$\mathcal{V} := \{V \in C(\mathbb{R}^d) : \|V\|_{\mathcal{V}} < \infty\}.$$

In the remainder of this subsection, we establish the well-definedness of $\mathcal{L}$ in Proposition B.2 and prove the consistency of the minimax formulation $\max_{V \in \mathcal{V}} \min_{T \in \mathcal{T}} \mathcal{L}(T, V)$ in Proposition B.3.

Proposition B.2. *The objective functional $\mathcal{L} : \mathcal{T} \times \mathcal{V} \to \mathbb{R}$ is well-defined.*

Proof. The proof is almost direct. First, we show that $\mathcal{L}$ is well-defined with respect to the equivalence classes of the chosen norms. Suppose $\|T_1 - T_2\|_{\mathcal{T}} = 0$ and $\|V_1 - V_2\|_{\mathcal{V}} = 0$. The former implies $T_1 = T_2$ for $(\mu \times \xi)$-almost surely. For the latter, since $\mathcal{V} \subset C(\mathbb{R}^d)$, the condition

$\|V_1 - V_2\|_{\mathcal{V}} = 0$ directly implies $V_1(y) = V_2(y)$ for all $y \in \mathbb{R}^d$. These conditions collectively ensure that $\mathcal{L}(T_1, V_1) = \mathcal{L}(T_2, V_2)$.

Next, we verify that $\mathcal{L}(T, V)$ is finite for every $(T, V) \in \mathcal{T} \times \mathcal{V}$. The growth condition on c and the definition of $\|V\|_{\mathcal{V}}$ give

$$\begin{aligned}|\mathcal{L}(T,V)| &\le \int c(x,T(x,z))\,d(\mu\times\xi) + \int |V(y)|\,d\nu(y) + \int |V(T(x,z))|\,d(\mu\times\xi)\\ &\le \alpha\left(\int \|x\|^r\,d\mu + \int \|T(x,z)\|^r\,d(\mu\times\xi)\right)\\ &\quad + \|V\|_{\mathcal{V}}\left(2 + \int \|y\|^r\,d\nu + \int \|T(x,z)\|^r\,d(\mu\times\xi)\right).\end{aligned}$$

Since $\mu, \nu \in \mathcal{P}_r(\mathbb{R}^d)$ and $T \in L^r(\mu \times \xi)$, all integrals are finite. Thus $\mathcal{L}(T, V)$ is finite. □

Proposition B.3. *Suppose c satisfies* (21) *and is lower-semicontinuous, and $\mu, \nu \in \mathcal{P}_r(\mathbb{R}^d)$. Then*

$$\mathrm{OT}_c(\mu,\nu) = \sup_{V\in\mathcal{V}} \inf_{T\in\mathcal{T}} \mathcal{L}(T,V),$$

where $\mathrm{OT}_c(\mu, \nu)$ is the minimum value of the Kantorovich problem (4)*.*

Proof. For $V \in \mathcal{V}$, define its c-transform

$$V^c(x) := \inf_{y\in\mathbb{R}^d} \big(c(x,y) - V(y)\big).$$

For any fixed potential $V \in \mathcal{V}$, the interchange rule between integration and minimization [Rockafellar and Wets, 1998, Theorem 14.60], applied to the normal integrand $c(x, y) - V(y)$, gives

$$\begin{aligned}&\inf_{T\in\mathcal{T}} \int_{\mathcal{X}\times\mathcal{Z}} \big[c(x,T(x,z)) - V(T(x,z))\big]\,d(\mu\times\xi)\\ &\quad= \int_{\mathcal{X}\times\mathcal{Z}} \inf_{y\in\mathbb{R}^d} \big(c(x,y) - V(y)\big)\,d(\mu\times\xi)\\ &\quad= \int_{\mathcal{X}} V^c(x)\,d\mu(x),\end{aligned}$$

where the identity is understood in the extended-real sense if necessary. Consequently,

$$\inf_{T\in\mathcal{T}} \mathcal{L}(T,V) = \int V^c\,d\mu + \int V\,d\nu.$$

For every x, y,

$$V^c(x) + V(y) \le c(x,y).$$

Therefore, for every $\pi \in \Pi(\mu, \nu)$,

$$\int V^c\,d\mu + \int V\,d\nu \le \int c(x,y)\,d\pi(x,y).$$

Taking the infimum over π yields

$$\inf_{T\in\mathcal{T}} \mathcal{L}(T,V) \le \mathrm{OT}_c(\mu,\nu) \qquad \forall V \in \mathcal{V}.$$

Thus

$$\sup_{V\in\mathcal{V}} \inf_{T\in\mathcal{T}} \mathcal{L}(T,V) \le \mathrm{OT}_c(\mu,\nu).$$

For the reverse inequality, note that $\mathcal{V}$ contains $C_b(\mathbb{R}^d)$, the space of bounded continuous functions on $\mathbb{R}^d$. By Kantorovich duality for nonnegative lower-semicontinuous costs on Polish spaces [Villani, 2008],

$$\mathrm{OT}_c(\mu,\nu) = \sup_{V\in C_b(\mathbb{R}^d)} \left\{\int V^c\,d\mu + \int V\,d\nu\right\}.$$

It follows that

$$\mathrm{OT}_c(\mu,\nu) \le \sup_{V\in\mathcal{V}} \inf_{T\in\mathcal{T}} \mathcal{L}(T,V).$$

Combining the two inequalities proves the proposition. □

Although the maximin value is preserved, an optimal Kantorovich potential need not belong to $\mathcal{V}$. For example, consider the quadratic-cost case $c(x, y) = \frac{1}{2}\|x - y\|^2$ (corresponding to $r = 2$), $\mu = \mathcal{N}(0, 1)$, and $\nu = T_\sharp\mu$ with the optimal map $T(x) = x^{1/3}$. The derivative of the optimal potential satisfies $V'(y) = y - T^{-1}(y) = y - y^3$, giving $V(y) = \frac{1}{2}y^2 - \frac{1}{4}y^4$ up to an additive constant. Its quartic growth exceeds the quadratic growth permitted in $\mathcal{V}$. This example shows that the consistency of the maximin value does not by itself guarantee attainment by a potential in the chosen space.

In contrast, V is guaranteed to stay within $\mathcal{V}$ if the target measure ν has bounded support. In this case, V is Lipschitz continuous and its gradient is bounded, meaning it grows at most linearly [Santambrogio, 2015, Proposition 1.17]. Also, for Gaussian-to-Gaussian transport, the optimal V is a quadratic form, which naturally belongs to $\mathcal{V}$.

B.2 Proof of Theorem 3.3

Theorem B.4 (Restated Theorem 3.3 with general cost c)**.** *Let c be a nonnegative lower-semicontinuous cost satisfying* (21)*, and let $\mu, \nu \in \mathcal{P}_r(\mathbb{R}^d)$. If a global solution $(T^\star, V^\star)$ of Semi-Dual OT formulation exists in $\mathcal{T} \times \mathcal{V}$, then $(T^\star, V^\star)$ is non-spurious if and only if $\nabla_V \mathcal{L}(T^\star, V^\star) = 0$, where $\nabla_V \mathcal{L}(T^\star, \cdot)$ is the Fréchet derivative in the Banach space $\mathcal{V}$.*

Proof. Since $\mathcal{L}(T, V)$ is affine in V, its Fréchet derivative is

$$D_V \mathcal{L}(T, V)[\varphi] = \int_{\mathcal{Y}} \varphi(y)\, d\nu(y) - \int_{\mathcal{X}\times\mathcal{Z}} \varphi(T(x, z))\, d(\mu \times \xi)(x, z). \tag{22}$$

Indeed, this is a bounded linear functional on $\mathcal{V}$, since

$$\begin{aligned} |D_V \mathcal{L}(T, V)[\varphi]| &\le \|\varphi\|_{\mathcal{V}} \left(2 + \int \|y\|^r\, d\nu(y) + \int \|T(x, z)\|^r\, d(\mu \times \xi)\right) \\ &< \infty, \end{aligned}$$

where we used $\nu \in \mathcal{P}_r(\mathbb{R}^d)$ and $T \in \mathcal{T}$.

Consequently,

$$D_V \mathcal{L}(T^\star, V^\star) = 0 \iff \int \varphi\, d\nu = \int \varphi\, d(T^\star_\sharp(\mu \times \xi)) \quad \forall \varphi \in \mathcal{V}.$$

Since $C_c(\mathbb{R}^d)$ (set of continuous functions with compact support) is contained in $\mathcal{V}$ and separates Borel probability measures. Thus the integral condition is equivalent to $T^\star_\sharp(\mu \times \xi) = \nu$.

Finally, when this condition holds, the terms involving $V^\star$ cancel, giving

$$\mathcal{L}(T^\star, V^\star) = \int_{\mathcal{X}\times\mathcal{Z}} c(x, T^\star(x, z))\, d(\mu \times \xi).$$

Since $(T^\star, V^\star)$ is a global maximin point, Proposition B.3 implies that this value equals $\mathrm{OT}_c(\mu, \nu)$. Therefore, $T^\star$ is a stochastic OT map. □

C Proofs in Section 4

Proposition C.1 (Restated proposition 4.1)**.** *If $c(x, y) = \alpha\|x - y\|^2$ is a quadratic cost, then the functional $\mathcal{L} : \mathcal{T} \times (\mathcal{V} \cap \mathcal{S}_c) \to \mathbb{R}$ is convex-concave. Moreover, the problem* (16) *is a constrained convex-concave maximin optimization.*

Proof. To prove that the functional $\mathcal{L}(T, V)$ is convex-concave on $\mathcal{T} \times (\mathcal{V} \cap \mathcal{S}_c)$, we must show that:

1. For any fixed $T \in \mathcal{T}$, the mapping $V \mapsto \mathcal{L}(T, V)$ is concave.
2. For any fixed $V \in \mathcal{V} \cap \mathcal{S}_c$, the mapping $T \mapsto \mathcal{L}(T, V)$ is convex.
3. The set $\mathcal{T} \times (\mathcal{V} \cap \mathcal{S}_c)$ is convex.

Step 1. $V \mapsto \mathcal{L}(T, V)$ **is concave.** For a fixed $T \in \mathcal{T}$, the Semi-Dual loss is defined as:

$$\mathcal{L}(T, V) = \int_{\mathcal{X}\times\mathcal{Z}} [c(x, T(x, z)) - V(T(x, z))]\, d(\mu \times \xi) + \int_{\mathcal{Y}} V(y)\, d\nu(y).$$

Observe that the terms involving V are:

$$J_T(V) = \int_{\mathcal{Y}} V(y)\, d\nu(y) - \int_{\mathcal{X}\times\mathcal{Z}} V(T(x, z))\, d(\mu \times \xi).$$

This expression is a linear functional of V. Since every linear functional is both convex and concave, the mapping $V \mapsto \mathcal{L}(T, V)$ is concave (specifically, it is affine).

Step 2. $T \mapsto \mathcal{L}(T, V)$ **is convex.** For a fixed $V \in \mathcal{V} \cap \mathcal{S}_c$, let $h(x, y) := c(x, y) - V(y)$. Substituting the quadratic cost $c(x, y) = \alpha\|x - y\|^2$, we have:

$$h(x, y) = \alpha\|x\|^2 - 2\alpha\langle x, y\rangle + \alpha\|y\|^2 - V(y).$$

By the definition of the set $\mathcal{S}_c$ in the context of quadratic Semi-Dual OT, it consists of functions V such that the map $y \mapsto \alpha\|y\|^2 - V(y)$ is convex. Since the term $\alpha\|x\|^2 - 2\alpha\langle x, y\rangle$ is linear in y (and thus convex), the sum $h(x, \cdot)$ is a convex function for each $x \in \mathcal{X}$. Now, consider the loss as a function of T:

$$\mathcal{L}(T, V) = \int_{\mathcal{X}\times\mathcal{Z}} h(x, T(x, z))\, d(\mu \times \xi) + \text{constant in } T.$$

For any $T_1, T_2 \in \mathcal{T}$ and $\lambda \in [0, 1]$, the convexity of $h(x, \cdot)$ implies:

$$h(x, \lambda T_1 + (1 - \lambda)T_2) \le \lambda h(x, T_1) + (1 - \lambda)h(x, T_2).$$

Integrating both sides over $(\mathcal{X} \times \mathcal{Z}, \mu \times \xi)$ preserves the inequality. Thus, $T \mapsto \mathcal{L}(T, V)$ is convex.

Step 3. The domain is a convex set. The problem (16) is given by $\max_{V\in\mathcal{V}\cap\mathcal{S}_c} \min_{T\in\mathcal{T}} \mathcal{L}(T, V)$.

- The domain $\mathcal{T}$ and $\mathcal{V}$ are vector spaces, which are convex sets.
- If $V_1, V_2 \in \mathcal{S}_c$, then $g_i(y) = \alpha\|y\|^2 - V_i(y)$ are lower semi-continuous convex functions. Any convex combination $\lambda g_1 + (1 - \lambda)g_2 = \alpha\|y\|^2 - (\lambda V_1 + (1 - \lambda)V_2)$ is also a lower semi-continuous convex function, so $\lambda V_1 + (1 - \lambda)V_2 \in \mathcal{S}_c$. This implies that $\mathcal{S}_c$ is convex.

Since the objective $\mathcal{L}$ is convex-concave and the constraint sets $\mathcal{T}$ and $\mathcal{V} \cap \mathcal{S}_c$ are convex, the problem is a constrained convex-concave maximin optimization. □

Theorem C.2 (Restated Theorem 4.2)**.** *Let $c(x, y) = \alpha\|x - y\|^2$ and $\mathcal{L}$ be the Semi-Dual loss. Suppose Assumption 3.2 holds. If there is non-deterministic OT map $T^\star$, i.e.,*

$$\mu\left(\{x \mid T^\star(x, \cdot)_\sharp\xi \text{ is non-Dirac measure almost surely}\}\right) > 0,$$

then the pair $(\tilde{T}, V^\star)$, where $\tilde{T}(x, z) := \mathbb{E}_{z'\sim\xi}[T^\star(x, z')]$ is an averaged map and $V^\star$ is an optimal potential, is a stationary spurious global solution of the constrained problem (16)*.*

Proof. Define

$$\tilde{T}(x) := \int_{\mathcal{Z}} T^\star(x, z)\, d\xi(z),$$

and regard $\tilde{T}$ as a map on $\mathcal{X} \times \mathcal{Z}$ that is constant in z.

Step 1. $\tilde{T} \in \mathcal{T}$**.** By Jensen's inequality,

$$\|\tilde{T}(x)\|^2 \le \int_{\mathcal{Z}} \|T^\star(x, z)\|^2\, d\xi(z).$$

Integrating with respect to μ gives

$$\int_{\mathcal{X}} \|\tilde{T}(x)\|^2\, d\mu(x) \le \int_{\mathcal{X}\times\mathcal{Z}} \|T^\star(x, z)\|^2\, d(\mu \times \xi)(x, z) < \infty.$$

Thus $\tilde{T}$ is square-integrable. Moreover, by the $L^2(\xi)$-continuity of $x \mapsto T^\star(x, \cdot)$, for any sequence $x_n \to x$,

$$\begin{aligned}\|\tilde{T}(x_n) - \tilde{T}(x)\|^2 &= \left\| \int_{\mathcal{Z}} \big(T^\star(x_n, z) - T^\star(x, z)\big)\, d\xi(z) \right\|^2 \\ &\leq \int_{\mathcal{Z}} \|T^\star(x_n, z) - T^\star(x, z)\|^2\, d\xi(z) \longrightarrow 0.\end{aligned}$$

Hence $\tilde{T}$ is continuous and belongs to $\mathcal{T}$.

Step 2. Global optimality of $(\tilde{T}, V^\star)$. Fix any $V \in \mathcal{V} \cap \mathcal{S}_c$. By the definition of $\mathcal{S}_c$, we can write $V(y) = \alpha\|y\|^2 - u(y)$ with u convex. Therefore,

$$h_V(x, y) := c(x, y) - V(y) = \alpha\|x\|^2 - 2\alpha\langle x, y\rangle + u(y)$$

is convex in y. Applying Jensen's inequality to u gives

$$\begin{aligned}h_V(x, \tilde{T}(x)) &= \alpha\|x\|^2 - 2\alpha \left\langle x, \int_{\mathcal{Z}} T^\star(x, z)\, d\xi(z) \right\rangle + u\left(\int_{\mathcal{Z}} T^\star(x, z)\, d\xi(z) \right) \\ &\leq \int_{\mathcal{Z}} \left[\alpha\|x\|^2 - 2\alpha\langle x, T^\star(x, z)\rangle + u(T^\star(x, z))\right]\, d\xi(z) \\ &= \int_{\mathcal{Z}} h_V(x, T^\star(x, z))\, d\xi(z).\end{aligned}$$

After integrating with respect to μ and adding $\int V\, d\nu$, we obtain

$$\mathcal{L}(\tilde{T}, V) \leq \mathcal{L}(T^\star, V) = \mathrm{OT}_c(\mu, \nu).$$

The equality follows because $T^\star$ is an OT map.

At $V = V^\star$, optimality of $(T^\star, V^\star)$ also gives

$$\mathcal{L}(\tilde{T}, V^\star) \geq \min_{T \in \mathcal{T}} \mathcal{L}(T, V^\star) = \mathcal{L}(T^\star, V^\star) = \mathrm{OT}_c(\mu, \nu).$$

Consequently,

$$\mathcal{L}(\tilde{T}, V^\star) = \mathrm{OT}_c(\mu, \nu), \qquad \tilde{T} \in \arg\min_{T \in \mathcal{T}} \mathcal{L}(T, V^\star).$$

Since $V^\star$ is an optimal potential, this proves that $(\tilde{T}, V^\star)$ is a global solution of (16).

Step 3. Stationarity of $(\tilde{T}, V^\star)$. Since $\mathcal{L}$ is affine in V, for every $V \in \mathcal{V} \cap \mathcal{S}_c$,

$$\begin{aligned}\left\langle \nabla_V \mathcal{L}(\tilde{T}, V^\star), V - V^\star \right\rangle &= \int_{\mathcal{Y}} (V - V^\star)(y)\, d\nu(y) - \int_{\mathcal{X}} (V - V^\star)(\tilde{T}(x))\, d\mu(x) \\ &= \mathcal{L}(\tilde{T}, V) - \mathcal{L}(\tilde{T}, V^\star) \\ &\leq 0,\end{aligned}$$

where the inequality follows from Step 2. This is precisely the stationary condition (18).

Step 4. Spuriousness of $(\tilde{T}, V^\star)$. Using $T^\star_\sharp(\mu \times \xi) = \nu$ and the definition of $\tilde{T}$, we have

$$\begin{aligned}&\int_{\mathcal{Y}} \|y\|^2\, d\nu(y) - \int_{\mathcal{Y}} \|y\|^2\, d(\tilde{T}_\sharp \mu)(y) \\ &\quad = \int_{\mathcal{X}} \left[\int_{\mathcal{Z}} \|T^\star(x, z)\|^2\, d\xi(z) - \|\tilde{T}(x)\|^2 \right]\, d\mu(x) \\ &\quad = \int_{\mathcal{X} \times \mathcal{Z}} \|T^\star(x, z) - \tilde{T}(x)\|^2\, d(\mu \times \xi)(x, z) > 0.\end{aligned}$$

The strict inequality follows because $T^\star(x, \cdot)_\sharp \xi$ is non-Dirac on a set of positive μ-measure. Thus $\tilde{T}_\sharp \mu \neq \nu$. Since $\tilde{T}$ is constant in z, this also implies $\tilde{T}_\sharp(\mu \times \xi) \neq \nu$. Therefore, $(\tilde{T}, V^\star)$ is a stationary spurious global solution. □

D Proofs in Section 5

Proposition D.1 (Restated proposition 5.1)**.** *Under the same assumptions as in Theorem 3.3, the set of global solutions to the Lagrangian-based formulation* (20) *is characterized by*

$$\{(T^\star, V) \in \mathcal{T} \times \mathcal{V} \mid T^\star \textit{is a (stochastic) OT map } \}.$$

Proof. Fix $T \in \mathcal{T}$ and let $\lambda_T = T_\sharp(\mu \times \xi)$. If $\lambda_T \neq \nu$, bounded continuous functions separate probability measures, so there exists bounded function $f \in \mathcal{V}$ such that $\int f \, d(\nu - \lambda_T) > 0$. Taking $V = tf$ and letting $t \to \infty$ makes the potential term arbitrarily large. If $\lambda_T = \nu$, the potential terms cancel for every V. Thus,

$$\sup_{V \in \mathcal{V}} \mathcal{L}(T, V) = \begin{cases} \displaystyle\int_{\mathcal{X} \times \mathcal{Z}} c(x, T(x, z)) \, d(\mu \times \xi)(x, z), & T_\sharp(\mu \times \xi) = \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

By Assumption 3.2, minimizing this expression over $T \in \mathcal{T}$ yields $\mathrm{OT}_c(\mu, \nu)$, with minimizers precisely the stochastic OT maps. For every such map, $\mathcal{L}(T, V)$ is independent of V, so every $V \in \mathcal{V}$ attains the inner supremum. This proves the claimed characterization. □

E Related Work: Generative Adversarial Networks

In this section, we elucidate the theoretical connection between Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] and neural OT. To emphasize the connection to GANs, we focus here on deterministic maps $T : \mathcal{X} \to \mathcal{Y}$, though our results extend naturally to the stochastic case.

Unlike Neural OT, GANs primarily enforce the push-forward constraint $T_\sharp \mu = \nu$ without explicitly minimizing a transport cost. This is typically achieved by minimizing a divergence via a minimax formulation:

$$\min_T \left\{ \mathrm{Div}(\nu, T_\sharp \mu) := \sup_{V \in \mathcal{F}} \int_{\mathcal{Y}} \ell(V(y)) \, (\nu - T_\sharp \mu)(dy) \right\}, \tag{23}$$

where the choice of the function ℓ and the constraint set $\mathcal{F}$ determines the specific divergence being optimized (e.g., Jensen-Shannon or Wasserstein) [Goodfellow et al., 2014, Arjovsky et al., 2017].

To bridge Semi-Dual OT and GANs, several studies [Xie et al., 2019, Liu et al., 2021, Lu et al., 2020, Leygonie et al., 2019, Choi et al., 2024] utilize a hybrid objective that minimizes a weighted sum of the transport cost and a divergence:

$$\min_{T : \mathcal{X} \to \mathcal{Y}} \int_{\mathcal{X}} c(x, T(x)) d\mu(x) + \lambda \cdot \mathrm{Div}(\nu, T_\sharp \mu) \tag{24}$$

where $\lambda > 0$ is a regularizer parameter. However, the global solutions may not satisfy $T_\sharp \mu = \nu$ [Gazdieva et al., 2025, Theorem 4.1], yielding another type of spurious solutions.

Recently, the Zero-Infinity GAN was introduced by Lee and Kim [2025], which is based on the following zero-infinity distance

$$\max_V \int_{\mathcal{Y}} V(y) \, (\nu - T_\sharp \mu) dy = \mathrm{Div}_{0/\infty}(T_\sharp \mu, \nu) := \begin{cases} 0 & \text{if } T_\sharp \mu = \nu \text{ almost surely} \\ \infty & \text{otherwise} \end{cases}. \tag{25}$$

This corresponds to the maximization form of the divergence in (23) obtained by letting $\ell(t) = t$ and removing the constraint set $\mathcal{F}$.

By substituting (25) into (24) with $\lambda = 1$, we recover the Lagrangian-based minimax formulation in (20), thereby establishing a close connection between formulations learning OT maps and transport maps more generally. It is also worth noting that the Lagrangian term in (19) coincides exactly with the maximization form of the zero-infinity distance (25).

F Pseudo Code for Two-Timescale Extragradient

In this section, we describe the two-timescale extragradient (TTS-EG) algorithm and explain how we choose the step sizes γ_T and γ_V.

Algorithm 1 Two-Timescale Extragradient (TTS-EG)

Require: Learning rates γ_T, γ_V, Batch size B, Initial parameters θ_1, ψ_1

for $t = 1, \ldots, K$ **do**

  Sample $\{x_i\}_{i=1}^B \sim \mu$, $\{y_i\}_{i=1}^B \sim \nu$, and $\{z_j\}_{j=1}^B \sim \xi$

  Step 1: Extrapolation (Look-ahead)

  Define empirical loss at (θ_t, ψ_t):

$$\hat{\mathcal{L}}_t \leftarrow \frac{1}{B}\sum_{i=1}^B \left[c(x_i, T_{\theta_t}(x_i, z_i)) - V_{\psi_t}(T_{\theta_t}(x_i, z_i)) + V_{\psi_t}(y_i)\right]$$

  $\bar{\theta}_t \leftarrow \theta_t - \gamma_T \cdot \nabla_\theta \hat{\mathcal{L}}_t$

  $\bar{\psi}_t \leftarrow \psi_t + \gamma_V \cdot \nabla_\psi \hat{\mathcal{L}}_t$

  Step 2: Actual Update

  Define empirical loss at extrapolated point $(\bar{\theta}_t, \bar{\psi}_t)$:

$$\bar{\mathcal{L}}_t \leftarrow \frac{1}{B}\sum_{i=1}^B \left[c(x_i, T_{\bar{\theta}_t}(x_i, z_i)) - V_{\bar{\psi}_t}(T_{\bar{\theta}_t}(x_i, z_i)) + V_{\bar{\psi}_t}(y_i)\right]$$

  $\theta_{t+1} \leftarrow \theta_t - \gamma_T \cdot \nabla_{\bar{\theta}} \bar{\mathcal{L}}_t$

  $\psi_{t+1} \leftarrow \psi_t + \gamma_V \cdot \nabla_{\bar{\psi}} \bar{\mathcal{L}}_t$

end for

return θ_{K+1}, ψ_{K+1}

As discussed in Section 2, TTS-EG can avoid convergence to strict non-minimax points under suitable conditions, including a sufficiently large ratio of the inner step size to the outer step size. We follow this ordering in both OT formulations.

In the Semi-Dual formulation

$$\max_{V \in \mathcal{V}} \min_{T \in \mathcal{T}} \mathcal{L}(T, V),$$

T is the inner variable, so we use $\gamma_T > \gamma_V$. In the Lagrangian-based formulation

$$\min_{T \in \mathcal{T}} \max_{V \in \mathcal{V}} \mathcal{L}(T, V),$$

V is the inner variable, so we use $\gamma_V > \gamma_T$. Setting $\gamma_T = \gamma_V$ recovers standard EG.

G Experiments Details

Our implementation is publicly available at:

`https://github.com/niyanda01/NOT-without-Spurious-Solutions`

G.1 2D Synthetic Data

In this section, we describe the experimental setup for the 2D synthetic data experiments shown in Figs. 1, 2, 3 and Table 1. These datasets, adopted from Choi et al. [2025], are known to exhibit spurious solutions, and have been reported to be difficult to train on without additional regularization or smoothing.

Dataset Description. Throughout this paragraph, let $x, y \in \mathbb{R}^2$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

- **Perpendicular:** We generate $x \sim \mu$ by sampling $x_1 \sim U([-1, 1])$ and setting $x_2 = 0$. Similarly, we generate $y \sim \nu$ by setting $y_1 = 0$ and sampling $y_2 \sim U([-1, 1])$.
- **Horizontal:** We generate $x \sim \mu$ by sampling $x_1 \sim U([-1, 1])$ and setting $x_2 = 0$. Similarly, we generate $y \sim \nu$ by sampling $y_1 \sim U([-1, 1])$ and setting $y_2 = 1$.
- **One-to-Many:** We generate $x \sim \mu$ by sampling $x_1 \sim U([-1, 1])$ and setting $x_2 = 0$. Similarly, we generate $y \sim \nu$ by sampling $y_1 \sim U([-1, 1])$ and $y_2 \in \{-1, 1\}$ uniformly at random.
- **Grid:** Let
$$\mathbb{P} := \mathrm{Cat}\left(\left\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\right\}, \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)\right).$$
We generate $x \sim \mu$ by sampling $x_1 \sim U([-1, 1])$ and $x_2 \sim \mathbb{P}$. Similarly, we generate $y \sim \nu$ by sampling $y_1 \sim \mathbb{P}$ and $y_2 \sim U([-1, 1])$.

Architectures. The transport map T is stochastic to ensure the existence of global solutions for One-to-Many dataset. For every source sample x, a noise sample $z \sim \mathcal{N}(0, \sigma^2)$ is drawn independently, and $(x, z) \in \mathbb{R}^3$ is passed through an MLP with two hidden layers of width 64 and ReLU activations. The noise scale s is set per dataset (Table 3). The potential V is an MLP of the same form on $y \in \mathbb{R}^2$ without imposing c-concavity constraint on V.

	Perpendicular	Horizontal	One-to-Many	Grid
Step-size schedule	cosine	constant	cosine	cosine
Noise scale s	3	1	10	1
Gradient evaluations	200K	200K	200K	1M
$W_2^2(\mu, \nu)$	$2/3$	1	1	$1/24$
D_{target} threshold	0.002	0.002	0.08	0.010
D_{cost} threshold	0.01	0.01	0.03	0.02

Table 3: Training setup and evaluation thresholds of the 2D experiments. $W_2^2(\mu, \nu)$ is computed with the squared Euclidean cost $\|x - y\|^2$. A run finds the OT map when D_{target} and D_{cost} are both below the thresholds.

Optimization. We compare five methods; GDmax on the semi-dual formulation, GDmax with smoothing, equal-timescale EG, and TTS-EG on the semi-dual and Lagrangian formulations. All methods minimize over T and maximize over V the empirical loss of Algorithm F with the cost $c(x, y) = \frac{1}{2}\|x - y\|^2$. Every gradient evaluation uses 1,024 samples from μ and 1,024 samples from ν.

- **EG and TTS-EG** take plain gradient steps (SGD without momentum). The extrapolation and update steps of an iteration use the same samples of x and y, and z is redrawn at each evaluation of T.

Method	γ_T	γ_V	D_{cost} (↓)	D_{target} (↓)	OT map found
Data 1: Perpendicular					
GDmax	3×10^{-4}	3×10^{-4}	0.1303 ± 0.0313	0.5634 ± 0.0981	0/5
GDmax + Smoothing	3×10^{-4}	3×10^{-4}	0.1031 ± 0.0060	0.0114 ± 0.0008	0/5
EG	3×10^{-3}	3×10^{-3}	0.0118 ± 0.0167	0.0023 ± 0.0038	4/5
TTS-EG (Semi-Dual)	6×10^{-3}	3×10^{-3}	0.0025 ± 0.0013	0.0003 ± 0.0001	5/5
TTS-EG (Lagrangian)	1×10^{-3}	3×10^{-3}	$\mathbf{0.0018 \pm 0.0015}$	$\mathbf{0.0005 \pm 0.0001}$	**5/5**
Data 2: Horizontal					
GDmax	1×10^{-4}	1×10^{-4}	0.0025 ± 0.0016	0.0007 ± 0.0000	5/5
GDmax + Smoothing	1×10^{-4}	1×10^{-4}	0.0028 ± 0.0011	0.0005 ± 0.0003	5/5
EG	1×10^{-3}	1×10^{-3}	$\mathbf{0.0003 \pm 0.0002}$	$\mathbf{0.0007 \pm 0.0000}$	**5/5**
TTS-EG (Semi-Dual)	1×10^{-3}	1×10^{-4}	0.0005 ± 0.0006	0.0008 ± 0.0000	5/5
TTS-EG (Lagrangian)	1×10^{-4}	1×10^{-3}	0.0037 ± 0.0019	0.0007 ± 0.0000	5/5
Data 3: One-to-Many					
GDmax	1×10^{-4}	1×10^{-4}	0.6126 ± 0.7502	1.4156 ± 1.7103	3/5
GDmax + Smoothing	1×10^{-4}	1×10^{-4}	0.0062 ± 0.0034	0.0541 ± 0.0306	3/5
EG	5×10^{-4}	5×10^{-4}	0.0144 ± 0.0062	0.0196 ± 0.0061	5/5
TTS-EG (Semi-Dual)	5×10^{-4}	$\frac{5}{3} \times 10^{-4}$	0.0221 ± 0.0022	0.0332 ± 0.0055	5/5
TTS-EG (Lagrangian)	1×10^{-4}	5×10^{-4}	$\mathbf{0.0070 \pm 0.0038}$	$\mathbf{0.0218 \pm 0.0060}$	**5/5**
Data 4: Grid					
GDmax	1×10^{-4}	1×10^{-4}	0.0086 ± 0.0061	0.0371 ± 0.0126	0/5
GDmax +Smoothing	1×10^{-4}	1×10^{-4}	0.0055 ± 0.0013	0.0082 ± 0.0030	3/5
EG	5×10^{-4}	5×10^{-4}	0.0232 ± 0.0020	0.0216 ± 0.0021	0/5
TTS-EG (Semi-Dual)	7×10^{-3}	7×10^{-4}	$\mathbf{0.0069 \pm 0.0029}$	$\mathbf{0.0083 \pm 0.0027}$	**4/5**
TTS-EG (Lagrangian)	2×10^{-4}	6×10^{-4}	0.0372 ± 0.0413	0.0097 ± 0.0031	2/5

Table 4: Quantitative comparison on 2D synthetic data. D_{cost} and D_{target} are the mean $\pm$ standard deviation over the seeds that did not diverge. The last column counts the seeds, out of five, that found the OT map. Both metrics are highlighted in bold for the method that achieves the lowest $D_{\text{cost}} + D_{\text{target}}$ among those with the highest number of success.

- **GDmax** takes $K = 10$ Adam steps on T for every Adam step on V, with $(\beta_1, \beta_2) = (0.9, 0.999)$.

All methods compared after the same number of gradient evaluations: 200K, or 1M on Grid. An EG iteration evaluates the gradient twice, and a GDmax iteration evaluates it $K + 1 = 11$ times, which gives the iteration counts in Table 3. Within a dataset, all methods use the same step-size schedule and only the step sizes γ_T and γ_V differ. Under the cosine schedule, both step sizes at iteration t of N are $\frac{\gamma}{2}\big(1 + \cos\frac{\pi t}{N}\big)$, where γ is the value in Table 4. The step sizes of EG and TTS-EG were tuned separately for each dataset.

Evaluations. We train each setting with five random seeds. After training, the transport cost $\mathbb{E}\|x - T(x, z)\|^2$ is averaged over $10 \times 2{,}048$ samples of $x \sim \mu$, and D_{cost} is its distance to $W_2^2(\mu, \nu)$. D_{target} is the exact empirical W_2^2 between 4,096 samples of $T_\sharp(\mu \times \xi)$ and 4,096 samples of ν, computed with POT [Flamary et al., 2021] and averaged over three draws. A run counts the found OT map when both metrics are below the thresholds of Table 3. A run diverges when its parameters or losses become non-finite, and a diverged run counts as not found. Table 4 reports the mean and standard deviation of each metric over the seeds that did not diverge, whether or not they found the OT map.

Table 4 shows the results, and Figure 5 show the learned maps of each settings.

- **Perpendicular.** GDmax sends all the mass to a single point of ν on every seed. EG finds the OT map on four seeds, and both TTS-EG variants on all five.
- **Horizontal.** All methods find the OT map on every seed.
- **One-to-Many.** EG and both TTS-EG variants find the OT map on every seed. GDmax finds it on three seeds; on the other two, the mass collapse to one line.

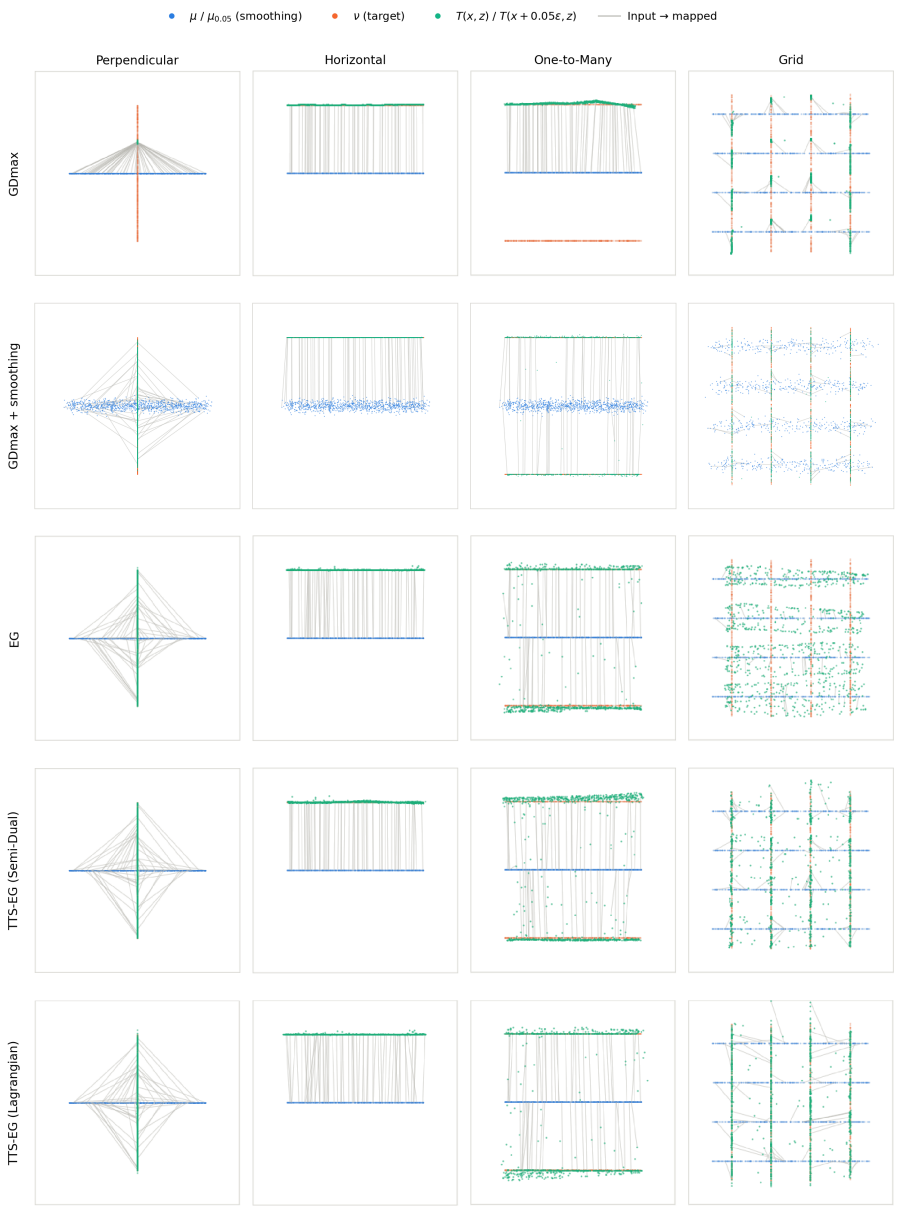


Figure 5: Transport maps learned on the 2D synthetic datasets (rows) by each method (columns), with the settings of Table 4. Each panel shows 1,000 samples of μ (blue) and ν (orange), $T(x, z)$ for the same x (green), and 60 of the pairs $x \to T(x, z)$ (gray).

- **Grid.** This is the hardest dataset. GDmax gathers the mass into clumps on the columns of ν, and EG stalls with the mass spread between the rows of μ. TTS-EG with Semi-dual formulation finds the OT map on four seeds. TTS-EG with Lagrangian-based formulation finds it on two seeds but diverges at one seed.

Compared with GDmax + Smoothing, both TTS-EG variants find the OT map on more seeds in the Perpendicular and One-to-Many settings (Table 4). TTS-EG with Semi-Dual also succeeds on more seeds in the Grid setting. Figure 6 shows the training curves of all methods across the four settings.

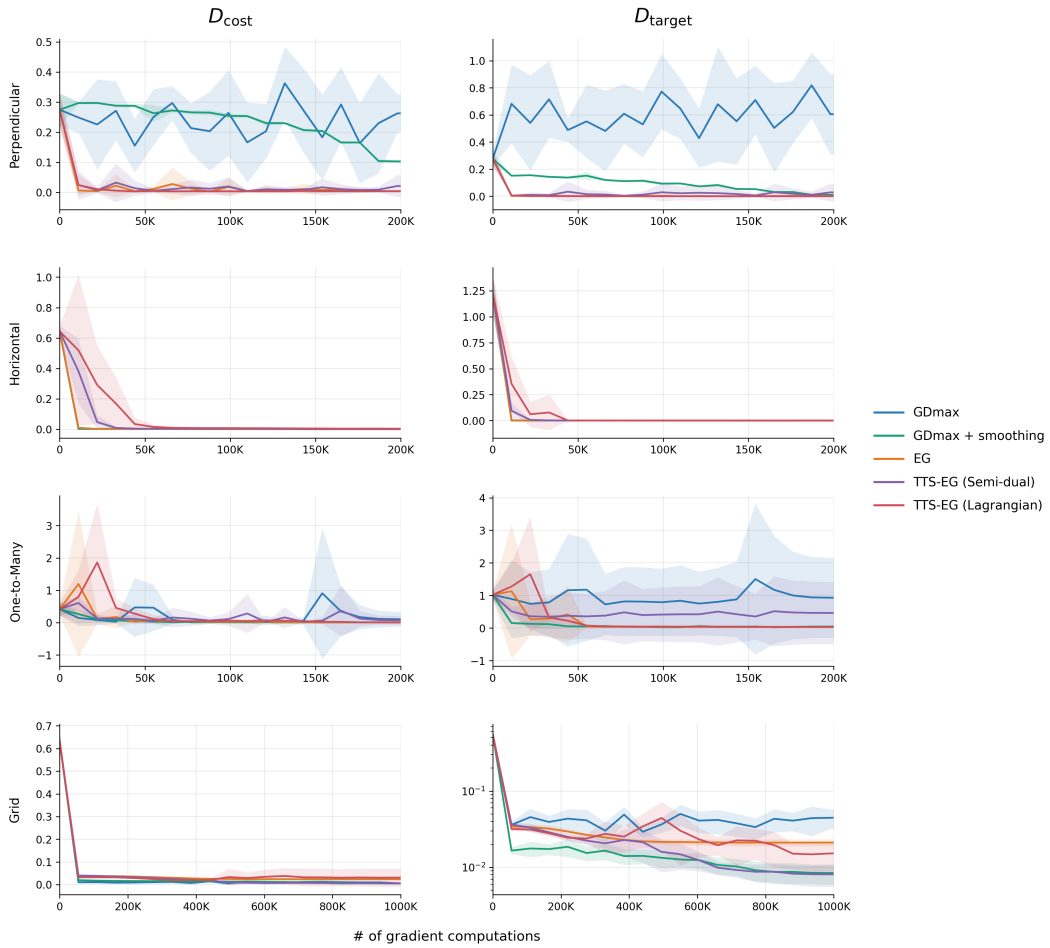


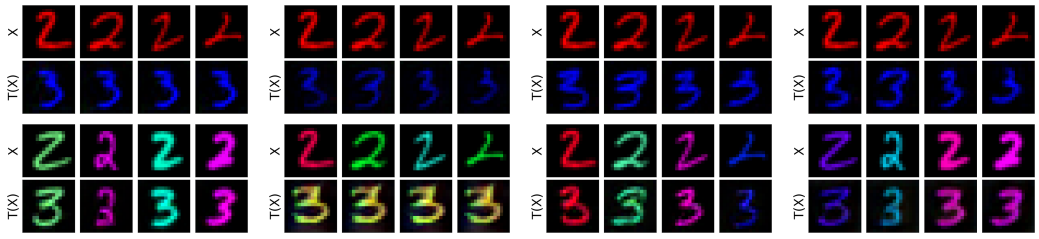
Figure 6: Training curves on four 2D synthetic datasets. The left and right columns show D_{cost} and D_{target}, respectively, against the number of gradient computations. Lines show means, and shaded bands indicate $\pm$ one standard deviation, computed over the non-divergent runs among five independent runs per method.

G.2 MNIST Data

We evaluate deterministic neural optimal transport maps on three unpaired MNIST digit translation tasks.

Dataset description. All images are represented in RGB and normalized to $[-1, 1]$. We consider the following settings:

- **Red 2 to Blue 3.** The source consists of digit 2 images rendered in the red channel, and the target consists of digit 3 images rendered in the blue channel. Images are resized to 16×16 pixels. This construction makes the source and target distributions to be almost orthogonal in the RGB color space, facilitating the emergence of spurious global solutions.
- **Multi-colored 2 to 3.** The source and target consist of digit 2 and digit 3 images, respectively, resized to 16×16 pixels. For each image in both the source and target sets, we sample a hue uniformly from $[0°, 360°)$ and map the original grayscale intensity to the corresponding RGB values.
- **Red even to Blue odd.** The source contains all even digits, $\{0, 2, 4, 6, 8\}$, rendered in red, while the target contains all odd digits, $\{1, 3, 5, 7, 9\}$, rendered in blue. Images are resized to 32×32 pixels.



(a) GDmax (Semi-Dual) (b) Extragradient (c) TTS-EG (Semi-Dual) (d) TTS-EG (Lagrangian)

Figure 7: Comparison of $2 \to 3$ transport maps across different optimization methods. TTS-EG (both Semi-dual and Lagrangian) identifies the optimal transport map.

Architectures. We use the following architectures for T and V:

- **Red 2 to Blue 3.** We use a residual CNN with three residual blocks for T and a four-block CNN for V, both with 128 hidden channels.
- **Multi-colored 2 to 3.** We use a seven-layer CNN for T and a four-block CNN for V, both with 128 hidden channels.
- **Red even to Blue odd.** We use a two-level U-Net for T and a ResNet with spectral normalization for V, both with 32 base channels.

We impose no c-concavity constraint on V. Full architectural details are provided in the released code.

Optimization. We compare GDmax on the semi-dual formulation, EG, and TTS-EG on the semi-dual and Lagrangian formulations. The step sizes and GDmax inner-loop counts are given in Table 5.

We use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$ for the $2 \to 3$ tasks. For Red even to Blue odd, we use $(\beta_1, \beta_2) = (0, 0.9)$ and train for $50K$ iterations with a batch size of 120. The learning rates follow a cosine decay schedule, updated every 500 iterations.

Evaluations. For our evaluation, we utilized 10% of the total dataset as the test set. All reported results represent the average values obtained from five independent experimental trials.

Table 5 shows that, in Red 2 to Blue 3, both TTS-EG variants achieve lower D_{target} than GDmax. TTS-EG with Semi-Dual loss attains the lowest D_{cost}, while TTS-EG with Lagrangian loss attains the lowest D_{target}. GDmax yields a relatively low D_{cost} but a high D_{target}, consistent with the spurious-solution failure case of the semi-dual formulation.

In Multi-colored 2 to 3, EG performs worse than both TTS-EG variants, particularly in D_{target}. However, GDmax also performs well, achieving a D_{target} comparable to that of TTS-EG with Semi-Dual loss. A possible explanation is that spurious solutions are absent in this setting.

To further compare EG and TTS-EG in a setting with spurious solutions, we extend the orthogonal red-to-blue construction to the Red even to Blue odd task. In this setting, both TTS-EG variants achieve lower D_{cost} and D_{target} than EG and GDmax. TTS-EG achieves the lower values in both formulation than EG and GDmax.

G.3 Handbags and Shoes Data

To examine whether our theoretical findings extend to more complex datasets, we learn deterministic transport maps from handbag photographs to shoe photographs.

Dataset description. We use Amazon Handbags from Zhu et al. [2016] and Yu and Grauman [2014], provided as `edges2handbags` and `edges2shoes` in pix2pix Isola et al. [2017]. Images are resized to 64×64 pixels and normalized to $[-1, 1]$, with handbags rendered in red and shoes in blue as in the Red-to-Blue MNIST settings.

Method	K	γ_T	γ_V	D_{cost} (↓)	D_{target} (↓)
Data 1: Red 2 to Blue 3					
GDmax	2	1×10^{-4}	1×10^{-4}	0.0233 ± 0.0316	0.0821 ± 0.0075
GDmax + smoothing	10	1×10^{-4}	1×10^{-4}	0.0447 ± 0.0478	0.0488 ± 0.0278
EG	–	1×10^{-4}	1×10^{-4}	0.0292 ± 0.0215	0.0488 ± 0.0236
TTS-EG (Semi-Dual)	–	2×10^{-4}	1×10^{-4}	$\mathbf{0.0171 \pm 0.0093}$	$\mathbf{0.0449 \pm 0.0065}$
TTS-EG (Lagrangian)	–	2×10^{-4}	5×10^{-4}	0.0343 ± 0.0256	0.0374 ± 0.0055
Data 2: Multi-colored 2 to 3					
GDmax	5	2×10^{-4}	2×10^{-4}	0.0432 ± 0.0042	0.0556 ± 0.0010
GDmax + smoothing	10	1×10^{-4}	1×10^{-4}	$\mathbf{0.0282 \pm 0.0060}$	$\mathbf{0.0616 \pm 0.0023}$
EG	–	2×10^{-4}	2×10^{-4}	0.0535 ± 0.0201	0.1399 ± 0.0411
TTS-EG (Semi-Dual)	–	5×10^{-4}	1×10^{-4}	0.0481 ± 0.0051	0.0559 ± 0.0014
TTS-EG (Lagrangian)	–	2×10^{-4}	5×10^{-4}	0.0422 ± 0.0087	0.0603 ± 0.0024
Data 3: Red even to Blue odd					
GDmax	5	1×10^{-4}	1×10^{-4}	0.0084 ± 0.0093	0.1269 ± 0.0179
GDmax + smoothing	10	1×10^{-4}	1×10^{-4}	0.0019 ± 0.0019	0.0579 ± 0.0009
EG	–	1×10^{-4}	1×10^{-4}	0.0366 ± 0.0360	0.1249 ± 0.0262
TTS-EG (Semi-Dual)	–	2×10^{-4}	1×10^{-4}	0.0045 ± 0.0039	0.0585 ± 0.0316
TTS-EG (Lagrangian)	–	1×10^{-4}	2×10^{-4}	$\mathbf{0.0019 \pm 0.0010}$	$\mathbf{0.0400 \pm 0.0008}$

Table 5: Hyperparameters and quantitative comparison on MNIST. K denotes the number of GDmax inner updates, and γ_T and γ_V denote the learning rates for T and V, respectively. Results are mean $\pm$ standard deviation over five runs. Both metrics are highlighted in bold for the method that achieves the lowest $D_{\text{cost}} + D_{\text{target}}$ among those with the highest number of success.

Method	K	γ_T	γ_V
GDmax	5	1×10^{-5}	1×10^{-5}
GDmax + smoothing	5	1×10^{-5}	1×10^{-5}
EG	–	1×10^{-5}	1×10^{-5}
TTS-EG (Semi-Dual)	–	2×10^{-5}	1×10^{-5}
TTS-EG (Lagrangian)	–	1×10^{-5}	5×10^{-5}

Table 6: Hyperparameters for Red Handbags to Blue Shoes. K is the number of GDmax inner updates; γ_T and γ_V denote the peak learning rates. GDmax + smoothing uses constant Gaussian source-noise with $\sigma = 1$.

Architectures. We use a ResNet encoder–decoder transport network T and a spectral-normalized ResNet potential V, both with base channel width 64 and approximately 4.8 million parameters.

Optimization. We train for $100K$ iterations with batch size 32 using Adam with $(\beta_1, \beta_2) = (0, 0.9)$, weight decay 10^{-10}, and gradient norm clipping at 10. The transport network is deterministic, and gradient penalties are disabled. The training cost is $c(x, y) = \|x - y\|_2^2/d$, where $d = 3 \times 64 \times 64$. Learning rates warm up over the first 500 iterations and then follow cosine decay to 25% of their peak values at $100K$. The peak learning rates and GDmax inner-loop count are listed in Table 6.

Evaluations. We report a single run per method, with both metrics reported at the checkpoint attaining the lowest observed D_{target}. As shown in Table 2, all EG-based methods achieve lower D_{target} than GDmax.

Computation resources Experiments were conducted on a GPU cluster, using one GPU per training run. On an NVIDIA RTX 6000 Ada GPU, TTS-EG with batch size 32 required approximately 0.34 seconds per training iteration.

G.4 Licenses and Assets

In this work, we utilize several existing assets and build upon established baselines. We have respected their terms of use and credited the original creators as follows:

- **Baseline Codebase**: Our implementation of Neural Optimal Transport (NOT) baselines is based on the official code provided by Korotin et al. [2023]. The codebase is used under the MIT License (Copyright (c) 2023 Alexander).
- **Dataset**: We use the publicly available MNIST dataset LeCun et al. [1998], which is provided under the Creative Commons Attribution-Share Alike 3.0 license. We also use Amazon Handbags from Zhu et al. [2016] and UT Zappos50K Yu and Grauman [2014]. UT Zappos50K is provided for academic, non-commercial use only.
- **Deep Learning Framework**: Our experiments are implemented in **PyTorch** Paszke et al. [2019], which is released under the BSD-style license.
- **Optimal Transport Library**: We utilize the **POT (Python Optimal Transport)** library Flamary et al. [2021] for evaluating Wasserstein distances and performing discrete optimal transport calculations. It is released under the MIT License.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our theoretical findings on spurious solutions and the effectiveness of TTS-EG, all of which are directly supported by the proofs and experimental results in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the primary limitations of our work, including the gap between theoretical functional spaces and practical parameter spaces, in the Conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are stated with their respective assumptions in the main text, and complete, rigorous proofs for all theorems and propositions are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide comprehensive details necessary for reproduction, including the specific neural network architectures, all optimization hyperparameters (summarized in Appendix G), and the precise protocols for dataset generation and normalization.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code will be made available through a GitHub repository. Detailed instructions for environment setup and reproduction of the experimental results are provided in Appendix G.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer:[Yes]

   Justification: We specify all experimental details, including data generation protocols, hyperparameter configurations (summarized in Table 5), and optimizer settings (TTS-EG), in Section 6 and Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation calculated across multiple independent runs with different random seeds. These error bars are included in all numerical tables (e.g., Table 5) and figures to ensure the statistical significance of our findings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on a GPU cluster, and we provide detailed information regarding the GPU models, average execution time per run, and total compute resources utilized in Appendix G.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: Since our work focuses on the theoretical aspects of optimal transports, we believe it does not have a direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly credit all existing assets, including the MNIST dataset, the PyTorch framework, and the POT (Python Optimal Transport) library. Additionally, our baseline implementation is built upon the codebase of Korotin et al. [2023], which is used under the MIT License. Detailed citations and license information are provided in Appendix G.4.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the source code for our proposed algorithm and experiments as a new asset. Detailed documentation, including environment setup and usage instructions, is included in the supplementary material and referenced in Appendix G

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This study did not involve any crowdsourcing experiments or research with human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This research does not involve human subjects or sensitive personal data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used to assist with manuscript writing and code implementation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.