STABILITY OF THE MONGE MAP IN SEMI-DUAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper shows that the semi-dual formulation of the optimal transport problem has a degenerate saddle-point structure, and that its numerical solution is equivalent to solving a constrained optimization problem. We derive necessary and sufficient conditions for the convergence of Monge maps without requiring optimality of the dual potential. This analysis helps explain why, in practice, numerical algorithms often require more iterations to update the transport map than the potential.

1 INTRODUCTION

Recently, optimal transport (OT) has become a powerful tool in machine learning, statistics, and signal processing. From a theoretical perspective, many modern generative models — including variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion models, and flow-based models — can be interpreted as approximations of transport processes between probability distributions. For instance, GANs can be viewed as learning a map that transports a simple latent distribution (e.g., uniform or Gaussian) to complex data distribution (e.g., a highly nonlinear data generating distribution). Similarly diffusion models describe stochastic flows that gradually transform one distribution into another. These connections suggest that *optimal transport provides a unifying mathematical framework for generative modeling*, where learning corresponds to approximating transport maps or transport dynamics between distributions.

This work is motivated by neural optimal transport (NOT) algorithm introduced in Rout et al. (2021), which is closely related to the $(2, 2)$-WGAN from Mallasto et al. (2019) and other WGAN formulations. Later, NOT algorithm was extended to weak formulations in Korotin et al. (2023). These approaches rely on the semi-dual formulation of optimal transport[1] together with amortized optimization techniques introduced in Dam et al. (2019), later interpreted as measurable selection mechanisms in the theory of integrands.

Altogether, these works exploit the existence (and, for quadratic cost, uniqueness) of optimal transport maps, but do not provide convergence guarantees for the learned maps. Many papers containing convergence estimates implicitly assume that the transport map exists and that the optimization problem is well-posed; however, these assumptions are mathematically nontrivial and may fail in practice, leading to misleading or impractical conclusions (see Sec. 5).

In this paper, we show that the semi-dual optimal transport problem possesses a degenerate saddle-point structure: at the optimal transport map, the objective functional becomes independent of the potential. We provide an estimate of the transport map which is independent of the potential:

$$\|T - T^\star\|_{L_2}^2 \lesssim \left|\mathbb{E}[c(x, T(x))] - W_2(\mu, \nu)^2\right| + d_{KR}(T_\sharp\mu, \nu). \tag{1}$$

While many works assess convergence through the approximation gap and the discrepancy between generated and target distributions, to the best of our knowledge such estimates have not been formulated explicitly in the form of equation 1.

Early works on the numerical solution of the semi-dual OT problem approximated the conjugate potential using an inner optimization loop Taghvaei & Jalali (2019). Later, with neural network parametrizations it was observed that performing multiple updates of the transport map leads to

[1]This formulation is already present in the seminal work of Kantorovich (1942); the optimization itself is stated explicitly in Ekeland (2005), while the term "semi-dual" is introduced in Genevay et al. (2016).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

better performance Makkuva et al. (2020). We provide a theoretical explanation for this phenomenon by interpreting the problem as a constrained optimization task and relating it to a continuous-time two-timescale dynamical system.

2 Optimal Transport Problem

2.1 Monge Formulation

We will assume that X is a complete separable metric space. Such spaces are referred to as Polish spaces. In practice, we consider either $X = \mathbb{R}^n$ or a compact subset of $\mathbb{R}^n$. By $\mathcal{P}(X)$ we denote the space of probability measures defined on the Borel σ-algebra $\mathcal{B}(X)$, i.e., the smallest σ-algebra generated by open subsets of X. $\mathcal{P}_p(X)$ denotes subspace of $\mathcal{P}(X)$ having pth moments.

Assume that Y is a Polish space and a lower semicontinuous (l.s.c.)[2] cost function is defined:$c\colon X \times Y \to [0, \infty)$. Later, we assume that X and Y are compact sets in $\mathbb{R}^n$ endowed with metric $d(x, y)$ and cost

$$c(x, y) = \frac{1}{p} d(x, y)^p. \tag{2}$$

Letting $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ be probability measures we can consider two probability spaces $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ with $\mathcal{X} = \mathcal{B}(X)$ and $\mathcal{Y} = \mathcal{B}(Y)$. For a $\mathcal{X}|\mathcal{Y}$-measurable function $T\colon X \to Y$, the measure ν such that $\nu(B) = \mu(T^{-1}(B))$ for all $B \in \mathcal{Y}$ is called a push-forward measure of μ by T. In this case, the following notation is used: $\nu = T_\sharp \mu$ $(= \mu \circ T^{-1})$.

Denote by $T(\mu, \nu)$ the space of measurable (Borel) functions $T\colon X \to Y$, for which ν is a push-forward measure of μ. Elements of $T(\mu, \nu)$ will be referred to as *transport maps*.

The Monge problem is to find a transport map $T^*\colon X \to Y$ that minimizes the transport cost

$$MP(\mu, \nu, c) = \inf_{T \in T(\mu,\nu)} \int_X c(x, T(x)) d\mu(x). \tag{3}$$

If the infimum is reached on some function T^*, then T^* is called *the optimal transport map.* Sometimes, the class $T(\mu, \nu)$ is empty (see Table 1).

2.2 Kantorovich Formulation

Denote by $\Pi(\mu, \nu)$ the set of probability measures π on the product space $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$ with prescribed marginal distributions: $\pi(A \times Y) = \mu(A)$, and $\pi(X \times B) = \nu(B), \quad \forall A \in \mathcal{X},\ B \in \mathcal{Y}$. Measures $\pi \in \Pi(\mu, \nu)$ are called *transport (or transportation) plans*.[3]

The Kantorovich problem consists in finding a transport plan π^* that minimizes the total transport cost

$$KP(\mu, \nu, c) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y). \tag{4}$$

The integral can be written as $\mathbb{E}[c(\xi, \eta)]$, where $(\xi, \eta) \sim \pi$. A plan π^* at which the minimum is attained is called an *optimal transport plan*. Under assumptions above[4] the infimum in equation 4 is actually attained (Villani, 2009, Theorem 5.10 (iii)).

2.3 Dual Problem

Problem equation 4 can be reformulated as a maximization problem (for generality we write supremum, but it is attained under very general assumptions:

$$DP(\mu, \nu, c) = \sup_{\substack{\varphi \in C(X), \psi \in C(Y) \\ \varphi(x) + \psi(y) \le c(x,y)}} \left(\int_X \varphi d\mu + \int_Y \psi d\nu \right) \tag{5}$$

[2]Continuous functions are l.s.c.

[3]The set $\Pi(\mu, \nu)$ is nonempty; for instance it contains the product measure $\mu \otimes \nu$.

[4]For unbounded domain additionally $c(x, y) \le a(x) + b(y)$ for some integrable functions $a \in L_1(\mu)$ and $b \in L_1(\nu)$.

where X and Y are compact sets, in general case φ and ψ should be absolutely integrable.

Under our assumptions (see, e.g., (Villani, 2009, Theorem 5.10 (iii))), $DP(\mu, \nu, c) = KP(\mu, \nu, c)$.

2.4 Semi-Dual Problem

For functions $\varphi\colon X \to \mathbb{R} \cup \{\pm\infty\}$ and $\psi\colon Y \to \mathbb{R} \cup \{\pm\infty\}$ define *c-transform*

$$\varphi^c(y) = \inf_{x \in X} [c(x, y) - \varphi(x)], \qquad \psi^c(x) = \inf_{y \in Y} [c(x, y) - \psi(y)]. \tag{6}$$

A function $\varphi(x)$ is said to be c-concave, if there exists a function $\psi(y)$, such that $\varphi(x) = \psi^c(x)$. For X and Y bounded subsets in $\mathbb{R}^n$ the dual problem of optimal transport can be formulated as semi-dual (see):[5]

$$SDP(\mu, \nu, c) = \sup_{\psi \in C(Y)} \left(\int_X \psi^c(x)\, d\mu + \int_Y \psi(y)\, d\nu \right). \tag{7}$$

2.5 Connection between the Monge and Kantorovich Problems

The following theorem establishes the relation between the Monge and Kantorovich formulations of optimal transport.

Theorem 1 (Pratelli (2007))**.** *Let X and Y be Polish spaces and let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ be probability measures. Assume that μ is nonatomic and that the cost function $c : X \times Y \to [0, \infty)$ is continuous. Then $MP(\mu, \nu, c) = KP(\mu, \nu, c)$.*

The theorem states that the infimum value of the Monge problem coincides with the minimum value of the Kantorovich problem. In particular, although the Monge problem may fail to admit a minimizing map, it is possible to approximate the optimal transport plan by transport maps whenever the source measure is nonatomic. In Table 1, we summarize the theoretical results on the existence of the optimal transport map according to Bogachev & Kolesnikov (2012).

Table 1: Existence and equivalence properties in optimal transport under different assumptions.

μ, ν	$c(x, y)$	$\pi^\star$	$\varphi^\star, \psi^\star$	$T^\star$	$KP = DP$	$KP = MP$
general	continuous	$\exists$	$\exists$	sometimes	yes	$\leq$
continuous	l.s.c.	$\exists$	$\exists$	sometimes	yes	$\leq$
μ non-atomic	continuous	$\exists$	$\exists$	sometimes	yes	yes
μ no $(n-1)$ mass	$\frac{1}{p}\lvert x - y\rvert^p,\ p > 1$	$\exists!$	$\exists$	$\exists!$	yes	yes
continuous	$\lvert x - y\rvert$	$\exists$	$\exists$	$\exists$	yes	yes

When X and Y are subsets of $\mathbb{R}^n$ and cost is quadratic $c(x, y) = \frac{1}{2}\|x - y\|^2$, the Kantorovich problem admits additional structure.[6]

Theorem 2 (Brenier–McCann)**.** *Let X and Y be compact sets in $\mathbb{R}^n$ or $X = Y = \mathbb{R}^n$, $c(x, y) = \frac{1}{2}\|x - y\|^2$, and suppose μ vanishes on all sets of Hausdorff dimension at most $n - 1$ (for example, if μ is absolutely continuous with respect to Lebesgue measure). Then there exists a Borel map $T\colon X \to Y$ such that $\nu = T_\sharp \mu$. Moreover: (i) $T = \nabla u$ for some convex function u; (ii) T is μ-a.e. unique; (iii) T is the unique solution to the Monge problem.*

If ν satisfies the same condition, then there exists an inverse map $S = \nabla v$, where v is the Legendre–Fenchel dual of u, such that $S \circ T(x) = x$ and $T \circ S(y) = y$ almost surely.

This theorem provides insights into situations in which the Monge map may fail to exist. (Villani, 2009, Theorem 5.10) says that the support[7] of the optimal plan belongs to a c-superdifferential of the potential:

$$\operatorname{sp} \mu \subset \partial^c \varphi = \{(x, y) \in X \times Y : \varphi(x) + \varphi^c(y) = c(x, y)\}. \tag{8}$$

[5] For unbounded domains, we should assume $\psi \in L_1(d\nu)$.

[6] Brenier (1991) for compact sets and McCann (1995) in case of $X = Y = \mathbb{R}^n$.

[7] In case of a separable metric space X, *support of a measure* $\mu \in \mathcal{P}(X)$ can be defined as the smallest closed set $B \subset X$ with $\mu(B) = 1$.

In case of quadratic cost, a c-concave function can be represented as $\varphi(x) = \frac{|x|^2}{2} - u(x)$, where u is convex function. Because a convex function is Lipschitz, the potential φ is Lipschitz, that means it is differentiable almost everywhere, and the c-superdifferential coincides with $\nabla\varphi$ at almost every point. Consequently, the support of the optimal plan $\operatorname{sp}\pi$ is concentrated on the graph of $\nabla\varphi$, up to a set of measure zero where the superdifferential may be multivalued. In this case, the set of points where the transport map is non-unique has zero measure.

In contrast, if the set of points where φ is not differentiable has positive measure, then the c-superdifferential is genuinely multivalued on a set of positive measure. As a result, the support of π cannot be represented as the graph of a function, and therefore no Monge map exists.

3 Towards a Minimax Formulation

Given that many applications with bounded domains, e.g., in images the components belong to $[0, 1]$ or $[0, 255]$, from now we assume that X and Y are compact subsets in $\mathbb{R}^n$.

3.1 Saddle-Point Formulation

By interchange theorem (see Rockafellar & Wets (2009, Th. 14.60) or Rockafellar (1976)), one can rewrite (7) in a form that involves measurable selections:

$$SDP(\mu, \nu,) = \sup_{\psi \in C(Y)} \left(\inf_{t \in \mathcal{M}(X,Y)} \int_X \big[c(x, t(x)) - \psi(t(x))\big] d\mu + \int_Y \psi(y)\, d\nu \right). \tag{9}$$

Where $\mathcal{M}(X, Y)$ denotes all measurable functions from X to Y. It can be replaced by L_p if the integral converges ($p \geq 1$).

In the quadratic case $c(x, y) = \frac{1}{2}\|x - y\|^2$, one expects (and can show under the stated assumptions) that the infimum in (9) is attained at the optimal Monge map $t(x) = T^\star(x)$, while the supremum is attained at the Kantorovich potential $\psi^\star$. The following theorem formalizes it.

Theorem 3. *Let X and Y be compact subsets in $\mathbb{R}^n$, and let μ, ν be probability measures both absolutely continuous with respect to the Lebesgue measure. Let the cost $c\colon X \times Y \to \mathbb{R}$ be l.s.c. For measurable maps $t\colon X \to Y$ and $\psi\colon Y \to \mathbb{R}$ $(\cup\{\pm\infty\})$, such that $\psi \in C(Y)$, define*

$$F(\psi, t) = \int_X c(x, t(x))\, d\mu + \int_Y \psi(y)\, d(\nu - t_\sharp\mu). \tag{10}$$

Then:

(i) The minimax value satisfies

$$\sup_\psi \inf_t F(\psi, t) = \sup_\psi \left(\int_X \psi^c\, d\mu + \int_Y \psi\, d\nu \right), \tag{11}$$

which coincides with the Kantorovich dual problem.

(ii) For all ψ $\inf_t F(\psi, t) \leq \inf_t F(\psi^{cc}, t)$. with equality reached on c-concave functions ψ only.

(iii) If $t_\sharp\mu \neq \nu$,

$$\sup_\psi F(\psi, t) = +\infty. \tag{12}$$

(iv) If in addition the cost $c(x, y) = \frac{1}{p}|x - y|^p, \quad 1 < p < +\infty$. There exists an optimal potential $\psi^\star$ and a measurable map $T^\star$ such that for all ψ and t

$$F(\psi^\star, t) \geq F(\psi^\star, T^\star) = F(\psi, T^\star). \tag{13}$$

Moreover, $T^\star(x) = \arg\min_y [c(x, y) - \psi^\star(y)]$ μ-a.e. and $T^\star$ coincides with the Brenier optimal transport map from μ to ν

$$T^\star(x) = x - |\nabla\psi^{\star c}(x)|^{q-2}\nabla\psi^{\star c}(x) \quad \mu\text{-a.e.}, \quad \frac{1}{q} = 1 - \frac{1}{p}. \tag{14}$$

(v) The map $T^\star$ is unique μ-almost everywhere and is the unique minimizer of $t \mapsto F(\psi^\star, t)$.

The equality in *(iv)* reveals a fundamental structural property of the formulation: flatness of the functional in the dual variable. More precisely, at the optimal transport map $t = T^\star$, the value of the functional becomes independent of ψ. This degeneracy closely parallels well-known instability phenomena in generative adversarial networks, where the discriminator may continue to change without affecting the objective once the generator approaches optimality. The works we survey either implicitly treat the saddle point as strict or use formulations that do not foreground the constrained-optimization structure (Section 5.5) making the degeneracy apparent; we are not aware of an explicit treatment elsewhere in the semi-dual optimal transport literature.

This observation has important consequences for existing convergence results. In particular, several works — including Fan et al. (2023, Assumption 3), Rout et al. (2021, Theorem 4.3) (convexity assumption), Tarasov et al. (2026, Theorem 4.6), and Vacher & Vialard (2021, Lemma 1) — rely on properties derived under the assumption that the potential ψ is optimal. However, due to the flatness of the objective, optimality of ψ is neither necessary nor enforced by the optimization dynamics once the transport map is close to optimal. As a result, convergence of the map can be achieved without convergence of the potential, as demonstrated in Section 5.

Connection (14) is used in Makkuva et al. (2020), Korotin et al. (2019), Taghvaei & Jalali (2019) and other works (see Table 2).

4 Stability of transport maps

4.1 Convergence

Recall the notion of the Kantorovich–Rubinstein distance in the space of probability measures $\mathcal{P}(X)$.

$$d_{KR}(\mu, \nu) = \sup_{\substack{\|\varphi\| \leq 1, \\ |\varphi(x) - \varphi(y)| \leq d(x,y)}} \left(\int_X \varphi(x)\, d\mu(x) - \int_X \varphi(x)\, d\nu(x) \right). \tag{15}$$

Kantorovich introduced metric in $\mathcal{P}_1(X)$ as $W_1(\mu, \nu) = KP\big(\mu, \nu, d(x, y)\big)$ and using dual formulation (5) showed that

$$W_1(\mu, \nu) = \sup_{|\varphi(x) - \varphi(y)| \leq d(x,y)} \left(\int_X \varphi(x)\, d\mu(x) - \int_Y \varphi(y)\, d\nu(y) \right). \tag{16}$$

It has been generalized to *Kantorovich power metric*[8] $W_p(\mu, \nu)^p = KP\big(\mu, \nu, \frac{1}{p} d(x, y)^p\big)$. Furthermore, d_{KR} is the metric in $\mathcal{P}(X)$ and convergence in it is equivalent to weak convergence. Similarly, W_p is a metric in $\mathcal{P}_p(X)$ and convergence in it in addition to weak convergence guarantees the convergence of the first p moments.

Theorem 4. *Let the assumptions of Theorem 3 hold, and $c(x, y) = d(x, y)^2/2$. In addition, assume that measures μ and ν have densities $p \in C^{0,\alpha}(X)$ and $q \in C^{0,\alpha}(Y)$ with $\alpha \in (0, 1)$. Additionally, p, q are bounded away from zero on X, Y, and X, Y have $C^{2,\alpha}$ smooth boundaries and are uniformly convex. Assume that $t_k \in L^2(\mu)$ and the family $\{\psi_k\}$ is uniformly L-Lipschitz.*[9] *Then the following estimate is true*

$$\|t_k - T^\star\|^2_{L^2(\mu)} \leq C \left(\big| F(t_k, \psi_k) - W_2(\mu, \nu)^2 \big| + d_{KR}(t_{k\sharp}\mu, \nu) \right). \tag{17}$$

In practice, it is rarely possible to compute the exact Kantorovich distance. Instead, we rely on observable quantities derived from the optimization objective.

[8] The distance W_1 (denoted as W in Kantorovich (1942)) was interpreted by Kantorovich as the minimal "work" required to transport mass. In the modern literature, the notation W_p is often mistakenly referred to as the Wasserstein distance. This misattribution was mentioned in multiple publications including L. Vasershtein's friend remark: "It is especially ironic to find the Kantorovich metric called the Vasershtein metric... Vasershtein's interesting article Vaserstein (1969)... does contain in passing a definition of the Kantorovich metric... But there is no definition of power metrics" Vershik (2013).

[9] In practice, parametrization $t_{\theta_k}, \psi_{\phi_k}$ by neural networks with parameters θ_k and ϕ_k is used.

We first observe that the functional F consists of two components with distinct roles. Next we notice that the first term acts as a stabilizing force, favoring maps close to the identity, while the second term enforces the marginal constraint $t_\sharp\mu = \nu$ preventing collapse to the trivial solution. Finally, since the constraint is only enforced approximately, there is no guarantee that the transport cost reaches the optimal value at $F(t, \psi) = W_2(\mu, \nu)^2$. Therefore, convergence cannot be assessed using a single quantity. Simultaneous improvement in both quantities provides empirical evidence of convergence toward the optimal transport map.

Remark 1. *Some non-smooth domains can be considered in Theorem 4. For example, multidimensional cubes $X = Y = [0, 1]^d$ due to the mirroring argument.*

Remark 2. *The reverse estimate is true under the same assumptions, i.e., from $t_k \to T^\star$ in $L^2(\mu)$ and ψ_k is uniformly L-Lipschitz it follows that $F(t_k, \psi_k) \to W_2(\mu, \nu)^2$ and $d_{KR}(t_{k\sharp}\mu, \nu) \to 0$.*

4.2 Hyperparameters of the Numerical Algorithm

In the NOT algorithm by Korotin et al. (2021), several gradient descent steps are performed to update t before a single gradient ascent step is applied to ψ. Statement (iii) of Theorem 3 provides insight into this design choice.

During training, the constraint $t_\sharp\mu = \nu$ is not satisfied so the maximization problem is unbounded. Assume that $d(t_\sharp\mu) = \rho_t(y)dy$ and $d\nu = q(y)dy$. If we look at the gradient, $\nabla_\psi F = \rho_t - q$, we see that the gradient ascent in ψ causes an increase in that direction without any restoring force, due to absence of concavity of F in ψ. This can cause instability if the variable t is not sufficiently close to optimal value. On the other hand, performing of multiple steps of optimization in t reduces both the transport cost and the constraint violation $\|\rho_t - q\|$, which results in smaller step in ψ.

This naturally leads to a two-timescale interpretation of the training dynamics: the primal variable t evolves on a faster timescale, stabilizing the system, while the dual variable ψ evolves more slowly to enforce the constraint (see Sec. 5 for dynamical system analogy).

One step of update looks as follows:

$$\begin{aligned} &t_0^{new} = t^{old},\ \ t_{k+1}^{new} = t_k^{new} - \eta_{t_{k+1}} \nabla_t F(t_k^{new}, \psi^{old}),\ k = 0, \dots, K-1,\ \ t^{new} = t_{K-1}^{new}, \\ &\psi^{new} = \psi^{old} + \eta_\psi \nabla F(t^{new}, \psi^{old}). \end{aligned} \tag{18}$$

How to choose K and η_t/η_ψ? We discuss this in the next section with further experimentation in Sec. E. Also note, if $\eta_\psi \ll K\eta_t$, the system behaves as $\min_t F(t, \psi)$ with slow varying ψ. The value $\kappa = K\eta_t/\eta_\psi$ characterizes the relative time scales of the transport map and potential updates.

5 Experimental Results and Discussion

5.1 Different formulations of NOT and connection with GANs

We classify static neural optimal transport (NOT) solvers into the following groups:

I. min–max optimization over the transport map T and the potential φ (or ψ);

II. min–max optimization over φ (or ψ) with the transport map defined as $T = \nabla\varphi$ (or $S = \nabla\psi$) Makkuva et al. (2020);

III. regularized (and possibly unbalanced) formulations. The effect of regularization and the bias it introduces require separate investigation.

In the case of quadratic cost, two equivalent formulations are commonly used:

1. Minimization of the squared distance;
2. Maximization of the dot product, also known as maximum crossentropy functional (MCF).

In the latter case, optimization is performed over convex functions Tarasov et al. (2026). All the results derived in this paper remain valid under this formulation, except for statement (iii) of Theorem 3 (see Sec. B).

Table 2: NOT solvers and related models

Group	Model	Objective
I	WGAN	$\inf_{G} \sup_{\varphi \in \mathrm{Lip}_1(Y)} \Big(\mathbb{E}_{y\sim\nu}[\varphi(y)] - \mathbb{E}_{x\sim\mu}[\varphi(G(x))] \Big)$
I	(p,q)-WGAN	$\sup_{\psi \in C(Y)} \inf_{G} \Big(\mathbb{E}_{x\sim\mu}\big[\frac{1}{p} d_q(x, G(x))^p\big] + \mathbb{E}_{y\sim\nu}[\psi(y)] - \mathbb{E}_{x\sim\mu}[\psi(G(x))] \Big)$
I	NOT	$\sup_{\psi \in C(Y)} \inf_{t} \Big(\mathbb{E}_{x\sim\mu}\big[c(x, t(x))\big] + \mathbb{E}_{y\sim\nu}[\psi(y)] - \mathbb{E}_{x\sim\mu}[\psi(t(x))] \Big)$
I	MCF	$\inf_{\substack{v \in C(Y) \\ v \text{ convex}}} \sup_{t} \Big(\mathbb{E}_{x\sim\mu}[\langle x, t(x)\rangle] + \mathbb{E}_{y\sim\nu}[v(y)] - \mathbb{E}_{x\sim\mu}[v(t(x))] \Big)$
I	MCF	$\inf_{\substack{u \in C(X) \\ u \text{ convex}}} \sup_{s} \Big(\mathbb{E}_{y\sim\nu}[\langle s(y), y\rangle] + \mathbb{E}_{x\sim\mu}[u(x)] - \mathbb{E}_{y\sim\nu}[u(s(y))] \Big)$
II	MCF	$\inf_{\substack{u \in C(X) \\ u \text{ convex}}} \sup_{\substack{v \in C(Y) \\ v \text{ convex}}} \Big(\mathbb{E}_{y\sim\nu}[\langle \nabla v(y), y\rangle] + \mathbb{E}_{x\sim\mu}[u(x)] - \mathbb{E}_{y\sim\nu}[u(\nabla v(y))] \Big)$

The connection between NOT and generative adversarial networks (GANs) can be understood through the existence of measurable transport maps. In particular, classical results originating from von Neumann (1938) imply that there exists a measurable function G mapping a simple distribution μ (e.g., uniform $U[0,1]$) to an arbitrary non-atomic distribution ν. In practice, classical GANs approximate such a map using neural networks and train it by minimizing the Jensen–Shannon divergence[10] between $G_\sharp\mu$ and ν :

$$\min_{G \in \mathcal{M}(X,Y)} \sup_{D\colon Y \to (0,1)} \left(\mathbb{E}_{y\sim\nu}[\log D(y)] + \mathbb{E}_{x\sim\mu}[\log(1 - D(G(x)))] \right). \tag{19}$$

WGAN replaces the divergence with the Kantorovich distance $W_1(\nu, G_\sharp\mu)$:

$$\min_{G \in \mathcal{M}(X,Y)} \sup_{\varphi \in \mathrm{Lip}_1(Y)} \left(\mathbb{E}_{y\sim\nu}[\varphi(y)] - \mathbb{E}_{x\sim\mu}[\varphi(G(x))] \right). \tag{20}$$

If $X, Y \subset \mathbb{R}^n$ have the same dimensionality, the optimal transport formulation can be used:

$$W_1(\mu, \nu) = \sup_{\psi \in C(Y)} \inf_{t} \left(\int_X |x - t(x)|\, d\mu - \int_X \psi(t(x))\, d\mu + \int_Y \psi(y)\, d\nu \right). \tag{21}$$

In this setting, minimizing the WGAN objective enforces the marginal condition $G_\sharp\mu = \nu$. The optimal transport map $T^\star$, when it exists, is one particular generator satisfying this condition. More generally, (q,p)-WGAN Mallasto et al. (2019) corresponds to the optimal transport problem with cost $c(x,y) = \frac{1}{p} d_q(x,y)^p$, where $d_q(x,y) = \left(\sum_{i=1}^n |x_i - y_i|^q\right)^{1/q}$.

We summarize the training objectives for comparison in Table 2. All these formulations share a common structural feature: they define degenerate saddle-point problems, where the discriminator (or potential) is not uniquely determined when the generator (or transport map) is close to optimal.

Finally, we note an important distinction: in GANs and WGANs, the generator G typically maps a low-dimensional latent distribution to a high-dimensional data space. While this guarantees the existence of such a map, it does not ensure that it coincides with an optimal transport map. In contrast, (q,p)-WGAN and NOT formulations operate on spaces of equal dimensionality, where the metric as a cost function is used.

5.2 Connection with Flow-Matching

Flow-matching approximates OT trajectories with an ODE (Lipman et al. (2022)). NOT can be viewed as the solution of ODE in one step. The continuous limit of the descrete optimiszation scheme

[10] Up to a constant, because of

$$JS(\nu || G_\sharp\mu) = \sup_{D\colon X \to [0,1]} \left(\int_X \log D(x) d\nu + \int_X \log(1 - D(x)) d(G_\sharp\mu) \right) + \log 2.$$

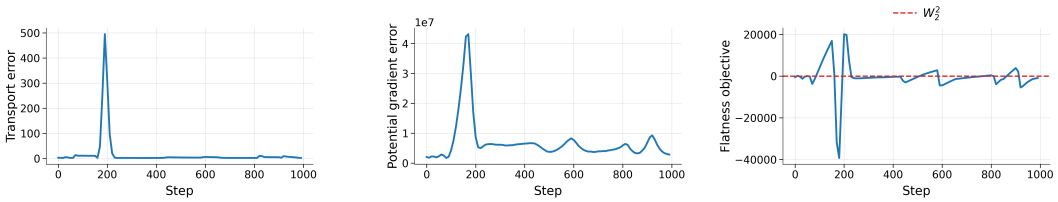


Figure 1: Convergence of transport map with divergence of potential.

is the following dynamic system

$$\begin{cases} \dfrac{\partial \rho_s}{\partial s} + \nabla(\rho_s v_s) = 0, \text{ where } v_s(y) = -(y - t_s^{-1}(y)) + \nabla\psi_\tau(y), \\ \dfrac{\partial \psi_\tau}{\partial \tau} = \kappa^{-1}(q(y) - \rho_s(y)). \end{cases} \tag{22}$$

The importance of two-timescale updates has been widely observed in GAN training, where multiple discriminator updates per generator step are required for stability Arjovsky et al. (2017); Gulrajani et al. (2017). This behavior was formalized in the two time-scale update rule Heusel et al. (2017), which shows that convergence depends on the relative step sizes of the two players. Similar empirical strategies are used in neural optimal transport Korotin et al. (2021), although an explicit characterization of the effective time-scale ratio is still lacking. From dynamic system (22) it follows that for large κ, t evolves faster. In the case of ψ, it evolves faster for small values of κ.

5.3 Numerical Results

Fig. 1 presents NOT results on two one-dimensional Gaussian distributions. The left plot shows the evolution of $\|t - T^\star\|^2_{L_2(\nu)}$, the middle plot shows $\|\nabla\psi - \nabla\psi^\star\|^2_{L_2(\nu)}$, and the right plot shows the values of $F(t, \psi)$. The figure illustrates that convergence of t_θ may occur while ψ_ϕ remains far from the optimal potential $\psi^\star$ (see Sec. D for details on the data used in the experiments).

We evaluate the performance of NOT in case of two 2D Gaussians with $K \in \{1,\ 5,\ 10,\ 15,\ 20\}$ and $\eta_\psi \in \{\eta_t,\ 0.1\eta_t,\ 0.01\eta_t\}$. The results in Fig. 2 demonstrate the importance of the value $K\eta_t/\eta_\psi$: for large K small η_t/η_ψ the transport map is approximated better, for small K and large η_t/η_ψ the potential is learned better. For experiments on more complex data see Sec. E. These examples illustrate that for the approximation of $T^\star$ the potential shouldn't be optimal.

5.4 Comparison with Previous Results

Different estimates exist in the literature on NOT, they assume λ-strong convexity for v like Tarasov et al. (2026), or even c-concavity of ψ like Fan et al. (2023). We relax ψ to uniformly Lipschitz functions and in contrast to previous work our estimate does not depend on ψ (the second term of $F(t, \psi)$ can be estimated by d_{KR}).

We emphasize that the existence of an infimum does not guarantee the existence of a minimizer (Table 1), a distinction that has led to incorrect claims in the literature, e.g., (Korotin et al., 2023, Lemma 1). We also caution against inferring optimality of t and ψ from the condition $F(t, \psi) \approx F(t^\star, \psi^\star)$: as illustrated in Fig. 1, this implication does not hold in general.

In Choi et al. (2025), spurious solutions are discussed. They rely on the convergence of the semi-dual problem, which is not guaranteed due to the degeneracy of the saddle-point (Theorem 3). They consider distributions concentrated on $(n-1)$-dimensional manifolds, when the Monge map dos not exist. Real data usually contain noise, or when their geometry is known, one can use OT on manifolds. We argue that spurious solutions in the semi-dual formulation arise from the degeneracy. Near optimal transport maps, the dual potential becomes non-identifiable: different potentials yield distinct gradients. As a result, optimization may converge to stationary points that do not correspond to $W_2(\mu, \nu)^2$, a phenomenon known as *gradient deviation* Korotin et al. (2021).

This viewpoint also explains the empirical observation of Makkuva et al. (2020)[Remark 3.4] that unconstrained neural parameterizations of the transport map (e.g., ∇v in Table 2) can outperform

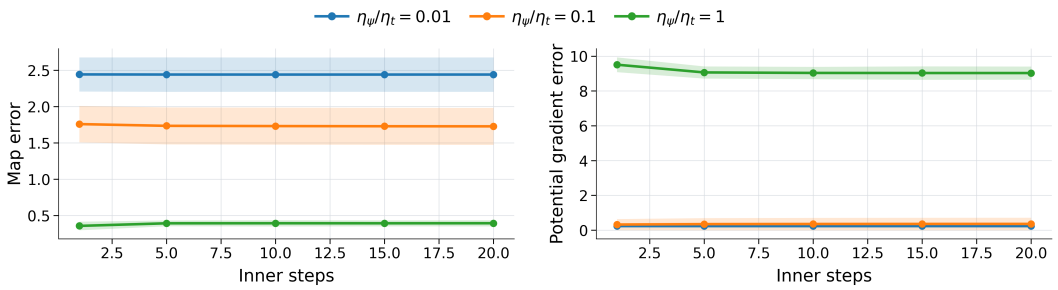


Figure 2: $\|t - T^\star\|^2_{L_2(\mu)}$ and $\|\nabla\psi - \nabla\psi^\star\|^2_{L_2(\nu)}$ with respect to K and η_ψ/η_t.

convex architectures. Imposing structural constraints on the potential, such as convexity or Lipschitz continuity, reduces degeneracy by eliminating flat directions. However, it also alters the optimization landscape by relaxing the exact dual constraints and may prevent perfect separation (12), thereby introducing approximation bias. To make the saddle-point strict, one can add a regularization term on ψ, see, e.g., Korotin et al. (2019), but it results in a different from OT generator.

5.5 Constrained Optimization Formulation and Connection with Unbalanced OT

Consider functions $f\colon L_2(X) \to \mathbb{R}$ and $g\colon L_2(X) \to \mathcal{M}(X)$ and the problem

$$f(t) = \int_X \|x - t(x)\|^2 d\mu \to \min_{t \in L_2(X)}, \qquad g(t) = t_\sharp\mu - \nu = 0. \tag{23}$$

Its Lagrangian coincides with F and one could expect the following minimax formulation

$$\inf_{t \in L_2(X)} \sup_{\psi \in C(X)} F(\psi, t). \tag{24}$$

In this Monge formulation, $g(t) = 0$ is the constrain and ψ is the dual variable, which can be arbitrary, when the constrain is satisfied. The Lagrangian is not convex, and the theory of saddle-point can not be applied (in particular, interchange of infimum and supremum). We consider constrain in week sense, the case of absolutely continuous densities was studied in Evans (1999).

Even though problems (11) and (24) are not equivalent, in the alternating scheme the order of $\sup$ and $\inf$ is interchangeable. Therefore, NOT can be viewed as a constrained optimization problem that makes it similar to unbalanced OT, where the constraint $t_\sharp\mu = \nu$ is relaxed via penalization.

5.6 Limitations

Our analysis relies on the existence of an optimal Monge map (see Table 1). In Theorem 4, we further assume that ψ is uniformly Lipschitz, which is reasonable in practice—for instance, when neural networks are regularized via gradient clipping or spectral normalization.

Stronger assumptions are required to characterize the second variation of $F(t, \mu)$ because of its degeneracy. In particular, convexity of the supports of μ and ν, together with their absolute continuity, plays a crucial role, as demonstrated in Figalli et al. (2013).

Finally, finite-sample approximations impose intrinsic limitations: lower bounds on the accuracy of empirical optimal transport estimators are known (see (Hütter & Rigollet, 2021, Theorem 2)).

AI use statement

In this work, we have not used generative AI tools for any of the required-disclosure tasks: generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology

or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results.

Additionally, we used generative AI tools (ChatGPT) to edit the manuscript for readability, and to help identify relevant prior work (Claude), format citations, and suggest keywords for the submission. We have reviewed all AI-assisted work: all edited text was checked against the original meaning and technical accuracy by the authors, and all AI-suggested literature references were independently verified against the original sources before citing. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

An anonymized version of our full codebase, configuration files, and trained model weights is included as supplementary material, enabling exact reproduction of all experiments (Sec. C); this will be released publicly under the MIT license upon acceptance. Section D gives complete descriptions of the data-generation procedures used to produce Fig. 1 and Fig. 2, and of the ICNN-based benchmark construction used in our additional experiments (Sec. D), including all distributional parameters, network architectures, and training hyperparameters needed to regenerate these datasets from scratch. Complete proofs of Theorems 3 and 4 and Remark 2 are given in Appendix A.

References

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 2017. URL https://arxiv.org/abs/1701.07875.

Vladimir I. Bogachev and Aleksandr V. Kolesnikov. The Monge–Kantorovich problem: achievements, connections, and perspectives. *Russian Math. Surveys*, 67(5):785–890, 2012.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44:375–417, 1991.

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. *NerISP*, 2025.

Nhan Dam, Quan Hoang, Trung Le, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Three-player wasserstein gan via amortised duality. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 2202–2208, 2019. doi: 10.24963/ijcai.2019/305. URL https://doi.org/10.24963/ijcai.2019/305.

Ivar Ekeland. An optimal matching problem. *ESAIM: COCV*, 11(1):57–71, 2005.

Lawrence C. Evans. Partial differential equations and Monge–Kantorovich mass transfer. In *Current developments in mathematics. Papers from the conference held in Cambridge, MA, USA, 1997*, pp. 65–126. Boston, MA: International Press, 1999.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

Alessio Figalli, Young-Heon Kim, and Robert J. McCann. Hölder continuity and injectivity of optimal maps. *Archive for Rational Mechanics and Analysis*, 209(3):747–795, 2013.

Wilfrid Gangbo and Robert McCann. The geometry of optimal transportation. *Acta Math.*, 177(2): 113–161, 1996.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 3440–3448, 2016. URL https://arxiv.org/abs/1605.08527.

Nicola Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, 2011. doi: 10.1017/S001309150800117X.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 5767–5777, 2017. URL https://arxiv.org/abs/1704.00028.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *Ann. Statist.*, 49(2):1166–1194, 2021.

Leonid V. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, 37(7-8):227–229, 1942. doi: 10.1007/s10958-006-0049-2. English Translation in J. Math. Sci. (N. Y.), 133(4) (2006), 1381–1382.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark. In *NeurIPS*, 2021.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *Proc. ICLR*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Le Matt. Flow matching for generative modeling. *https://arxiv.org/abs/2210.02747*, 2022.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. *ICML*, 2020.

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q,p)-Wasserstein GANs: comparing ground metrics for Wasserstein GANs. *arXiv preprint arXiv:1902.03642*, 2019.

Robert McCann. Existence and uniqueness of monotone measure-preserving map. *Duke Mathematical Journal*, 80(2), 1995.

Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc.*, 51(4):527–580, 2014. doi: https://doi.org/10.1090/S0273-0979-2014-01459-4.

Aldo Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 43(1): 1–13, 2007.

R. Tyrrell Rockafellar. Integral functionals, normal integrands and measurable selections. In J.-P. Gossez, E. J. Lami Dozo, J. Mawhin, and L. Waelbroeck (eds.), *Nonlinear Operators and the Calculus of Variations*, volume 543 of *Lecture Notes in Mathematics*, pp. 157–207. Springer, Berlin, Heidelberg, 1976. doi: 10.1007/BFb0087260.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 2009.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. *arXiv preprint arXiv:2110.02999*, 2021.

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

Roman Tarasov, Petr Mokrov, Milena Gazdieva, Evgeny Burnaev, and Alexander Korotin. A statistical learning perspective on semi-dual adversarial neural optimal transport solvers. In *Proc. ICLR*, 2026. URL https://arxiv.org/abs/2502.01310v2.

Adrien Vacher and François-Xavier Vialard. Parameter tuning and model selection in optimal transport with semi-dual Brenier formulation. *arXiv:2112.07275*, 2021.

Leonid N. Vaserstein. Markov processes over denumerable products of spaces describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Anatoly M. Vershik. Long history of the Monge–Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013. doi: 10.1007/s00283-013-9380-x.

Cédric Villani. *Optimal Transport. Old and New*. Springer, Berlin, 2009.

John von Neumann. On rings of operators. Reduction theory. *Annals of Mathematics*, 50(2):401–485, 1938.

A Proofs of Theorems 3 and 4 and Remark 2

Proof of Theorem 3. **Step 1: Rewriting the functional.** We write

$$F(\psi, t) = \int_X \big[c(x, t(x)) - \psi(t(x))\big]\, d\mu + \int_Y \psi(y)\, d\nu.$$

Step 2: Minimization over t. Fix ψ. Set

$$f(x, y) := c(x, y) - \psi(y).$$

Then

$$\inf_t \int_X f(x, t(x))\, d\mu.$$

Since the problem is separable in x, by a measurable selection argument (e.g., Rockafellar–Wets, Theorem 14.60),

$$\inf_t \int_X f(x, t(x))\, d\mu = \int_X \inf_{y \in Y} f(x, y)\, d\mu = \int_X \psi^c(x)\, d\mu.$$

Therefore,

$$\inf_t F(\psi, t) = \int_X \psi^c(x)\, d\mu + \int_Y \psi(y)\, d\nu.$$

Step 3: Proof of (i). Taking the supremum over ψ, we obtain

$$\sup_\psi \inf_t F(\psi, t) = \sup_\psi \left(\int_X \psi^c\, d\mu + \int_Y \psi\, d\nu \right),$$

which is exactly the Kantorovich dual problem.

Step 4: Proof of (ii). Using Step 2,

$$\inf_t F(\psi, t) = \int_X \psi^c\, d\mu + \int_Y \psi\, d\nu,$$

$$\inf_t F(\psi^{cc}, t) = \int_X (\psi^{cc})^c\, d\mu + \int_Y \psi^{cc}\, d\nu.$$

Since $(\psi^{cc})^c = \psi^c$, we obtain

$$\inf_t F(\psi^{cc}, t) - \inf_t F(\psi, t) = \int_Y (\psi^{cc} - \psi)\, d\nu \geq 0.$$

Equality holds if and only if $\psi = \psi^{cc}$ ν-a.e., i.e., ψ is c-concave.

Step 5: Proof of (iii). Suppose $t_{\#}\mu \neq \nu$. Then the signed measure

$$\sigma := \nu - t_{\#}\mu$$

is a nonzero signed Radon measure on the compact metric space Y, the Riesz representation theorem guarantees the existence of a continuous function $\tilde{\psi} \in C(Y)$ such that $\int_Y \tilde{\psi}\, d\sigma > 0$. Let $\psi_M = M\tilde{\psi}$, then

$$F(\psi_M, t) = \int c(x, t(x))\, d\mu + M \int \tilde{\psi}\, d\sigma,$$

which tends to $+\infty$ as $M \to \infty$, we obtain

$$\sup_{\psi} F(\psi, t) = +\infty.$$

Step 6: Proof of (iv). Let $\psi^\star$ be an optimal Kantorovich potential. Then $\psi^\star$ is c-concave, and there exists an optimal transport map $T^\star$ such that

$$T^\star(x) \in \arg\min_y \{c(x, y) - \psi^\star(y)\} \quad \mu\text{-a.e.}$$

Hence,

$$c(x, T^\star(x)) - \psi^\star(T^\star(x)) = \psi^{\star c}(x),$$

and therefore

$$F(\psi^\star, T^\star) = \int_X \psi^{\star c}(x)\, d\mu + \int_Y \psi^\star(y)\, d\nu.$$

For any t, using $\psi^{\star c}(x) \le c(x, t(x)) - \psi^\star(t(x))$, we obtain

$$F(\psi^\star, t) \ge F(\psi^\star, T^\star).$$

If $t = T^\star$, then $t_{\#}\mu = \nu$, and thus

$$F(\psi, T^\star) = \int c(x, T^\star(x))\, d\mu,$$

which is independent of ψ. Hence

$$F(\psi^\star, T^\star) = F(\psi, T^\star) \quad \text{for all } \psi.$$

Formula (14) follows from Gangbo & McCann (1996).

Step 7: Proof of (v). Fix $\psi^\star$. Then minimizing $t \mapsto F(\psi^\star, t)$ is equivalent to minimizing

$$\int_X \big[c(x, t(x)) - \psi^\star(t(x))\big]\, d\mu.$$

By Step 2, any minimizer satisfies

$$t(x) \in \arg\min_y \{c(x, y) - \psi^\star(y)\} \quad \mu\text{-a.e.}$$

For the quadratic cost, this minimizer is unique μ-almost everywhere (by Brenier's theorem), hence $t = T^\star$ μ-a.e. and the minimizer is unique.

□

Proof of Remark 2. **Step 1.** Use the pointwise inequality

$$\begin{aligned} |c(x,\, t_k(x)) - c(x,\, T^\star(x))| &= \left| \frac{1}{2}|x - t_k(x)|^2 - \frac{1}{2}|x - T^\star(x)| \right| \\ &= \frac{1}{2}\, |\langle t_k(x) - T^\star(x),\, t_k(x) + T^\star(x) - 2x\rangle| \\ &\le \frac{1}{2}|t_k(x) - T^\star(x)|\, |t_k(x) + T^\star(x) - 2x|. \end{aligned} \tag{25}$$

Since X and Y are bounded, there exists $C > 0$ such that $|t_k(x) + T^\star(x) - 2x| \le 2C$ for every $x \in X$. Hence, by the Cauchy–Schwarz inequality,

$$\left| \int c(x,\, t_k(x))\, d\mu(x) - W_2(\mu,\, \nu)^2 \right| \le C \int |t_k - T^\star|\, d\mu(x) \le C \|t_k - T^\star\|_{L^2(\mu)}, \tag{26}$$

where we used the fact that μ is a probability measure.

Step 2. Since $T^\star_\sharp \mu = \nu$, consider the coupling $\pi_k = (t_k,\, T^\star)_\sharp \mu$. By the definition,

$$W_2(t_{k\sharp}\mu,\, \nu)^2 \le \int |t_k(x) - T^\star(x)|\, d\mu(x) = \|t_k - T^\star\|^2_{L^2(\mu)}. \tag{27}$$

Therefore,

$$d_{KR}(t_{k\sharp}\mu,\, \nu) \le W_1(t_{k\sharp}\mu,\, \nu) \le W_2(t_{k\sharp}\mu,\, \nu) \le \|t_k - T^\star\|_{L^2(\mu)}. \tag{28}$$

Step 3. Set $g_k = \psi_k - \psi^\star$ and $c_k = (\max g_k + \min g_k)/2$. Then

$$\begin{aligned} \left| \int (\psi_k - \psi^\star)\, d(t_{k\sharp}\mu - \nu) \right| &= \left| \int (g_k - c_k)\, d(t_{k\sharp}\mu - \nu) \right| \\ &= M \left| \int \frac{g_k - c_k}{M}\, d(t_{k\sharp}\mu - \nu) \right| \le M d_{KR}(t_{k\sharp},\, \nu), \end{aligned} \tag{29}$$

where $M = \max(2L',\, \mathrm{diam}(Y)L')$. By L' we denote a common Lipschitz constant for ψ_k and $\psi^\star$.

Step 4. From

$$F(\psi_k,\, t_k) - W_2(\mu,\, \nu)^2 = \int c(x,\, t_k(x))\, d\mu - W_2(\mu,\, \nu)^2 + \int (\psi_k - \psi^\star)\, d(t_{k\sharp}\mu - \nu) \tag{30}$$

we have

$$\begin{aligned} \left|F(\psi_k,\, t_k) - W_2(\mu,\, \nu)^2\right| &\le \left| \int c(x,\, t_k(x))\, d\mu - W_2(\mu,\, \nu)^2 \right| + \left| \int (\psi_k - \psi^\star)\, d(t_{k\sharp}\mu - \nu) \right| \\ &\le (C + M)\|t_k - T^\star\|_{L^2(\mu)}. \end{aligned} \tag{31}$$

Altogether, this gives that $F(\psi_k,\, t_k) \to W_2(\mu,\, \nu)^2$ in $L^2(\mu)$ and $d_{KR}(t_{k\sharp}\mu,\, \nu) \to 0$. □

Proof of Theorem 4. First, note that $\psi^\star \in \mathrm{Lip}_L$ for some $L > 0$, because it is differentiable. Therefore, class $\psi \in C(Y)$ can be narrowed to $\psi \in \mathrm{Lip}_L$. Recall that $T^\star(x) = \nabla u(x)$ and $S^\star(y) = \nabla v(y)$, where

$$u(x) = \frac{|x|^2}{2} - \varphi^\star(x), \qquad v(y) = \frac{|y|^2}{2} - \psi^\star(y). \tag{32}$$

Step 1. (Philippis & Figalli, 2014, Theorem 3.3) guaranties $u \in C^2(\overline{X})$ that gives the estimate $\|u\|_\infty < C$ for some $C > 0$. Then, from the boundedness of $\det D^2 u$ away from zero and infinity, the positive definetness of the Hessian follows $D^2 v \ge \lambda I$ (see also (Gigli, 2011, Corollary 3.2)).[11] Therefore, v is λ-strongly convex, i.e.,

$$v(y_2) \ge v(y_1) + \langle \nabla v(y_1), y_2 - y_1 \rangle + \frac{\lambda}{2}|y_2 - y_1|^2, \qquad \forall y_1, y_2 \in Y. \tag{33}$$

Step 2.

$$\begin{aligned} F(t, \psi^\star) - F(T^\star, \psi^\star) =& \frac{1}{2}\int |x - t(x)|^2 d\mu - \int \psi^\star(t(x)) d\mu + \int \psi^\star(y) d\nu \\ & - \frac{1}{2}\int |x - T^\star(x)|^2 d\mu + \int \psi^\star(T^\star(x)) d\mu - \int \psi^\star(y) d\nu \\ =& \int \left[-\langle x, t - T^\star \rangle + v(t(x)) - v(T^\star(x)) \right] d\mu \\ \overset{(33)}{\ge}& \int \left[-\langle x, t - T^\star \rangle + \langle \nabla v(T^\star(x)), t - T^\star \rangle + \frac{\lambda}{2}|t - T^\star(x)|^2 \right] d\mu \\ =& \frac{\lambda}{2}\|t - T^\star\|^2_{L^2(\mu)}. \end{aligned} \tag{34}$$

[11] They prove $\lambda I \le D^2 u \le \Lambda I$, we note that $D^2 v = (D^2 u)^{-1}$ and redenote $1/\Lambda$ by λ.

We uses the fact that $\nabla v(T^\star(x)) = x$.

Step 3. We need the following estimate:

$$\int_Y (\psi^\star - \psi) d(t_\sharp \mu - \nu) \leq M d_{KR}(t_\sharp \mu, \nu) \tag{35}$$

with $M = \max(2L',\, L' \cdot \mathrm{diam}(Y))$, where $L' = \max\{L,\, \|\psi^\star\|_{Lip}\}$ is a common Lipschitz constant. For each ψ define $g = \psi^\star - \psi$. Since $\psi, \psi^\star$ are L'-Lipschitz, g is $2L'$-Lipschitz. Since $t_\sharp \mu - \nu$ has total mass zero,

$$\int (g - c) d(t_{k\sharp} \mu - \nu) = \int g d(t_{k\sharp} \mu - \nu) \tag{36}$$

for any constant c. We take $c = \frac{1}{2}(\max g + \min g)$. With this choice,

$$\|g - c\|_\infty \leq \tfrac{1}{2}(\max g - \min g) \leq \tfrac{1}{2} \cdot 2L' \cdot \mathrm{diam}(Y) = L' \cdot \mathrm{diam}(Y). \tag{37}$$

The function $(g - c)/M$ satisfies both $|(g - c)/M| \leq 1$ and is 1-Lipschitz, so it is admissible in the supremum defining d_{KR}. Hence

$$\int_Y (\psi^\star - \psi) d(t_\sharp \mu - \nu) = M \int \frac{g - c}{M} d(t_\sharp \mu - \nu) \leq M d_{KR}(t_\sharp \mu, \nu). \tag{38}$$

Step 4. Notice that

$$F(t, \psi) = F(t, \psi^\star) + \int (\psi - \psi^\star) d(t_\sharp \mu - \nu), \tag{39}$$

then

$$\begin{aligned} \frac{\lambda}{2} \|t - T^\star\|^2_{L^2(\mu)} &\overset{(34)}{\leq} F(t, \psi^\star) - F(T^\star, \psi^\star) \\ &\overset{(39)}{=} F(t, \psi) - F(T^\star, \psi^\star) - \int (\psi - \psi^\star) d(t_\sharp \mu - \nu) \\ &\overset{(35)}{\leq} F(t, \psi) - W_2(\mu, \nu)^2 + M d_{KR}(t_\sharp \mu, \nu). \end{aligned} \tag{40}$$

Because $\lambda > 0$ and does not depend on t or ψ, we can divide the inequality by $\lambda/2$ to obtain the final result. □

B Alternative Formulation of NOT in Case of Quadratic Cost

If we look at the Kantorovich problem in case of quadratic cost

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \left(\frac{|x|^2}{2} + \frac{|y|^2}{2} - \langle x, y \rangle \right) d\pi = \int_X \frac{\|x\|^2}{2} d\mu + \int_Y \frac{\|y\|^2}{2} d\nu - \sup_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} \langle x, y \rangle d\pi \tag{41}$$

we can connect the following maximum correlation functional to the Monge formulation:

$$\sup_{t \in T(\mu,\nu)} \int_X \langle x, t(x) \rangle d\mu. \tag{42}$$

Therefore, by duality,

$$\int_X \frac{\|x\|^2}{2} d\mu + \int_Y \frac{\|y\|^2}{2} d\nu - \sup_{t \in T(\mu,\nu)} \int_X (x, t(x)) d\mu = \sup_{\varphi(x) + \psi(y) \leq \|x - y\|^2} \left(\int_X \varphi d\mu + \int_Y \psi d\nu \right), \tag{43}$$

or

$$MCF(\mu, \nu) = \inf_{u(x) + v(y) \geq \langle x, y \rangle} \left(\int_X u d\mu + \int_Y v d\nu \right). \tag{44}$$

where $u(x) = \frac{|w|^2}{2} - \phi(x)$ and $v(y) = \frac{|y|^2}{2} - \psi(y)$ are convex functions.

Using semi-duality, we have

$$MCF(\mu,\nu) = \inf_{\substack{u\in C(Y)\\ u \text{ convex}}} \left(\int_X u(x)\, d\mu + \int_Y \sup_{x\in X} (\langle x, y\rangle - u(x))\, d\nu \right) \tag{45}$$

$$= \inf_{\substack{v\in C(Y)\\ v \text{ convex}}} \left(\int_X \sup_{y\in Y} (\langle x, y\rangle - v(x))\, d\mu + \int_Y g\, d\nu \right). \tag{46}$$

And using interchange theorem,

$$MCF(\mu,\nu) = \inf_{\substack{u\in C(X)\\ u \text{ convex}}} \sup_{s\in\mathcal{M}(Y,X)} \Big(\int_X u(x) d\mu + \int_Y \big[\langle s, y\rangle - u(s(y))\big] d\nu \Big) \tag{47}$$

$$= \inf_{\substack{v\in C(Y)\\ v \text{ convex}}} \sup_{t\in\mathcal{M}(Y,X)} \Big(\int_X \big[\langle x, t\rangle - v(t(x))\big] d\mu + \int_Y v(y) d\nu \Big). \tag{48}$$

Note that

$$SDP(\mu,\nu) = \int_X \frac{\|x\|^2}{2} d\mu + \int_Y \frac{\|y\|^2}{2} d\nu - MCF(\mu,\nu). \tag{49}$$

Consider the functional

$$\tilde{F}(t, v) = \int_X \langle x, t(x)\rangle d\mu + \int_Y v(y) d(\nu - t_\sharp \mu). \tag{50}$$

The same analysis can be applied as to $F(t, \psi)$.

C Code Availability

Upon acceptance, we will release the full codebase, trained weights, and generated data under the MIT license on our GitHub repository and project page. Due to double-blind anonymity constraints, we do not include the public repository link at submission time; instead, the submitted supplementary material contains the code needed to reproduce this work.

D Data Generation

Figures in Sec. 5 To obtain Fig. 1, we consider one-dimensional Gaussian distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(2, 1.5^2)$. Both t and ψ are parameterized by MLPs and trained until convergence. We then perturb ψ with noise and continue training for an additional 1000 steps. For Fig. 2, we consider two-dimensional Gaussians $\mathcal{N}(0, I)$ and $\mathcal{N}(\mathbf{1}, 2^2 I)$, and run NOT with 30 different random seeds. The corresponding models and trained weights are available on the project page.

D.1 Additional Experiments

We use an ICNN-generated continuous OT benchmark, following the analytic-map benchmark idea of Korotin et al. (2021). The source distribution $\mu \subset \mathbb{R}^{192}$ is a 24-component Gaussian mixture:

$$X \sim \sum_{j=1}^{24} \pi_j \mathcal{N}\big(m_j,\ 0.35^2 \operatorname{diag}(s_j^2)\big)\,. \tag{51}$$

The component means m_j are random directions normalized to radius 2.5. Mixture weights are unequal, with $\pi = \operatorname{softmax}(\ell)$ and $\ell_j \sim \mathcal{N}(0, 1.25^2)$. Each component has anisotropic diagonal scale s_j, where $\log s_{jk}$ is sampled with standard deviation 0.9 and centered within each component.

The ground-truth Brenier potential is a randomly initialized input-convex neural network $\Phi_0 : \mathbb{R}^{192} \to \mathbb{R}$ with five hidden layers of width 192, softplus activations, initialization standard deviation 0.14, and strong-convexity coefficient 0.05. We calibrate its scale on 4096 source samples:

$$\Phi = a\Phi_0, \qquad a = \frac{\sqrt{\mathbb{E}\|X\|^2}}{\sqrt{\mathbb{E}\|\nabla\Phi_0(X)\|^2}}. \tag{52}$$

The target distribution is then defined by the exact pushforward:

$$T^\star(x) = \nabla\Phi(x), \qquad \nu = T^\star_\# \mu. \tag{53}$$

Training batches are resampled every epoch with batch size 512 and 128 batches per epoch; validation and test splits contain 8192 samples each. In the timescale sweep we use $K \in \{1, 2, 5, 10, 20\}$, ratios $\eta_\psi/\eta_T \in \{0.02, 0.05, 0.1, 0.25, 0.5, 1\}$, $\eta_T = 5 \cdot 10^{-4}$, 4096 outer iterations, and disable solver noise.

Target-potential diagnostics. Because the dataset is generated by a known convex potential Φ, we also know the canonical target-side potential on paired samples. For $y = T^\star(x) = \nabla\Phi(x)$, the convex conjugate satisfies[12]

$$\Phi^*(y) = \langle x, y\rangle - \Phi(x), \qquad \nabla\Phi^*(y) = x. \tag{54}$$

Thus, for max-correlation-style methods we compare the learned target potential v to $v^\star = \Phi^*$. For the semi-dual quadratic-cost potential used by OTP, the corresponding c-concave potential is

$$\psi^\star(y) = \frac{1}{2}\|y\|^2 - \Phi^*(y), \qquad \nabla\psi^\star(y) = y - x. \tag{55}$$

Potential values are compared only after centering, since potentials are identified up to an additive constant. We report centered target-potential MSE and target-potential gradient MSE on 512 validation pairs.

Computation of d_{KR}. We approximate $d_{KR}(T_\#\mu, \nu) \approx W_1(T_\#\mu, \nu)$ using empirical measures $\hat{\mu}_T = \frac{1}{n}\sum_{i=1}^n \delta_{T(x_i)}$ and $\hat{\nu} = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$:

$$\widehat{d}_{KR} = \min_{\pi \in S_n} \frac{1}{n}\sum_{i=1}^n \left\|T(x_i) - y_{\pi(i)}\right\|_2, \tag{56}$$

where the optimal permutation π is computed by linear assignment, with $n \le 1024$.

E Experiments

E.1 OTP

OTP is the direct-map semi-dual NOT solver: an MLP map T_θ is trained against a c-concave DenseICNN target potential ψ_ϕ using the quadratic semi-dual objective (Korotin et al., 2023; Rout et al., 2021). In the sweep, K is the number of map updates per potential update. The results are shown in Figure 3.

E.2 Monge Map

MongeMap follows the weak-form neural Monge-map formulation of Fan et al. (2023). It uses a direct MLP map and an unconstrained MLP potential, trained with the same alternating map/potential schedule. The results are shown in Figure 4.

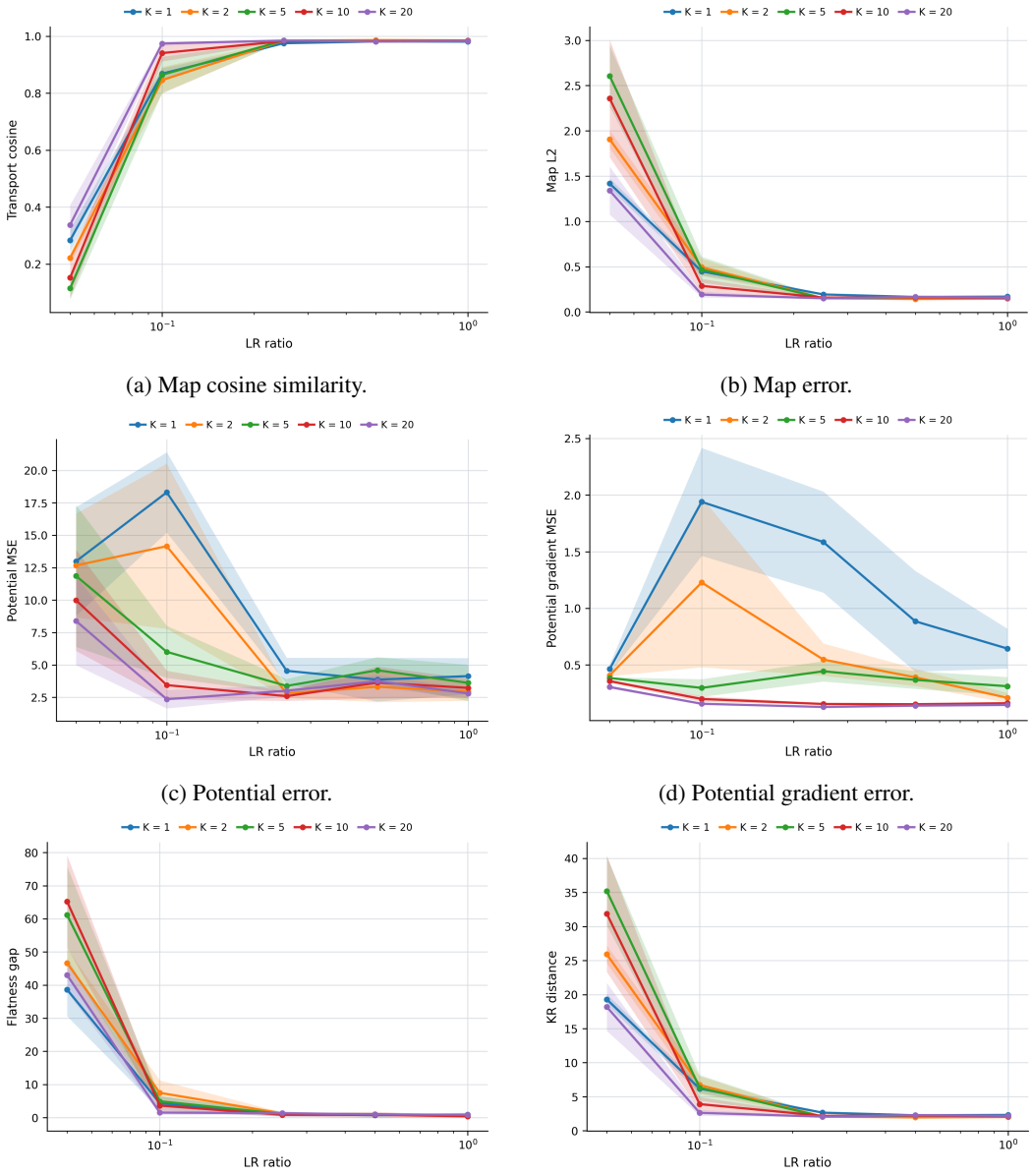
E.3 Max Correlation

MaxCorr uses the maximum-correlation form of quadratic OT:

$$\langle x, T(x)\rangle - g(T(x)) + \mathbb{E}_{y\sim\nu} g(y). \tag{57}$$

It uses a direct MLP map and DenseICNN target potential, and is the unregularized max-correlation ablation of the OTM-style objective (Rout et al., 2021; Tarasov et al., 2026). The results are shown in Figure 5.

[12]To avoid notational ambiguity, we use $*$ to denote the Legendre–Fenchel transform, while $\star$ denotes the optimal transport map and potential.



(a) Map cosine similarity.

(b) Map error.

(c) Potential error.

(d) Potential gradient error.

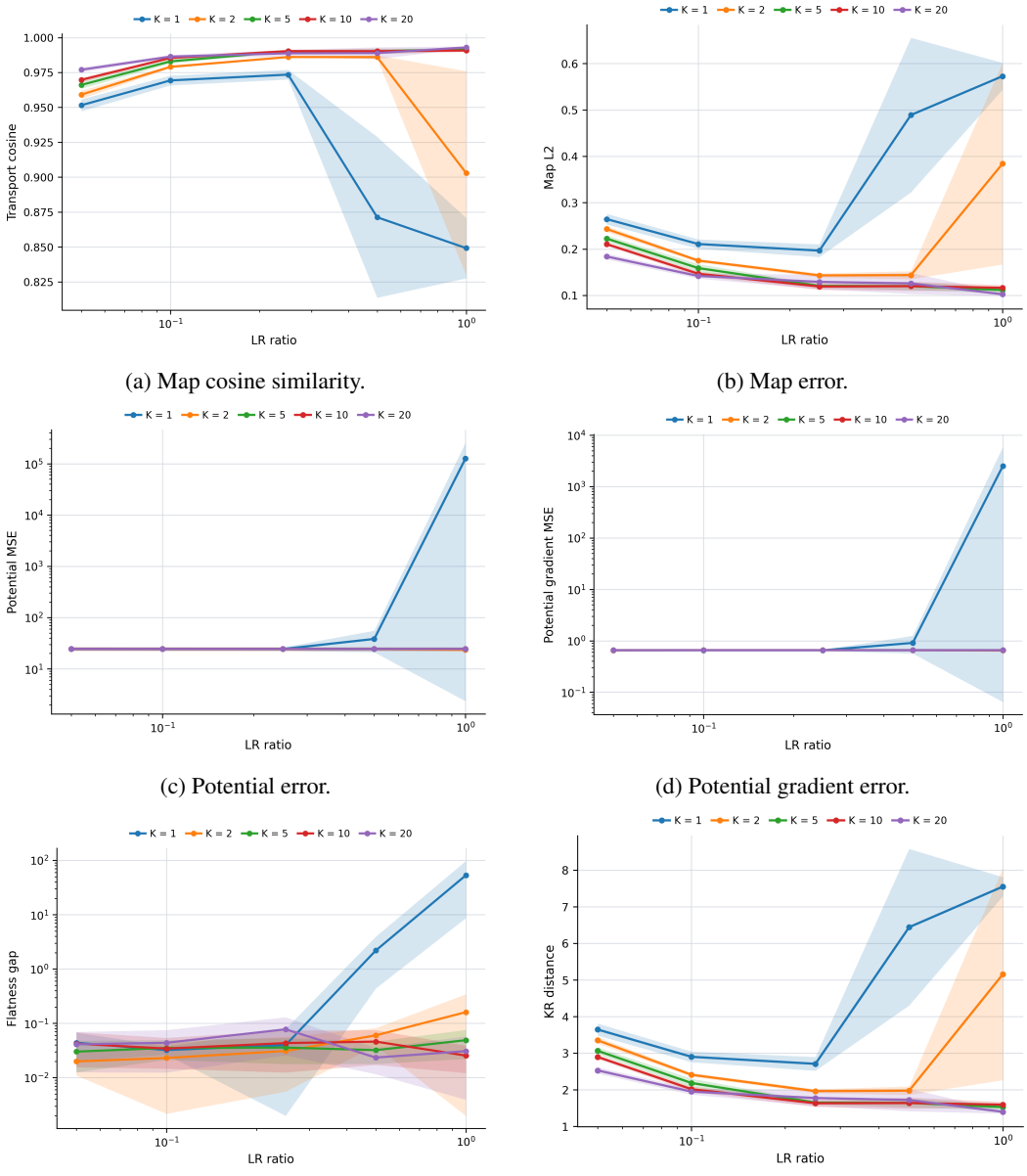
(e) Flatness.

(f) d_{KR}.

Figure 3: Convergence behavior of the transport map and potential for the OTP method.

E.4 OTM

OTM uses the same max-correlation objective, but adds the published gradient-optimality penalty from optimal transport modeling (Rout et al., 2021). In our setup this penalty has weight 0.1. The results are shown in Figure 6.



(a) Map cosine similarity.

(b) Map error.

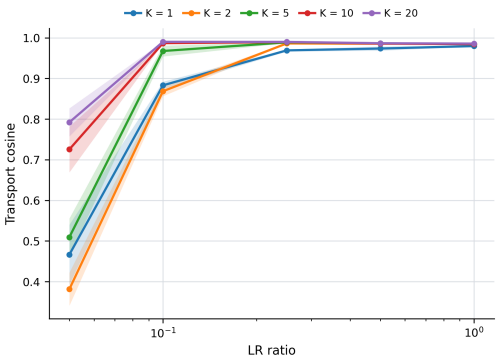
(c) Potential error.

(d) Potential gradient error.

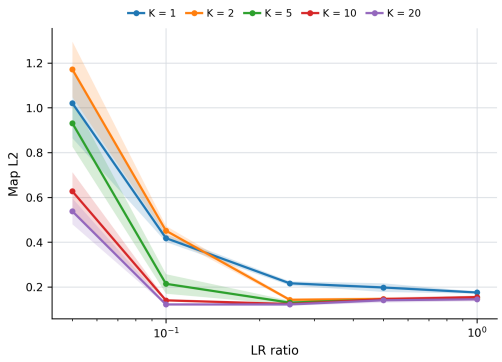
(e) Flatness.

(f) d_{KR}.

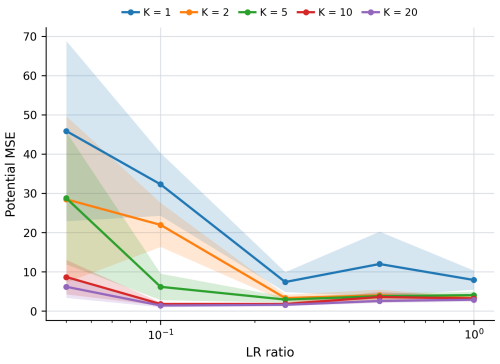
Figure 4: Convergence behavior of the transport map and potential for the Monge Map method.



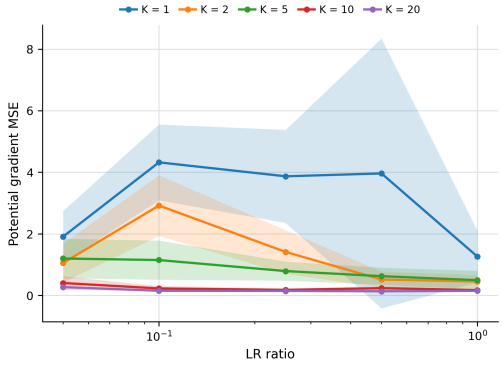
(a) Map cosine similarity.



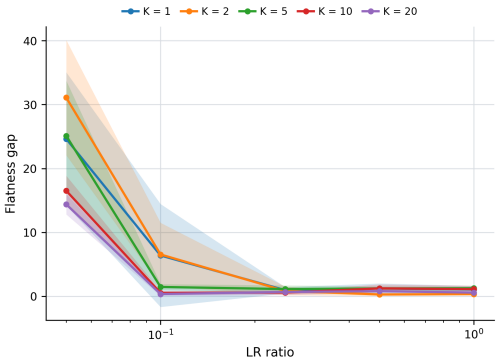
(b) Map error.



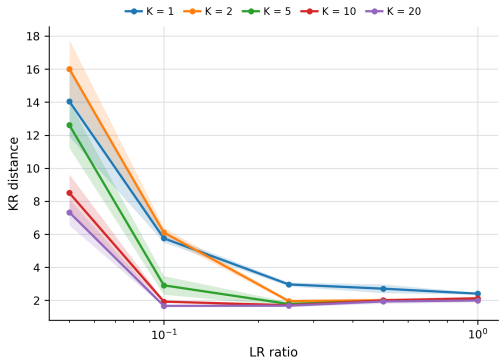
(c) Potential error.



(d) Potential gradient error.

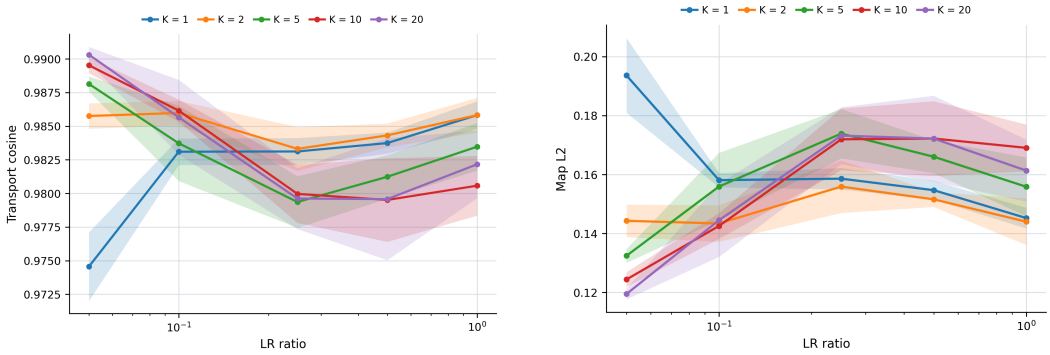


(e) Flatness.



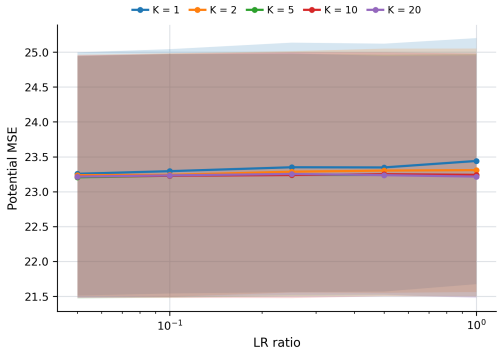
(f) d_{KR}.

Figure 5: Convergence behavior of the transport map and potential for the Max Correlation method.

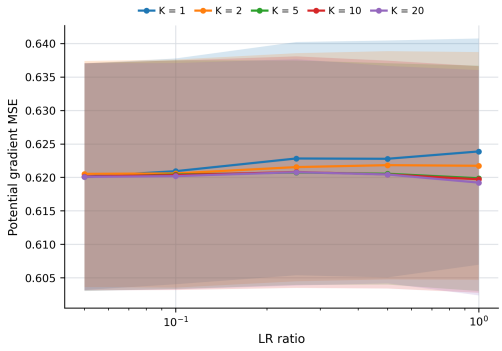


(a) Map cosine similarity.

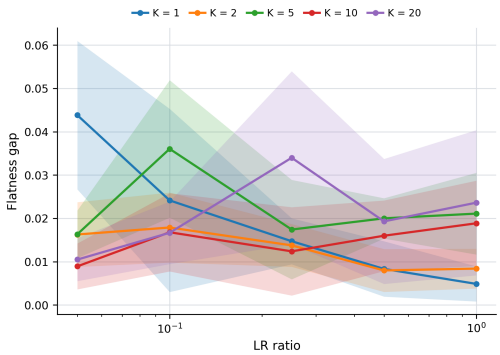
(b) Map error.



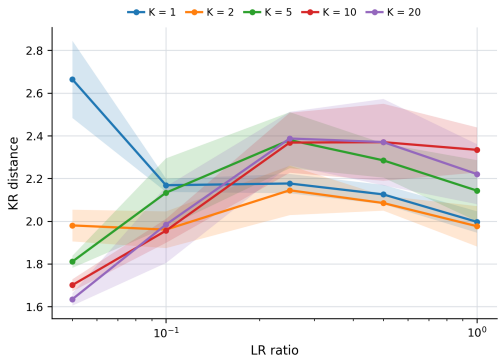
(c) Potential error.



(d) Potential gradient error.



(e) Flatness.



(f) d_{KR}.

Figure 6: Convergence behavior of the transport map and potential for the OTM method.