
Stability and Oracle Inequalities for Optimal Transport Maps between General Distributions

Shubo Li

Department of Statistics
The Pennsylvania State University
State College, PA 16803
sk16034@psu.edu

Yizhe Ding

Department of Statistics
The Pennsylvania State University
State College, PA 16803
yvd5212@psu.edu

Lingzhou Xue

Department of Statistics
The Pennsylvania State University
lxue@psu.edu

Runze Li

Department of Statistics
The Pennsylvania State University
rzli@psu.edu

Abstract

Optimal transport (OT) provides a powerful framework for comparing and transforming probability distributions, with wide applications in generative modeling, AI4Science and statistical inference. However, existing estimation theory typically requires stringent smoothness conditions on the underlying Brenier potentials and assumes bounded distribution supports, limiting practical applicability. In this paper, we introduce a unified theoretical framework for semi-dual OT map estimation that relaxes both of these restrictions. Building on sieved convex conjugate, our framework has two key contributions: (i) a new map stability bounds that holds without any second-order regularity assumptions on the true Brenier potentials, and (ii) an oracle inequality that cleanly decomposes the estimation error into statistical error, sieved bias, and approximation error. Specifically, our approximation error is measured in the L^∞ norm rather than Sobolev norm in the existing results, aligning more naturally with classical approximation theory. Leveraging these tools, we provide statistical error of semi-dual estimators with mild and verifiable conditions on the true OT map. Moreover, we establish the first theoretical guarantee for deep neural network OT map estimator between general distributions, with Tanh network function class as an example.

1 Introduction

Optimal Transport (OT) provides a powerful framework for transforming one probability measure into another. Concretely, given two distributions μ and ν on $\mathbb{R}^d$, the OT problem seeks a map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that transports μ to ν while minimizing a cost function. Beyond its elegant mathematical foundation, OT has found broad applications in AI4Science [39, 7, 29], computer vision [3, 23, 43, 12], and nonparametric statistical inference [10, 26, 24, 18], powering advances in scientific discovery and machine/deep learning. Despite its broad impact, rigorous theoretical guarantees for OT estimators under mild and realistic regularity conditions remain limited.

This paper aims to contribute to this direction by focusing on the dual-type OT map estimators, formally introduced in Section 2, which are based on the characterization of the optimal transport map as the gradient of a convex function (the so-called Brenier potential). The seminal work of [27] established minimax optimality for a wavelet-truncated dual-type OT estimator, but under restrictive conditions: μ is supported on $[0, 1]^d$ with a bounded density, ν is compactly supported, and the true

OT map is both α -Hölder smooth and strongly convex. These assumptions exclude many practical OT applications.

In parallel, Gunsilius [25] introduced an alternative analysis using the Poincaré inequality, a fundamental tool in PDEs and functional analysis. While still assuming compact and convex supports, this approach removed the need for Hölder smoothness or strong convexity of the Brenier potential. Building on this, [21] extended the theory to sub-exponential μ , under (α, a) -convexity and (β, a) -smoothness conditions on the Brenier potential. More recently, [20] relaxed the Poincaré inequality assumption to accommodate heavy-tailed μ and ν , and eliminated the (α, a) -convexity condition via a new sieved OT estimator, characterizing how tail thickness of μ and ν influences convergence.

We also note recent advances in plug-in type OT estimators, which directly compute the OT map between estimated distributions [10, 17, 34, 37, 5]. In parallel with these theoretical advances, significant efforts have focused on improving the computational efficiency of OT, including entropic regularization [15, 1, 22, 11, 19] and sum-of-squares formulations [35, 41]. Meanwhile, neural-network-based OT estimators have advanced rapidly, powering large-scale applications such as image generation and translation [33, 30, 4, 31, 12].

Despite these advances, a few core regularity assumptions on the true Brenier potential φ_0 and distributions still limit the applicability of dual-type estimators. To clarify these limitations, we decompose the overall estimation error into two components: statistical error and approximation error. The statistical error measures the discrepancy between the estimator and the best possible function in $\mathcal{F}$, reflecting the randomness from finite samples. The approximation error captures the gap between the true Brenier potential and its best approximation within $\mathcal{F}$, since we do not assume that the true Brenier potential φ_0 lies in $\mathcal{F}$. Existing analyses of each error component often rely on restrictive smoothness or convexity assumptions, which we discuss in more below.

First, bounding the statistical error in existing works requires a (β, b) -smoothness condition on φ_0 , i.e., a polynomial bound on its Hessian. While this condition is verifiable under strong log-concavity of μ and ν [8, 14, 13, 28], its validity beyond this class remains unclear. Moreover, Proposition 3.12 of [20] shows that this smoothness requirement constrains ν to have tails no much heavier than those of μ , excluding many practical scenarios, such as multivariate quantile [9, 10, 24], where we transport from $\mathbb{U}([0, 1]^d)$ to distributions with unbounded support, or transporting from a Gaussian (with exponential tails) to a Student's t -distribution (with polynomial tails).

Second, controlling the approximation error typically relies on (α, a) -convexity of φ_0 , i.e. a generalization of strong convexity, which similarly imposes constraints on the relative tail behavior of μ and ν [20]. For instance, OT map from an unbounded μ to a bounded ν fails this condition. This is especially problematic in generative modeling, where the OT map is learned from Gaussian noise to image data of bounded values [40, 30, 38, 32]. Moreover, existing approximation error bounds assume that both μ and ν have bounded support, [27, 21].

These limitations raise a fundamental question in broadening the scope of optimal transport:

Can one bound the statistical and approximation errors of dual-type OT estimators for general distributions without such restrictive assumptions?

In this work, we provide a positive answer to the question posed above. Our two core innovations are a novel map stability inequality and a refined oracle inequality. Building on these tools, we are the first to establish statistical and approximation error bounds for dual-type OT estimators under minimal regularity assumptions. Table 1 provides a side-by-side comparison of the regularity conditions required by our results versus those in the existing literature.

Table 1: Regularity assumptions on the true Brenier potential φ_0 and μ required to establish statistical and approximation error bounds ($\mathcal{E}_{stat}$ and $\mathcal{E}_{app}$) in prior work and in our results.

	Assumption on φ_0		Distributional Assumption	
	(β, b) -smooth (for $\mathcal{E}_{stat}$)	(α, a) -convex (for $\mathcal{E}_{app}$)	Tail thickness of μ	Poincaré inequality on μ
[27]	$b = 0$	$a = 0$	Compact	Implicitly Used
[25]	Not Required	Not Required	Compact	Required
[21]	Required	Required	Sub-exponential	Required
[20]	Required	Required	General	Poincaré-type inequality
Ours	Not Required	Not Required	General	Not Required

Specifically, we summarize our main contributions as follows:

1. A new map stability inequality without (β, b) -smoothness. Stability inequalities are key to linking OT estimation error to the semi-dual objective. Existing results rely on (β, b) -smoothness of φ_0 . In Proposition 3.3, we derive a new inequality that removes this requirement, assuming only a well-behaved OT map (Assumption 3.2), which holds unconditionally when ν is bounded and under mild conditions for unbounded distributions.

2. A novel oracle inequality without (α, a) -convexity. Proposition 3.4 provides an oracle inequality that avoids the (α, a) -convexity assumption on φ_0 and decomposes the estimation error into statistical error, sieved bias, and approximation error. Unlike prior works [21, 20], which involve harder-to-analyze gradient terms like $\inf_{\varphi \in \mathcal{F}} \|\nabla \varphi - \nabla \varphi_0\|_{L^2(\mu)}$, our bound relies on a more tractable supremum error over a compact region B : $\inf_{\varphi \in \mathcal{F}} \|\varphi - \varphi_0\|_{L^\infty(B)}$, making it more suitable for analysis using classical approximation theory.

3. First non-asymptotic bounds under minimal smoothness. With our new map stability inequality, Theorem 3.6 establishes the first non-asymptotic statistical error bound for dual-type OT estimators requiring only a first-order condition on φ_0 , without second-order smoothness assumptions.

4. First theoretical guarantee for neural OT estimators. With our new tools, we derive the first convergence guarantee for neural OT estimators using smooth networks. For tanh neural networks (TNNs), Theorem 3.7 shows that if $\varphi_0 \in C^\alpha(\mathbb{R}^d)$ for $\alpha \geq 2$, then the neural OT estimator $\nabla \hat{\varphi}_n$ between unbounded distributions satisfies:

$$\mathbb{E} \|\nabla \hat{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 \lesssim_{\log n} n^{-\frac{\alpha}{d+2\alpha}}. \quad (1)$$

Notably, this rate matches that of [25], but under significantly weaker assumptions: we do not require compact support for μ and ν , nor a Poincaré inequality on μ .

Notations and Terminologies Throughout this paper, we use the following notations. For $\alpha \in \mathbb{R}$, $\lfloor \alpha \rfloor = \{x \in \mathbb{Z} : x < \alpha\}$. For $v \in \mathbb{R}^d$, we write $\|v\|_2$ for its ℓ_2 norm and $\|v\|_\infty$ for its ℓ_∞ norm. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\|A\|_\infty := \max_{i,j} |A_{i,j}|$. If A is symmetric, $\|A\|_{op}$ denotes its operator norm, $\lambda_{\min}(A)$ denotes its smallest eigenvalue. For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and a measure μ , we write μf , $\mathbb{E}_\mu f(X)$, or $\int f d\mu$ for its integral, and $\|f\|_{L^2(\mu)} := (\int \|f\|_2^2 d\mu)^{1/2}$ for the L^2 norm. For $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and a set $B \subset \mathbb{R}^d$, we denote $\|f\|_{L^\infty(B)} := \sup_{x \in B} |f(x)|$.

We write $a \lesssim b$ to mean $a \leq Cb$ for some constant $C > 0$ independent of a and b , and $a \lesssim_{\log n} b$ to omit polylogarithmic factors in n . We write $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. We use $a \vee b$ and $a \wedge b$ to denote the maximum and minimum of a and b , respectively. For sequences a_n and b_n , we write $b_n = \mathcal{O}(a_n)$ if there exists a constant $c > 0$ such that $|b_n| \leq c a_n$, and write $b_n = \mathcal{O}_{\log}(a_n)$ if there exists constants $c_1, c_2 > 0$ such that $|b_n| \leq c_1 a_n \log(n)^{c_2}$. We use $\mathbb{I}$ to denote indicator function.

For $x \in \mathbb{R}^d$, write $\langle x \rangle := 1 + \|x\|_2$. When $x \in \mathbb{R}$, we set $\langle x \rangle := 1 + |x|$. We define $\log_+(x) := \max\{\log x, 1\}$. A function $\varphi \in C^2$ is called **(β, b) -smooth** if $\|\nabla^2 \varphi(x)\|_{op} \leq \beta \langle x \rangle^b$ for all x , and **(α, a) -convex** if $\lambda_{\min}(\nabla^2 \varphi(x)) \geq \alpha \langle x \rangle^a$.

A probability measure μ is **(λ, K) -sub-Weibull** if for $X \sim \mu$, $\mathbb{P}(\|X\|_2 \geq t) \lesssim \exp(-(t/K)^\lambda)$ for all $t > 0$. We refer to such a distribution as **λ -sub-Weibull**, or simply **sub-Weibull**.

2 Background on Optimal Transport

Let μ and ν be two probability measures on $\mathbb{R}^d$. A measurable map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is called a push-forward from μ to ν , if for any measurable set $A \subset \mathbb{R}^d$, $\mu(T^{-1}(A)) = \nu(A)$. We denote this by $\nu = T_\# \mu$. In Monge's problem, the optimal transport map is defined as the push-forward that minimizes the transport cost from μ to ν :

$$\min_{T: \mathbb{R}^d \rightarrow \mathbb{R}^d} \int_{\mathbb{R}^d} \|x - T(x)\|_2^2 \mu(dx), \quad \text{s.t. } T_\# \mu = \nu, \quad (2)$$

However, Monge's problem may not admit a solution, for example, when μ is discrete and ν is continuous. A natural relaxation is the Kantorovich formulation, which considers couplings $\pi \in \Pi(\mu, \nu)$ with marginals μ and ν :

$$\min_{\pi} \iint \|x - y\|_2^2 \pi(dx, dy), \quad \text{s.t. } \pi \in \Pi(\mu, \nu). \quad (3)$$

Kantorovich formulation is always feasible [42]. As a linear program, its semi-dual form reads:

$$\min_{\varphi \in L^1(\mu)} \left\{ \int \varphi d\mu + \int \varphi^* d\nu \right\} = \min_{\varphi \in L^1(\mu)} \mu\varphi + \nu\varphi^*. \quad (4)$$

where φ^* is the convex conjugate (a.k.a. Legendre-Fenchel conjugate) of φ :

$$\varphi^*(y) = \sup_{x \in \mathbb{R}^d} \{ \langle x, y \rangle - \varphi(x) \}, \quad \text{for all } y \in \mathbb{R}^d. \quad (5)$$

The equivalence between Monge's problem (2), Kantorovich's formulation (3), and its semi-dual form (4) was established in the seminal work of Brenier [6]:

Theorem 2.1 (Brenier's Theorem). *Let μ and ν be probability measures on $\mathbb{R}^d$, both with finite second moments, and assume that μ does not assign mass to small sets. Then there exists a convex function φ_0 that uniquely (up to constants) solves (4). Moreover, $\nabla \varphi_0$ solves Monge's problem (2), and $(Id \times \nabla \varphi_0)_\# \mu$ solves the Kantorovich formulation (3).*

Brenier's Theorem reveals that the OT map is the gradient of a convex function φ_0 , called the *Brenier potential*. Furthermore, since φ_0 is the minimizer of (4), we can estimate it using empirical measures μ_n and ν_m of μ and ν , respectively. For a chosen function class $\mathcal{F}$, we define the Brenier potential estimator and the associated dual-type OT estimator as:

$$\hat{\varphi}_{n,m} = \arg \min_{\varphi \in \mathcal{F}} \mu_n \varphi + \nu_m \varphi^*, \quad \text{with OT map estimator } \nabla \hat{\varphi}_{n,m}. \quad (6)$$

However, the supremum over the entire space $\mathbb{R}^d$ in the convex conjugate φ^* makes the optimization procedure unstable and sensitive. It also introduces theoretical challenges when analyzing the convergence of dual-type OT estimator $\hat{\varphi}_{n,m}$. To address this issue, [20] proposed the concept of a *sieved convex conjugate*, which restricts the supremum to a bounded region. Specifically, for some sieve radius $\tilde{R} > 0$, define:

$$\varphi_{\tilde{R}}^*(y) := \sup_{x: \|x\|_2 \leq \tilde{R}} \langle x, y \rangle - \varphi(x), \quad \text{for all } y \in \mathbb{R}^d. \quad (7)$$

Note that $\varphi^*(\nabla \varphi(x)) = \langle x, \nabla \varphi(x) \rangle - \varphi(x)$. By restricting x to $B(0, \tilde{R})$, we equivalently restrict the domain of $\varphi_{\tilde{R}}^*(y)$ to the compact set $\nabla \varphi(B(0, \tilde{R}))$. This compactness facilitates us to analyze the estimation error of sieved OT estimator.

Given a sieve radius $\tilde{R}$, the corresponding *sieved estimator* is defined as

$$\tilde{\varphi}_{n,m} := \arg \min_{\varphi \in \mathcal{F}} \mu_n \varphi + \nu_m \varphi_{\tilde{R}}^*, \quad \text{with sieved estimator } \nabla \tilde{\varphi}_{n,m}. \quad (8)$$

Note that when $\tilde{R} = \infty$, we recover the original dual-type estimator in Equation (6). In the following, we may abbreviate it to $\tilde{\varphi}_n$ when $n = m$.

3 Main Results

In Sections 3.1 and 3.2, we develop two core analytical tools: a new stability inequality for the sieved estimator (Proposition 3.3), and a refined oracle inequality that cleanly decomposes the estimation error (Proposition 3.4). Building on these results, Section 3.3 establishes non-asymptotic statistical error rates for general function classes (Theorem 3.6). Finally, in Section 3.4, we show how the overall estimation error of neural OT estimators between general distributions can be controlled, using tanh-activated neural networks a concrete example (Theorem 3.7).

3.1 A New Map Stability Inequality

From a statistical learning perspective, map stability inequalities are essential for linking the semi-dual objective to estimation error. For the original dual-type estimator, denote its objective is $\mathcal{S}_{\mu,\nu}(\varphi) := \mu\varphi + \nu\varphi^*$. Existing results [27, 21, 20] typically show that for all $\varphi \in \mathcal{F}$,

$$\|\nabla \varphi - \nabla \varphi_0\|_{L^2(\mu)}^2 \lesssim \mathcal{S}_{\mu,\nu}(\varphi) - \mathcal{S}_{\mu,\nu}(\varphi_0). \quad (9)$$

where $\nabla\varphi_0$ is the true OT map to be estimated.

However, these results apply only to the original estimator (6), not the sieved version in (8), which often performs better in practice. Moreover, while neither side of the inequality involves second-order derivatives, prior analyses still require (β, b) -smoothness of φ_0 , limiting their generality.

To accommodate sieved estimator and relax their assumptions, we consider the following conditions:

Assumption 3.1 (Envelope for Function Class). *Let $\mathcal{F} \subset C^2(\mathbb{R}^d)$ be a function class. We assume there exists a non-decreasing and pointwise finite function $U_2 : [0, \infty) \rightarrow [1, \infty)$ such that:*

$$\sup_{\varphi \in \mathcal{F}} \sup_{\|x\| \leq R} \|\nabla^2 \varphi(x)\|_{op} \leq U_2(R), \quad \sup_{\varphi \in \mathcal{F}} \|\nabla \varphi(0)\|_2 \leq U_2(R).$$

Assumption 3.1 depends only on $\mathcal{F}$ and is independent of φ_0, μ and ν . It is satisfied by many common function classes, for example,

1. Quadratic function class: When μ and ν are both Gaussian (or from the same elliptical family), the Brenier potential is quadratic and lies in the following quadratic family:

$$\mathcal{F}_{quad} = \{x \mapsto x^\top Bx + \langle b, x \rangle : B \in \mathbb{S}_+^d, \|B\|_{op} \leq r_1, \|b\|_2 \leq r_2\}.$$

2. Smooth Neural Networks: Neural network function classes with all parameters bounded and activation functions that are C^2 -smooth, such as sigmoid, tanh, or softmax.
3. Reproducing Kernel Hilbert Spaces (RKHS): Let $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a positive-definite kernel on its domain $\mathcal{X} \times \mathcal{X}$, and let $\mathcal{H}_K$ be the corresponding RKHS with norm $\|\cdot\|_{\mathcal{H}_K}$. We assume $\mathcal{F}$ is C^4 and take $\mathcal{F}$ to be the unit ball of $\mathcal{H}_K$, i.e., $\mathcal{F} = \{\varphi \in \mathcal{H}_K : \|\varphi\|_{\mathcal{H}_K} \leq 1\}$.

Assumption 3.1 and 3.2 together play a role similar to Assumption A.1 in [21], but are strictly weaker. As a result, the function classes discussed in Section 4 of [21], including parametric family, wavelet expansions, RKHS, and Barron Spaces, apply directly to our setting. We refer readers to Section 4 of [21] for a detailed discussion.

Assumption 3.2 (Envelope for OT map). *There is a function $u \in L^4(\mu)$, such that for all $\varphi \in \mathcal{F}$,*

$$\|\nabla \varphi(x) - \nabla \varphi_0(x)\|_2 \leq u(x), \quad \text{for all } x.$$

Define $U_1(R) := \sup_{\|x\|_2 \leq R} u(x)$. Additionally, let $\Phi \in L^2(\mu)$ be an envelope function for $\mathcal{F} \cup \{\varphi_0\}$, i.e.

$$|\varphi(x)| \leq \Phi(x), \quad \text{for all } x \in \mathbb{R}^d \text{ and all } \varphi \in \mathcal{F} \cup \varphi_0.$$

A sufficient condition for Assumption 3.2 is that both Proposition A.2 and Assumption 3.1 hold. Specifically, Proposition A.2 ensures $\|\nabla \varphi_0(x)\|_2$ to be bounded by some function $L_1(x)$. Meanwhile, Assumption 3.1 provides an envelope for $\|\nabla \varphi(x)\|_2$, denoted as $L_2(x)$. Then Assumption 3.2 follows via the triangle inequality $\|\nabla \varphi(x) - \nabla \varphi_0(x)\|_2 \leq L_1(x) + L_2(x)$. While Assumption 3.1 solely relies on $\mathcal{F}$, typical distributions for Proposition A.2 to hold include: μ can be normal or Student-t, and ν can be any measure with certain tail decay. We refer to Appendix A for further details and an explicit example verifying Assumption 3.2.

We are now ready to present our new stability inequality for the sieved estimator:

Proposition 3.3 (Map Stability Inequality). *Suppose Assumptions 3.1 and 3.2 hold. For any $\varepsilon > 0$, define R_ε and sieve radius $\tilde{R}_\varepsilon$ such that*

$$\mu(\|X\|_2 > R_\varepsilon) \leq \varepsilon, \quad \tilde{R}_\varepsilon \geq \sup_{\|x\|_2 \leq R_\varepsilon} \left\{ \|x\|_2 + \frac{u(x)}{U_2(\|x\|_2 + u(x))} \right\} \quad (10)$$

For any $\varphi \in \mathcal{F}$, define the truncated excess risk:

$$r_\varepsilon(\varphi) := \int_{B(0, R_\varepsilon)} \varphi(x) + \varphi_{\tilde{R}_\varepsilon}^*(\nabla \varphi_0(x)) \mu(dx) - \int_{B(0, R_\varepsilon)} \varphi_0(x) + \varphi_0^*(\nabla \varphi_0(x)) \mu(dx). \quad (11)$$

Then the following inequality holds: for any $\varphi \in \mathcal{F}$,

$$\|\nabla \varphi - \nabla \varphi_0\|_{L^2(\mu)}^2 \leq 2U_2(R_\varepsilon + U_1(R_\varepsilon)) \cdot r_\varepsilon(\varphi) + \|u\|_{L^4(\mu)}^2 \cdot \varepsilon^{\frac{1}{2}}, \quad (12)$$

We first clarify the roles of R_ε and $\tilde{R}_\varepsilon$, which enable a more tractable approximation error in the next section. Specifically, the hyper-balls $B(0, R_\varepsilon)$ and $\nabla\varphi_0(B(0, R_\varepsilon))$ serve as bounded pseudo-supports for μ and ν , respectively, over which the behavior of φ and its convex conjugate can be effectively controlled. Meanwhile, the sieve radius $\tilde{R}_\varepsilon$ is introduced to handle any mismatch between $\nabla\varphi(B(0, R_\varepsilon))$ and the pseudo-support of ν , by better regulating the range of the sieved conjugate. In practice, if $\nabla\varphi$ closely approximates $\nabla\varphi_0$ in L^∞ norm (so that $u \approx 0$), the mismatch is negligible and one can take $\tilde{R}_\varepsilon \approx R_\varepsilon$. In that case, it suffices to compute the convex conjugate over the pseudo-support of μ alone.

For simplicity, a sufficient choice for $\tilde{R}_\varepsilon$ is $\tilde{R}_\varepsilon = R_\varepsilon + U_1(R_\varepsilon)$. In particular, if μ is supported on $B(0, R)$, one may take $R_\varepsilon = R$.

Equation (12) decomposes the estimation error into two components. The first arises from the truncated excess risk over the high-probability region $B(0, R_\varepsilon)$, which can be well-controlled (see Proposition 3.4). Its coefficient $U_2(R_\varepsilon + U_1(R_\varepsilon))$ originates from a second-order Taylor expansion of φ , capturing the “uniform curvature” of $\mathcal{F}$ on the ball $B(0, R_\varepsilon + U_1(R_\varepsilon))$. The second component accounts for the residual risk outside this region. As a result, ε must be carefully selected to balance the trade-off: a smaller ε reduces the residual risk but increases the truncated excess risk.

We defer the proof of Proposition 3.3 to Appendix B.

Comparison with existing map stability results. The most closely related results are Proposition 10 of [27], Proposition 1 of [21], and Lemma 3.15 of [20], all of which require (β, b) -smoothness assumptions on the true Brenier potential φ_0 and the candidate class $\mathcal{F}$. In contrast, our Proposition 3.3 substantially relaxes these requirements by assuming only an upper bound on the true OT map $\nabla\varphi_0$ via function $u(x)$.

Another key distinction is the introduction of the sieved convex conjugate $\varphi_{\tilde{R}_\varepsilon}^*$, essential for providing a more accessible approximation error in the next section. This sieving remains necessary even when μ has bounded support, unless additional regularity conditions are imposed on μ and ν .

Compared to map stability results for plug-in type estimators, our bound also holds under milder assumptions. For instance, Theorem 2.1 of [17] requires $\nabla\varphi_0$ to be Lipschitz, while Theorem 3 of [5] assumes that the Brenier potential is strongly convex.

3.2 A New Oracle Inequality

Now, we present our second contribution of a new oracle inequality that helps cleanly decompose the estimation error into statistical error, sieved bias, and approximation error.

Proposition 3.4 (Oracle Inequality). *Under Assumptions 3.1 and 3.2, and let $R_\varepsilon, \tilde{R}_\varepsilon$ be as defined in Equation (10). For the sieved OT estimator $\tilde{\varphi}_{n,m}$ from Equation (8), its truncated excess risk $\tilde{r}_{n,m,\varepsilon} := r_\varepsilon(\tilde{\varphi}_{n,m})$, defined in Equation (11), admits the decomposition:*

$$\tilde{r}_{n,m,\varepsilon} := r_\varepsilon(\tilde{\varphi}_{n,m}) \leq \mathcal{E}_{stat} + \mathcal{E}_{sieve} + \mathcal{E}_{app}, \quad (13)$$

where

$$\mathcal{E}_{stat} := \sup_{f \in \bar{\mathcal{F}}} \int_{B(0, R_\varepsilon)} f d(\mu_n - \mu) + \sup_{g \in \bar{\mathcal{G}}} \int_{B(0, R_\varepsilon)} g d(\nu_m - \nu), \quad (14a)$$

$$\mathcal{E}_{sieve} := 2\mathbb{E}_{X \sim \mu_n} [\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)] + 2\mathbb{E}_{Y \sim \nu_m} [G(Y) \cdot \mathbb{I}(\|(\nabla\varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)], \quad (14b)$$

$$\mathcal{E}_{app} := 2 \inf_{\varphi \in \mathcal{F}} \|\varphi - \varphi_0\|_{L^\infty(B(0, \tilde{R}_\varepsilon))}. \quad (14c)$$

Here, $\bar{\mathcal{F}} := \{\varphi_1 - \varphi_2 : \varphi_1, \varphi_2 \in \mathcal{F}\}$, $\bar{\mathcal{G}} := \{\varphi_{1, \tilde{R}_\varepsilon}^* - \varphi_{2, \tilde{R}_\varepsilon}^* : \varphi_1, \varphi_2 \in \mathcal{F}\}$ and $G(y) := \tilde{R}_\varepsilon \|y\|_2 + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x)$.

Thus, with a careful choice of ε and R_ε (which can be selected in a distribution-free manner, as discussed in Section 3.3), Propositions 3.3 and 3.4 together provide a complete characterization of how the sample sizes, function class $\mathcal{F}$, and the sieved convex conjugate collectively influence the estimation error of the sieved dual-type estimators.

Specifically, this result decomposes the truncated excess risk into three components. The first term, $\mathcal{E}_{stat}$, reflects the statistical error from replacing the underlying distributions μ and ν with their

estimates μ_n and ν_m . Importantly, these estimates need not be empirical measures, allowing for smoothed or bootstrap-based alternatives. The second term, $\mathcal{E}_{sieve}$, accounts for the bias introduced by replacing the original convex conjugate (Equation (7)) with its sieved version. This bias vanishes when $R_\varepsilon = \infty$, where the sieved and original convex conjugates coincide.

The third term, $\mathcal{E}_{app}$, captures the approximation error of the function class $\mathcal{F}$ in representing the true potential φ_0 over the compact domain $B(0, \tilde{R}_\varepsilon)$. As usual, enlarging $\mathcal{F}$ reduces $\mathcal{E}_{app}$ at the cost of increasing $\mathcal{E}_{stat}$. Notably, we express the approximation error in the L^∞ norm on a compact set, aligning naturally with classical approximation theory. In contrast, prior works [27, 21, 20] typically rely on gradient-based errors like $\inf_{\varphi \in \mathcal{F}} \|\nabla \varphi - \nabla \varphi_0\|_{L^2(\mu)}$, which remain challenging to analyze for rich function classes such as deep neural networks.

Proof ideas of Propositions 3.4 The sieve technique allows us to treat $B(0, R_\varepsilon)$ and $\nabla \varphi_0(B(0, \tilde{R}_\varepsilon))$ as bounded pseudo-supports of μ and ν , or equivalently, as bounded pseudo-domains for φ_0 and φ_0^* . This reduction localizes the approximation problem to compact sets, where classical approximation theory applies. Another key benefit of this sieving approach is that it eliminates the need for the (α, a) -convexity assumption required to control the approximation error in prior works.

The sieve bias is then controlled via tail probability bounds, while the statistical error is handled using standard tools from empirical process theory. The full proof is deferred to Appendix C.

3.3 Non-Asymptotic Bounds of Statistical Error and Sieved Bias

In this section, we derive bounds for the statistical error $\mathcal{E}_{stat}$ and the sieved bias $\mathcal{E}_{sieve}$ in Proposition 3.4, under minimal assumptions on the source distribution μ . Notably, our results do not require smoothness of the true OT map $\nabla \varphi_0$, improving upon existing analyses.

We begin with the following assumption on the covering entropy of the function class $\mathcal{F}$:

Assumption 3.5 (Covering Entropy of $\mathcal{F}$). *For some $\eta \geq 0$, $\gamma \in [0, 2)$, $\gamma' \geq 1$ and $D_{\mathcal{F}} > 1$, the covering entropy of $\mathcal{F}$ satisfies: for any $h \geq 0$,*

$$\log \mathcal{N}(h, \mathcal{F}, L^\infty([-R, R]^d)) \leq D_{\mathcal{F}} \cdot h^{-\gamma} \cdot \log_+(1/h)^{\gamma'} \cdot R^\eta. \quad (15)$$

We focus on the Donsker regime ($\gamma \in [0, 2)$) for theoretical simplicity, which already captures a wide range of function classes, such as parametric families, wavelets, reproducing kernel Hilbert space (RKHS), and both shallow and deep neural networks. Here, $D_{\mathcal{F}}$ represents the “effective dimension” of $\mathcal{F}$, and $n/D_{\mathcal{F}}$ can be viewed as the effective sample size from the statistical learning theory perspective. Because most covering entropy results are stated for the $L^\infty([0, 1]^d)$ norm, we introduce an additional R^η factor to translate them to our unbounded setting in $L^\infty([-R, R]^d)$. Specifically, the exponent η measures how the envelope of $\mathcal{F}$ grows (e.g. $\eta = 2$ if the envelope scales like $x \mapsto x^2$).

Theorem 3.6 (Statistical Error with Empirical Measures). *Under assumptions of Proposition 3.4, and Assumption 3.5, the sieved estimator $\tilde{\varphi}_{n,m}$ from Equation (8), computed with empirical measures μ_n and ν_m , satisfies:*

$$\mathbb{E}[\mathcal{E}_{stat}] \lesssim \sqrt{\frac{D_{\mathcal{F}}}{n}} \left(R_\varepsilon^{\frac{\eta}{2}} + \sup_{x \in B(0, R_\varepsilon)} \Phi(x)^{\frac{1}{2}} \right) \log(n)^{\frac{\gamma'}{2}} + \sqrt{\frac{D_{\mathcal{F}}}{m}} \left(\tilde{R}_\varepsilon^{\frac{\eta}{2}} + M^{\frac{1}{2}} \right) \log(m)^{\frac{\gamma'}{2}}, \quad (16)$$

and

$$\mathbb{E}[\mathcal{E}_{sieve}] \leq 2(\|\Phi\|_{L^2(\mu)} + \|G\|_{L^2(\nu)}) \cdot \varepsilon^{\frac{1}{2}}, \quad (17)$$

where $M := \tilde{R}_\varepsilon \sup_{x \in B(0, R_\varepsilon)} \|\nabla \varphi_0(x)\|_2 + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x)$, and the suppressed constant depends only on γ, γ' .

We defer the proof to Appendix D, where we also establish corresponding high-probability guarantees.

In Theorem 3.6, we establish a standard convergence rate for estimation error, $\mathbb{E}[\mathcal{E}_{stat}] = \mathcal{O}_{\log}(\sqrt{D_{\mathcal{F}}/n} + \sqrt{D_{\mathcal{F}}/m})$, which is common in empirical process theory. Meanwhile, the multiplicative prefactors in $\mathcal{E}_{stat}$, $R_\varepsilon^{\frac{\eta}{2}} + \sup_{x \in B(0, R_\varepsilon)} \Phi(x)^{\frac{1}{2}}$ and $\tilde{R}_\varepsilon^{\frac{\eta}{2}} + M^{\frac{1}{2}}$, naturally capture the “diameter” of $\mathcal{F}$ on the pseudo-supports of μ and ν .

Moreover, it tells that the sieve bias $\mathcal{E}$ is governed by the tail probability ε defined in Equation (10).

Practical choice of R_ε and $\tilde{R}_\varepsilon$. As shown in [20], R_ε can be selected in a distribution-free manner. Here, we extend this idea to provide a practical, joint selection strategy for both R_ε and $\tilde{R}_\varepsilon$:

$$R_\varepsilon = \max_{1 \leq i \leq n} \|X_i\|_2, \quad \tilde{R}_\varepsilon = R_\varepsilon + C \max_{1 \leq j \leq m} \|Y_j\|_2, \quad (18)$$

where $C \geq 0$ is a tuning parameter designed to account for the potential mismatch between the estimated and true OT maps, specifically the discrepancy $\|\nabla \tilde{\varphi}_{n,m} - \nabla \varphi_0\|_{L^\infty(B(0, R_\varepsilon))}$.

The choice of R_ε can also be justified by our stability and oracle bounds (Propositions 3.3 and 3.4). In particular, if $n = m$, μ is sub-Weibull, and both $\Phi(\cdot)$, $\|\nabla \varphi_0(\cdot)\|_2$ grow at most polynomially, then the estimation error of the sieved estimator reads:

$$\mathbb{E} \|\nabla \tilde{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 = \mathcal{O}_{\log} \left((n/D_{\mathcal{F}})^{-\frac{1}{2}} + \mathcal{E}_{app} + \varepsilon^{\frac{1}{2}} \right). \quad (19)$$

By exchangability of X_i 's, choosing $R_\varepsilon := \max_{1 \leq i \leq n} \|X_i\|_2$ yields $\varepsilon \asymp n^{-1}$, so the sieved bias $\varepsilon^{\frac{1}{2}}$ is dominated by the statistical error term $(n/D_{\mathcal{F}})^{-\frac{1}{2}}$.

While a complete theoretical justification for our choice of $\tilde{R}_\varepsilon$ remains an open question, numerical results in Section 4 provide strong empirical support: our sieved estimator consistently outperforms the original dual-type estimator of [27, 25, 21] across a range of settings.

3.4 Estimation Error of Neural OT Estimator

In this section, we derive the first non-asymptotic error bounds for dual-type OT map estimators parameterized by deep neural networks between general distributions.

ReLU networks, despite their popularity, lack the smoothness needed to recover the gradient of the estimated Brenier potential. Moreover, our analysis (Assumption 3.1) requires $\mathcal{F}$ to be C^2 . Consequently, smooth activations are both practically and theoretically preferable.

While input convex neural networks (ICNNs) [2] exploit convexity of Brenier potential and have been applied at scale [33, 4, 31], their approximation theory remains underdeveloped. This gap limits rigorous control of approximation error in existing analyses of dual-type estimators [27, 25, 21, 20]. By contrast, our approach removes the convexity assumption on $\mathcal{F}$, allowing us to employ standard fully connected networks.

Accordingly, we adopt the tanh-activated fully-connected neural network (TNN) as our function class $\mathcal{F}$, due to its C^∞ smoothness and non-asymptotic universal approximation guarantees [16]:

$$f(x) = \mathcal{L}_L \circ \sigma \circ \mathcal{L}_{L-1} \circ \cdots \circ \sigma \circ \mathcal{L}_1(x), \quad (20)$$

where $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is applied entrywise, and $\mathcal{L}_\ell(x) = A_\ell x + b_\ell$ for $\ell \in [L]$, where $A_\ell \in \mathbb{R}^{p_{\ell+1} \times p_\ell}$, $b_\ell \in \mathbb{R}^{p_\ell}$, with $p_1 = d$, $p_{L+1} = 1$.

We define the TNN class $\mathcal{F}$ with bounded parameters as:

$$\mathcal{F}(L, W, \kappa) := \{f \text{ of form (20)} : \|A_\ell\|_\infty \vee \|b_\ell\|_\infty \leq \kappa, \quad p_\ell \leq W \text{ for all } \ell \in [L]\}, \quad (21)$$

where $\kappa > 0$ is a truncation threshold used for technical convenience.

Theorem 3.7 (OT Estimation via Tanh Neural Network). *Let $n = m$, and suppose μ is (λ, K) -sub-Weibull, and the true Brenier potential $\varphi_0 \in C^\alpha(\mathbb{R}^d)$ with $\alpha \geq 2$ is (β, b) -smooth. By setting $R_\varepsilon, \tilde{R}_\varepsilon$ according to Equation (10), there exists a deep TNN function class $\mathcal{F} := \mathcal{F}(L, W, \kappa)$ where L, W, κ depend on n, d, α, β, b and are of order $3 \leq L = \mathcal{O}(1)$, $W = \mathcal{O}(n^{\frac{d}{d+2\alpha}})$, $\kappa = \mathcal{O}(n^{\frac{d(d+(\lfloor \alpha \rfloor + 2)^2 + 4) + 2}{2(d+2\alpha)}})$, such that the sieved estimator $\nabla \tilde{\varphi}_n$ with sieve radius*

$$\tilde{R}_n = C_{K,d,\alpha} \cdot (\log n)^{\frac{1}{\lambda}}, \quad \text{for some constant } C_{K,d,\alpha} \geq 4K \frac{2\alpha}{d+2\alpha} + 2 \quad (22)$$

satisfies

$$\mathbb{E} \|\nabla \tilde{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 \lesssim_{\log n} n^{-\frac{\alpha}{d+2\alpha}}. \quad (23)$$

Remark 3.8 (On mild regularity conditions). *The sub-Weibull and (β, b) -smoothness assumptions are used in our work primarily for analytical convenience, not because they are fundamentally*

required for the methodology. Notably, both conditions are widely used in the OT literature, see [21, 20, 27, 25, 34] for (β, b) -smoothness and [21, 17, 20] for sub-Weibull condition.

The (β, b) -smoothness of φ_0 is needed solely to control approximation error through Lemma F.4, ensuring the error grows polynomially in $\tilde{R}_\varepsilon$. Meanwhile, the sub-Weibull μ guarantees that $R_\varepsilon, \tilde{R}_\varepsilon$ scale polynomially in $\log(1/\varepsilon)$. Together, these two results ensure that the relevant Sobolev norm is bounded by a polynomial in $\log(1/\varepsilon)$, and the consequent approximation error has the order of ε .

If (β, b) -smoothness were lifted, one could still obtain approximation bounds, but they may degrade to $O(\varepsilon^a)$ for some $a < 1$. This complicates the statistical error, approximation error, and sieved estimation bias-variance trade-off by introducing additional convergence exponent. Likewise, relaxing source distribution to be polynomially tails causes $R_\varepsilon, \tilde{R}_\varepsilon$ to scale polynomially in $1/\varepsilon$. While the proofs follow similar steps, the resulting convergence rate becomes much more complex (See Equation 3.10 in [20] for an illustration of the added technical complications).

Remark 3.9 (On the choice of activation). While ReLU networks are widely used in practice, our framework currently focuses on the tanh activation due to theoretical considerations. Specifically, our estimator is defined via the gradient of a potential function. Although PyTorch can handle ReLU networks computationally, the ReLU activation is not globally differentiable in mathematics, making this definition ill-posed. Moreover, even under a weak derivative framework, ReLU networks lack second-order weak derivatives, whereas the existing optimal transport theory fundamentally relies on second-order regularity for statistical analysis.

The rate in Theorem 3.7 matches that of [25], but under substantially milder assumptions. Although slower than the minimax rate in [27], this is expected: we do not assume (α, a) -convexity of the true Brenier potential φ_0 , compact support of μ and ν , or a Poincaré inequality on μ . Moreover, the estimators in [27, 25, 21] are either computationally intensive or NP-hard. In contrast, our deep TNN estimator is both implementable and scalable in practice.

We focus on the TNN architecture due to the current lack of approximation theory for general smooth neural networks. The proof of Theorem 3.7 is provided in Appendix E. With minor modifications, the result also holds with high probability.

4 Numerical Examples

In this section, we present numerical simulations to evaluate the performance of our sieved-TNN estimator and compare it against the original dual-type estimators from [27, 25, 21]. We consider four synthetic OT problems that pose challenges for existing theoretical frameworks:

- *Non- (β, b) -smooth maps:* $\mathcal{N}(0, 1) \rightarrow t_6$, $\text{Uniform}(0, 1) \rightarrow \mathcal{N}(0, 1)$.
- *Non- (α, a) -convex maps:* $t_6 \rightarrow \mathcal{N}(0, 1)$, $\mathcal{N}(0, 1) \rightarrow \text{Uniform}(0, 1)$.

We evaluate performance in dimensions $d = 5, 10, 20$, using i.i.d. samples $X_i \sim \mu$ and $Y_j \sim \nu$ with sample sizes $n = m \in \{64, 128, 256, 512, 1024, 2048\}$. Experiments were conducted on a server with Intel Xeon Gold 6342 processors, requiring approximately 4,000 CPU core hours in total.

Implementation setup We implement our estimator in PyTorch [36], following algorithms from [20] with objective function in Equation (8). The candidate class $\mathcal{F}$ consists of TNNs with two hidden layers of width 10 (for $d = 5$), 20 (for $d = 10$), and 30 (for $d = 20$). Training proceeds in two phases: a warm-up phase of $\max\{\lfloor 10000/n \rfloor, 50\}$ epochs at learning rate 5×10^{-3} , followed by 300 epochs at 10^{-3} . Mini-batch size is 64 for both μ and ν , and the sieved convex conjugate subproblem is solved using 300 inner iterations (see Algorithm 2 in [20]). In each trial, 10% of the samples are held out for validation. We defer implementation details to Appendix G.

Sieve radius As discussed in Equation (18), we choose the sieve radius $\tilde{R}$ as

$$\tilde{R} = \max_{1 \leq i \leq n} \|X_i\|_2 + C \cdot \max_{1 \leq j \leq n} \|Y_j\|_2, \quad C \in \{0, 1, 2, 3, \infty\}.$$

Here, $C = 0$ corresponds to the setting in [20], while $C = \infty$ recovers the original dual-type estimator studied in [27, 25, 21], enabling direct comparison.

Evaluation We assess the estimation error using unexplained variance proportion (UVP) [30]: $L^2\text{-UVP}(\nabla\hat{\varphi}_n) := \|\nabla\hat{\varphi}_n - \nabla\varphi_0\|_{L^2(\mu)}^2 / \text{Var}_\nu(\|Y\|_2)$. Lower values indicate better performance. For each experiment, we approximate the $L^2\text{-UVP}$ using an independent set of 10^6 samples.

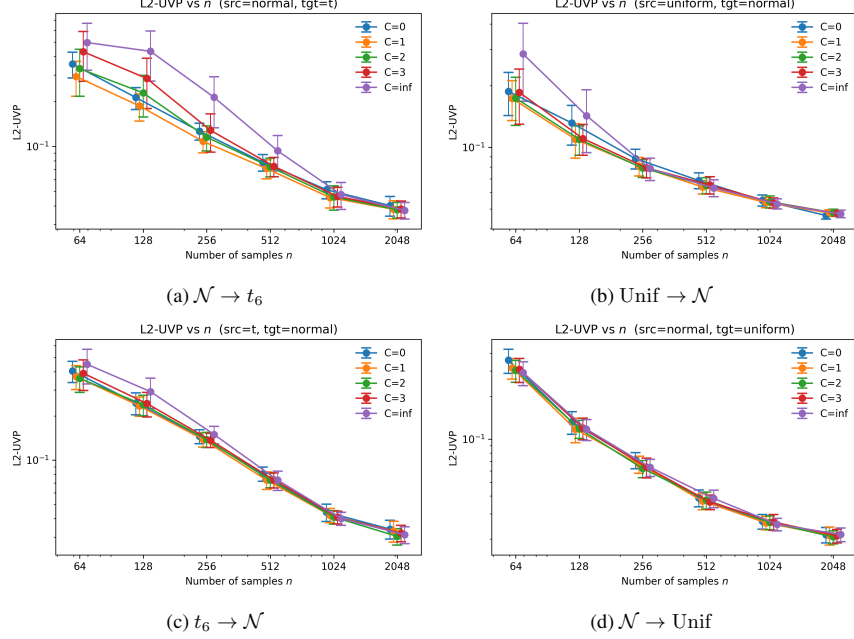


Figure 1: $L^2\text{-UVP}$ when $d = 5$. Each curve shows the mean $L^2\text{-UVP}$ over 50 random trials with one standard deviation.

Due to space constraints, we report only the $L^2\text{-UVP}$ error versus sample size n for $d = 5$ in Figure 1 here. The full set of numerical results is available in Appendix G.

Across all settings, our sieved TNN estimator consistently converges and outperforms the original dual-type estimator ($C = \infty$), particularly in the small-sample regime. As shown in Figure 4a, sieve-based estimators yield lower $L^2\text{-UVP}$ errors for small n , while the classical estimator only catches up as n grows large. This aligns with our theory: larger sieve radii may degrade convergence, and are only needed when the mismatching $\|\nabla\tilde{\varphi}_{n,m} - \nabla\varphi_0\|_{L^\infty}$, is large.

These results also offer practical guidance: a modest sieve constant (e.g., $C = 0$ or 1) suffices, and performance of our sieved estimator remains stable across a range of C .

5 Conclusion

In this paper, we develop new map stability and oracle inequalities for sieved dual-type OT estimators, without restrictive smoothness and convexity assumptions on the true Brenier potential φ_0 in the literature. Under mild regularity conditions, we establish the first non-asymptotic error bounds for neural OT estimators, using tanh networks as a concrete example. Numerical experiments further confirm the strong empirical performance of our sieved approach. Altogether, our unified framework advances the theoretical foundations of optimal transport and paves the way for future developments in both methodology and applications.

Acknowledgment: This research was partially supported the U.S. National Science Foundation under the grants DMS 2514400 and DMS 2210775, and by the U.S. National Institutes of Health under the grants R01GM163244 and 1R01GM152812.

References

- [1] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.
- [2] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [4] Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, and Evgeny Burnaev. Neural optimal transport with general cost functionals. In *International Conference on Learning Representations*, 2022.
- [5] Sivaraman Balakrishnan and Tudor Manole. Stability bounds for smooth optimal transport maps and their statistical implications. *arXiv preprint arXiv:2502.12326*, 2025.
- [6] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
- [7] 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.
- [8] Luis A Caffarelli. Monotonicity properties of optimal transportation and the fkg and related inequalities. *Communications in Mathematical Physics*, 214:547–563, 2000.
- [9] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–92, 2016.
- [10] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017.
- [11] Lenaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.
- [12] Jaemoo Choi, Yongxin Chen, and Jaewoong Choi. Improving neural optimal transport via displacement interpolation. *The Thirteenth International Conference on Learning Representations*, 2025.
- [13] Maria Colombo and Max Fathi. Bounds on optimal transport maps onto log-concave measures. *Journal of Differential Equations*, 271:1007–1022, 2021.
- [14] Maria Colombo, Alessio Figalli, and Yash Jhaveri. Lipschitz changes of variables between perturbations of log-concave measures. *arXiv preprint arXiv:1510.03687*, 2015.
- [15] M Cuturi. Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26(2):2292–2300, 2013.
- [16] Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.
- [17] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In *Advances in Neural Information Processing Systems*, volume 34, pages 29736–29753, 2021.
- [18] Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, 118(541):192–207, 2023.
- [19] Alex Delalande. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1619–1642. PMLR, 2022.

- [20] Yizhe Ding, Runze Li, and Lingzhou Xue. Statistical convergence rates of optimal transport map estimation between general distributions. *arXiv preprint arXiv:2412.08064*, 2024.
- [21] Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *The Annals of Statistics*, 53(3):963–988, 2025.
- [22] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm. In *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.
- [23] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.
- [24] Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *The Annals of Statistics*, 50(2):1012–1037, 2022.
- [25] Florian F Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, 38(2):381–417, 2022.
- [26] Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d : A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021.
- [27] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 2021.
- [28] Ksenia A Khudiakova, Jan Maas, and Francesco Pedrotti. L_∞ -optimal transport of anisotropic log-concave measures and exponential convergence in fisher’s infinitesimal model. *The Annals of Applied Probability*, 35(3):1913–1940, 2025.
- [29] Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Genot: Entropic (gromov) wasserstein flow matching with applications to single-cell genomics. *Advances in Neural Information Processing Systems*, 37:103897–103944, 2024.
- [30] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.
- [31] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. 2023.
- [32] Zezeng Li, Shenghao Li, Zhanpeng Wang, Na Lei, Zhongxuan Luo, and David Xianfeng Gu. Dpm-ot: a new diffusion probabilistic model based on optimal transport. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 22624–22633, 2023.
- [33] 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.
- [34] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.
- [35] Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.
- [36] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [37] Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning*, pages 28128–28150. PMLR, 2023.

- [38] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. 2022.
- [39] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.
- [40] Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.
- [41] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173. PMLR, 2021.
- [42] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.
- [43] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Chang-gong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021.

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: We provide exact Theorems/Propositions numbers for our main claims in the abstract and introduction, which correspond precisely with our four subsections in Section 3.

Guidelines:

- The answer NA 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 NA 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 explicitly discuss the limitations of our work. For example, in Section 3.3, we acknowledge the lack of theoretical justification for the practical choice of the sieve radius $\tilde{R}_\varepsilon$.

Regarding the convergence rate of the sieved TNN estimator in Theorem 3.7, we note that it is not minimax optimal as in [27], which is expected due to our substantially weaker assumptions. Furthermore, we explain our restriction to tanh networks and do not consider input convex neural networks (ICNNs), owing to the limited approximation theory available for general smooth neural networks.

Guidelines:

- The answer NA 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: Proof of our results are included in the appendix.

Guidelines:

- The answer NA 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 ensure full reproducibility by including a description of our sieved-TNN model and important training hyper-parameters setting in Section 4. Full description of our algorithm is provided in Appendix G. Meanwhile, as a synthetic simulation study, the data generating processes are provided in our code for reproducing consideration.

Guidelines:

- The answer NA 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: Full implementation code is available in the Supplementary Materials. All experimental data are synthetic and can be recreated exactly using the data-generation procedures and fixed random seeds detailed in Appendix G.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4 details the sieved-TNN architecture, data-splitting, and all key hyperparameters (learning rate, batch size, network depth, etc.). Appendix G provides the complete experimental design, including full algorithm pseudo-code, the process of training, the synthetic data-generation procedure with fixed random seeds, and the optimizer choice.

Guidelines:

- The answer NA 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: Every plotted result includes error bars representing one standard error of the mean over multiple independent runs. Since no statistical hypothesis tests are involved, the discussion of statistical significance is not applicable.

Guidelines:

- The answer NA 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: We briefly introduce the configuration for our server in Section 4, with full description available in the Appendix G.

Guidelines:

- The answer NA 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: We have all reviewed the NeurIPS Code of Ethics and striven to maintain and preserve anonymity.

Guidelines:

- The answer NA 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: [NA]

Justification: This work studies the theoretical properties of optimal transport and the authors cannot see any direct or potential societal impacts this work can bring.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: As a theoretical study on optimal transport, this work does not involve any data sets, and does not pose risks to existing optimal transport models for image translation or generation.

Guidelines:

- The answer NA 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: [NA]

Justification: We did not incorporate any third-party code, data, or models. We properly cited, respected, mentioned the algorithm from [20] and re-implemented it ourselves, so no external “assets” or licenses are involved.

Guidelines:

- The answer NA 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 full sieved-TNN implementation, including source code, README, and pseudo-code, as a new asset in the Supplementary Materials (with plans to publish it on GitHub upon acceptance).

Guidelines:

- The answer NA 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: [NA]

Justification: As a theoretical work, no human subjects are involved in it.

Guidelines:

- The answer NA 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: [NA]

Justification: This theoretical study does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA 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: [NA]

Justification: The theoretical and methodology advancements in this work are built on well-established mathematical tools, and LLM is not utilized for developing them.

Guidelines:

- The answer NA 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 (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.