Neural Estimation of Entropic Optimal Transport

Tao Wang
Applied Mathematics and Computational Science
University of Pennsylvania
Email: tawan@sas.upenn.edu

Ziv Goldfeld
School of Electrical and Computer Engineering
Cornell University
Email: goldfeld@cornell.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619399

***Abstract*—Optimal transport (OT) serves as a natural framework for comparing probability measures, with applications in statistics, machine learning, and applied mathematics. Alas, statistical estimation and exact computation of the OT distances suffer from the curse of dimensionality. To circumvent these issues, entropic regularization has emerged as a remedy that enables parametric estimation rates via plug-in and efficient computation using Sinkhorn iterations. Motivated by further scaling up entropic OT (EOT) to data dimensions and sample sizes that appear in modern machine learning applications, we propose a novel neural estimation approach. Our estimator parametrizes a semi-dual representation of the EOT distance by a neural network, approximates expectations by sample means, and optimizes the resulting empirical objective over parameter space. We establish non-asymptotic error bounds on the EOT neural estimator of the cost and optimal plan. Our bounds characterize the effective error in terms of neural network size and the number of samples, revealing optimal scaling laws that guarantee parametric convergence. The bounds hold for compactly supported distributions, and imply that the proposed estimator is minimax-rate optimal over that class. Numerical experiments validating our theory are also provided.**

I. Introduction

Optimal transport (OT) theory [1] provides a natural framework for comparing probability distributions. Specifically, given two Borel probability measures μ, ν on $\mathbb{R}^d$, the OT problem between them with cost function c is defined as

$$\mathsf{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\pi(x,y) \qquad (1)$$

where $\Pi(\mu,\nu)$ is the set of couplings between μ and ν. The special case is the p-Wasserstein distance for $p \in [1,\infty)$, is given by $\mathsf{W}_p(\mu,\nu) := \left(\mathsf{OT}_{\|\cdot\|^p}(\mu,\nu)\right)^{1/p}$. The Wasserstein distance has found applications in various fields, encompassing machine learning [2]–[4], statistics [5]–[7], and applied mathematics [8], [9]. This widespread applicability is driven by an array of desirable properties that the Wasserstein distance possesses, including its metric structure (W_p metrizes weak convergence plus convergence of p-th moments), a convenient dual form, robustness to support mismatch, and a rich geometry it induces on a space of probability measures.

Despite the aforementioned empirical progress, the OT problem suffers from the statistical and computational hardness issues. The estimation rate of the OT cost between distributions on $\mathbb{R}^d$ is generally $n^{-1/d}$ (without further assumptions) [10], which deteriorates exponentially with dimensions—a phenomenon known as the curse of dimensionality. Computationally, OT is a linear program (LP), solvable in $O(n^3\log(n))$ time for distribution on n points using interior point methods or min cost flow algorithms [11]. However, as statistical considerations mandate n to scale exponentially with d to get accurate estimates, the LP computational paradigm becomes infeasible when dimension is large. To circumvent these issues, entropic regularization has emerged as a popular alternative [12]

$$\mathsf{OT}_c^\varepsilon(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c\,d\pi + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi\|\mu\otimes\nu), \qquad (2)$$

where D_{KL} is the Kullback-Leibler divergence and $\varepsilon > 0$ is a regularization parameter. Empirical estimation of EOT enjoys the parametric $n^{-1/2}$ convergence rate in arbitrary dimension, under several settings [13], [14]. Computationally, EOT between discrete distributions can be efficiently solved via the Sinkhorn algorithm [12] in $O(n^2)$ time. However, even this quadratic time complexity is prohibitive when dealing with large and high-dimensional datasets that appear in modern machine learning tasks. Motivated to scale up EOT to such regimes, this work develops a novel neural estimation approach that is end-to-end trainable via backpropagation, compatible with minibatch-based optimization, and adheres to strong performance guarantees.

A. Contributions

We focus on the nominal case of the quadratic EOT distance, i.e., $c(x,y) = \frac{1}{2}\|x-y\|^2$. Thanks to the EOT semi-dual form, we have

$$\mathsf{OT}_c^\varepsilon(\mu,\nu) = \sup_{\varphi\in L^1(\mu)} \int_{\mathbb{R}^d} \varphi\,d\mu + \int_{\mathbb{R}^d} \varphi^{c,\varepsilon} d\,\nu, \qquad (3)$$

where $\varphi^{c,\varepsilon}$ is (c,ε)-transform of φ with respect to (w.r.t.) the cost function. We study regularity of optimal dual potentials φ and show that they belong to a Hölder class of arbitrary smoothness. Leveraging this, we define our neural estimator (NE) by parametrizing the dual potential using a neural network (NN), approximating expectations by sample means, and optimizing the resulting empirical objective over the NN parameters. Our approach yields not only an estimate of the EOT distance, but also a neural EOT plan that is induced by the learned NN. As the estimator is trainable via gradient methods using backpropagation and minibatches, it can seamlessly integrated into downstream tasks as a loss, a regularizer, or a discrepancy quantification module.

We provide formal guarantees on the quality of the NE of the EOT cost and the corresponding transportation plan.

Our analysis relies on non-asymptotic function approximation theorems and tools from empirical process theory to bound the two sources of error involved: function approximation and empirical estimation. Given n samples from the population distributions, we show that the effective error of a NE realized by a shallow NN of k neurons scales as

$$O\left(\mathrm{poly}(1/\varepsilon)\left(k^{-1/2}+n^{-1/2}\right)\right) \tag{4}$$

with the polynomial dependence on $1/\varepsilon$ explicitly characterized. This bound on the EOT cost estimation error holds for arbitrary, compactly supported distributions. This stands in struck contrast to existing neural estimation error bounds for other divergences [15]–[18], which typically require strong regularity assumptions on the population distributions (e.g., Hölder smoothness of densities). This is unnecessary in our setting thanks to the inherit regularity of dual EOT potentials for smooth cost functions, such as our quadratic cost.

The above bound reveals the optimal scaling of the NN and dataset sizes, namely $k \asymp n$, which achieves the parametric convergence rate of $n^{-1/2}$ and guarantees minimax-rate optimality of our NE. The explicitly characterized polynomial dependence on ε in our bound matches the bounds for EOT estimation via empirical plug-in [14], [19]. We also note that our neural estimation results readily extend to the EOT problem with general smooth cost functions. The developed NE is empirically tested on synthetic data, demonstrating its scalability to high dimensions and validating our theory.

B. Related Literature

Neural estimation is a popular approach for enhancing scalability. Prior research explored the tradeoffs between approximation and estimation errors in non-parametric regression [20]–[22] and density estimation [23], [24] tasks. More recently, neural estimation of statistical divergences and information measures has been gaining attention. The mutual information NE (MINE) was proposed in [25], and has seen various improvements since [26]–[29]. Extensions of the neural estimation approach to directed information were studied in [30]–[32]. Theoretical guarantees for f-divergence NEs, accounting for approximation and estimation errors, as we do here, were developed in [16], [17] (see also [15] for a related approach based on reproducing kernel Hilbert space parameterization). Neural estimation of the Stein discrepancy and the minimum mean squared error were considered in [33] and [34], respectively. Neural methods for approximate computation of the Wasserstein distances have been considered under the Wasserstein generative adversarial network (GAN) framework [2], [35], although these approaches are heuristic and lack formal guarantees. Utilizing entropic regularization, [36] studied a score-based generative neural EOT model, while an energy-based model was considered in [37].

II. Background and Preliminaries

A. Notation

Let $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ designate Euclidean norm and the inner product in $\mathbb{R}^d$, respectively. For $1 \le p < \infty$, the L^p space over $\mathcal{X} \subseteq \mathbb{R}^d$ with respect to (w.r.t.) the measure μ is denoted by $L^p(\mu)$, with $\|f\|_{p,\mu} \coloneqq \left(\int_{\mathcal{X}} |f|^p d\mu\right)^{1/p}$ representing the norm. We use $\|\cdot\|_{\infty,\mathcal{X}}$ for standard sup-norm on $\mathcal{X} \subseteq \mathbb{R}^d$ (i.e., when $p = \infty$). Slightly abusing notation, we also set $\|\mathcal{X}\| \coloneqq \sup_{x\in\mathcal{X}} \|x\|_\infty$. The class of Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{X})$. For $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with $\mu \ll \nu$, i.e., μ is absolutely continuous w.r.t. ν, we use $\frac{d\mu}{d\nu}$ for the Radon-Nikodym derivative of μ w.r.t. ν. The subset of probability measures that are absolutely continuous w.r.t. the Lebesgue measure is denoted by $\mathcal{P}_{\mathsf{ac}}(\mathcal{X})$. We use $\lesssim_x$ to denote inequalities up to constants that only depend on x; the subscript is dropped when the constant is universal. For $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

Some additional notation used for our derivations are as follows. For any multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\alpha| = \sum_{j=1}^d \alpha_j$ ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$), define the differential operator $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$ with $D^0 f = f$. We write $N(\delta, \mathcal{F}, \mathsf{d})$ for the δ-covering number of a function class $\mathcal{F}$ w.r.t. a metric d, and $N_{[]}(\delta, \mathcal{F}, \mathsf{d})$ for the bracketing number. For an open set $\mathcal{U} \subseteq \mathbb{R}^d$, $b \ge 0$, and an integer $m \ge 0$, let $\mathcal{C}_b^m(\mathcal{U}) \coloneqq \left\{f \in \mathcal{C}^m(\mathcal{U}) : \max_{\alpha:|\alpha|\le m} \|D^\alpha f\|_{\infty,\mathcal{U}} \le b\right\}$ denote the Hölder space of smoothness index m and radius b. The restriction of $f : \mathbb{R}^d \to \mathbb{R}$ to a subset $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $f\big|_{\mathcal{X}}$.

B. Entropic Optimal Transport

We briefly review basic definitions and results concerning EOT problems. Let $\mathcal{X} \subseteq \mathbb{R}^d$, given distributions $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the primal EOT formulation is obtained by regularizing the OT cost by the KL divergence,

$$\mathsf{OT}_c^\varepsilon(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\, d\pi + \varepsilon \mathsf{D}_{\mathsf{KL}}(\pi \| \mu \otimes \nu), \tag{5}$$

where $\varepsilon > 0$ is a regularization parameter and $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \coloneqq \int \log\left(\frac{d\mu}{d\nu}\right) d\mu$ if $\mu \ll \nu$ and $+\infty$ otherwise. Classical OT [1] is obtained from (5) by setting $\varepsilon = 0$. When $c \in L^1(\mu \otimes \nu)$, EOT admits the dual and semi-dual formulations, which are, respectively, given by

$$\begin{aligned}\mathsf{OT}_c^\varepsilon(\mu, \nu) = &\sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)} \int_{\mathcal{X}} \varphi d\mu + \int_{\mathcal{Y}} \psi d\nu \\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{\varphi\oplus\psi - c}{\varepsilon}} d\mu \otimes \nu + \varepsilon, \qquad (6)\\ = &\sup_{\varphi\in L^1(\mu)} \int_{\mathcal{X}} \varphi d\mu + \int_{\mathcal{Y}} \varphi^{c,\varepsilon} d\nu, \qquad (7)\end{aligned}$$

where we have defined $(\varphi \oplus \psi)(x, y) = \varphi(x) + \psi(y)$ and the (c, ε)-transform of φ is given by

$$\varphi^{c,\varepsilon} = -\varepsilon \log\left(\int_{\mathcal{X}} \exp\left(\frac{\varphi(x) - c(x,\cdot)}{\varepsilon}\right) d\mu(x)\right).$$

There exist functions (φ, ψ) that achieve the supremum in (6), which we call *EOT potentials*. These potentials are almost surely (a.s.) unique up to additive constants, i.e., if $(\tilde{\varphi}, \tilde{\psi})$ is

another pair of EOT potentials, then there exists a constant $a \in \mathbb{R}$ such that $\tilde{\varphi} = \varphi + a$ μ-a.s. and $\tilde{\psi} = \psi - a$ ν-a.s.

A pair $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ are EOT potentials if and only if they satisfy the Schrödinger system

$$\begin{aligned} \int_{\mathcal{X}} e^{\frac{\varphi(x)+\psi(\cdot)-c(x,\cdot)}{\varepsilon}} d\mu(x) = 1 \quad \nu\text{-a.s.} \\ \int_{\mathcal{Y}} e^{\frac{\varphi(\cdot)+\psi(y)-c(\cdot,y)}{\varepsilon}} d\nu(y) = 1 \quad \mu\text{-a.s.} \end{aligned} \tag{8}$$

Furthermore, φ solves the semi-dual from (7) if an only if $(\varphi, \varphi^{c,\varepsilon})$ is a solution to the full dual in (6). Given EOT potentials (φ, ψ), the unique EOT plan can be expressed in their terms as

$$d\pi^{\varepsilon}_{\star} = e^{\frac{\varphi \oplus \psi - c}{\varepsilon}} d\mu \otimes \nu.$$

Subject to smoothness assumptions on the cost function and the population distributions, various regularity properties of EOT potentials can be derived; cf., e.g., [38, Lemma 1].

III. Neural Estimation of EOT Cost and Plan

We consider compactly supported distributions $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and the quadratic cost function $c(x, y) = \frac{1}{2}\|x - y\|^2$ (henceforth dropping the subscript c). For simplicity, further assume that $\mathcal{X}, \mathcal{Y} \subseteq [-1, 1]^d$, although our results readily extend to arbitrary compact supports. We next describe the NE for the EOT distance and plan, followed by non-asymptotic performance guarantees for both objects. All proofs are deferred to the supplement [39].

A. EOT Neural Estimator

For $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, let $X^n := (X_1, \cdots, X_n)$ and $Y^n := (Y_1, \cdots, Y_n)$ be n independently and identically distributed (i.i.d.) samples from μ and ν, respectively. Further suppose that the sample sets are independent of each other. Denote the empirical measures induced by these samples as $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$.

Our NE is realized by a shallow ReLU NN (i.e., a single hidden layer) with k neurons, which defines the function class

$$\mathcal{F}_{k,d}(a) := \left\{ \begin{array}{l} f : \mathbb{R}^d \to \mathbb{R} : \\ f(x) = \sum_{i=1}^{k} \beta_i \phi\left(\langle w_i, x \rangle + b_i\right) + \langle w_0, x \rangle + b_0, \\ \max_{1 \le i \le k} |\beta_i| \le 2ak^{-1}, \ \|w_0\|_1 \le a \\ |b_0| \le a, \ \max_{1 \le i \le k} \|w_i\|_1 \vee |b_i| \le 1 \end{array} \right\}, \tag{9}$$

where $a \in \mathbb{R}_{\ge 0}$ specifies the parameter bounds and $\phi : \mathbb{R} \to \mathbb{R}_{\ge 0} : z \mapsto z \vee 0$ is the ReLU activation function, which acts on vectors component-wise.

We parametrize the semi-dual form of $\mathsf{OT}^{\varepsilon}(\mu, \nu)$ (see (7)) using a NN from the class $\mathcal{F}_{k,d}(a)$ and replace expectations with sample means. Specifically, the EOT distance NE is

$$\begin{aligned} \widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n, Y^n) := \max_{f \in \mathcal{F}_{k,d}(a)} \frac{1}{n} \sum_{i=1}^{n} f(X_i) \\ - \frac{\varepsilon}{n} \sum_{j=1}^{n} \log \left(\frac{1}{n} \sum_{i=1}^{n} \exp \left(\frac{f(X_i) - \frac{1}{2}\|X_i - Y_j\|^2}{\varepsilon} \right) \right). \end{aligned} \tag{10}$$

For any NN $f \in \mathcal{F}_{k,d}(a)$, we define the induced neural plan

$$d\pi^{\varepsilon}_f(x, y) := \frac{\exp\left(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right)}{\int_{\mathcal{X}} \exp\left(\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right) d\mu(x)} d\mu \otimes \nu(x, y). \tag{11}$$

Upon computing the NE in (10), the neural plan $d\pi^{\varepsilon}_{f_\star}$ induced by an optimal NN $f_\star \in \mathcal{F}_{k,d}(a)$ serves as an estimate of the true optimal plan $\pi^{\varepsilon}_\star$ that achieves the infimum in (5).

B. Performance Guarantees

We provide formal guarantees for the neural estimator of the EGW cost and the neural transportation plan defined above. Starting from the cost estimation setting, we establish two separate bounds on the effective (approximation plus estimation) error. The first is non-asymptotic and presents optimal convergence rates, but calibrates the NN parameters to a cumbersome dimension-dependent constant. Following that, we present an alternative bound that avoids the dependence on the implicit constant, at the expense of a polylogarithmic slow-down in the rate and a requirement that the NN size k is large enough.

Theorem 1 (EOT cost neural estimation; bound 1)**.** *There exists a constant $C > 0$ depending only on d, such that setting $a = C(1 + \varepsilon^{1-s})$ with $s = \lfloor d/2 \rfloor + 3$, we have*

$$\begin{aligned} &\sup_{(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})} \mathbb{E}\left[\left| \widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n, Y^n) - \mathsf{OT}^{\varepsilon}(\mu, \nu) \right| \right] \\ &\lesssim_d \left(1 + \frac{1}{\varepsilon^{\lfloor \frac{d}{2} \rfloor + 2}}\right) k^{-\frac{1}{2}} \\ &\quad + \min\left\{ 1 + \frac{1}{\varepsilon^{\lceil \frac{3d}{2} \rceil + 4}}, \left(1 + \frac{1}{\varepsilon^{\lfloor \frac{d}{2} \rfloor + 2}}\right) \sqrt{k} \right\} n^{-\frac{1}{2}}. \end{aligned} \tag{12}$$

The proof of Theorem 1 is given in Appendix A-A. We establish regularity of semi-dual EOT potentials (namely, $(\varphi, \varphi^{c,\varepsilon})$ in (7)), showing that they belong to a Hölder class of arbitrary smoothness. This, in turn, allows accurately approximating these dual potentials by NNs from the class $\mathcal{F}_{k,d}(a)$ with error $O(k^{-1/2})$, yielding the first term in the bound. To control the estimation error, we employ standard maximal inequalities from empirical process theory along with a bound on the covering or bracketing number of (c, ε)-transform of the NN class. The resulting empirical estimation error bound comprises the second term on the RHS above.

Remark 1 (Almost explicit expression for C)**.** *The expression of the constant C in Theorem 1 is cumbersome, but can*

nonetheless be evaluated. Indeed, one may express $C = C_s C_d \bar{c}_d$*, with explicit expressions for* C_d *and* $\bar{c}_d$ *given in* (19) *and* (22)*, respectively, while* C_s *is a combinatorial constant that arises from the multivariate Faa di Bruno formula (cf.* (36)-(37)*). The latter constant is quite convoluted and is the main reason we view* C *as implicit.*

Our next bound circumvents the dependence on C by letting the NN parameters grow with its size k. This bound, however, requires k to be large enough and entails additional polylog factors in the rate. The proof is similar to that of Theorem 1 and given in Appendix A-B.

Theorem 2 (EOT cost neural estimation; bound 2)**.** *Let* $\epsilon > 0$ *and set* $m_k = \log k \vee 1$*. Assuming* k *is sufficiently large, we have*

$$\sup_{(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} \mathbb{E}\left[\left|\widehat{\mathsf{OT}}^{\varepsilon}_{k,m_k}(X^n,Y^n) - \mathsf{OT}^{\varepsilon}(\mu,\nu)\right|\right] \lesssim_d \left(1+\frac{1}{\varepsilon^{\lfloor \frac{d}{2}\rfloor+2}}\right) k^{-\frac{1}{2}} + \min\left\{\left(1+\frac{1}{\varepsilon^{\lfloor \frac{d}{2}\rfloor}}\right)(\log k)^2, \sqrt{k}\log k\right\} n^{-\frac{1}{2}}. \tag{13}$$

Remark 2 (NN size)**.** *We can provide a partial account of the requirement that* k *is large enough. Specifically, for the bound to hold we need* k *to satisfy* $\log k \geq C(1+\varepsilon^{1-s})$*, where* C *is the constant from Theorem 1. It is, however, challenging to quantify the exact threshold on* k *required for the theorem to hold due to the implicit nature of* C*.*

Lastly, we move to account for the quality of the neural plan that is induced by the EOT NE (see (11)) by comparing it, in KL divergence, to the true EOT plan $\pi^{\varepsilon}_{\star}$.

Theorem 3 (EOT alignment plan neural estimation)**.** *Suppose that* $\mu \in \mathcal{P}_{\mathsf{ac}}(\mathcal{X})$*. Let* $\hat{f}_\star$ *be a maximizer of* $\widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n,Y^n)$ *from* (10)*, with* a *as defined in Theorem 1. Then, the induced neural plan* $\pi^{\varepsilon}_{\hat{f}_\star}$ *from* (11) *satisfies*

$$\mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(\pi^{\varepsilon}_{\star}\middle\|\pi^{\varepsilon}_{\hat{f}_\star}\right)\right] \lesssim_d \left(1+\frac{1}{\varepsilon^{\lfloor \frac{d}{2}\rfloor+3}}\right) k^{-\frac{1}{2}} + \min\left\{1+\frac{1}{\varepsilon^{\lceil \frac{3d}{2}\rceil+5}}, \left(1+\frac{1}{\varepsilon^{\lfloor \frac{d}{2}\rfloor+3}}\right)\sqrt{k}\right\} n^{-\frac{1}{2}}. \tag{14}$$

where π^{ε}_{*} *is optimal coupling of EOT problem* (5)*.*

Theorem 3 is proved in Appendix A-C. The key step in the derivation shows that the KL divergence between the alignment plans, in fact, equals the gap between the EOT cost $\mathsf{OT}^{\varepsilon}$ and its neural estimate from (10), up to a multiplicative ε factor. Having that, the result follows by invoking Theorem 1.

Remark 3 (Extension to Sigmoidal NNs)**.** *The results of this section readily extend to cover sigmoidal NNs, with a slight modification of some parameters. Specifically, one has to replace* s *from Theorem 1 with* $\tilde{s} = \lfloor d/2 \rfloor + 2$ *and consider the sigmoidal NN class, with nonlinearity* $\psi(z) = (1+e^{-z})^{-1}$ *(instead of ReLU) and parameters satisfying*

$$\max_{1\leq i\leq k}\|w_i\|_1 \vee |b_i| \leq k^{\frac{1}{2}}\log k, \ \max_{1\leq i\leq k}|\beta_i| \leq 2ak^{-1}, \ |b_0| \leq a, \ \|w_0\|_1 = 0.$$

The proofs of Theorems 1-3 then go through using the second part of Proposition 10 from [17], which relies on controlling the so-called Barron coefficient (cf. [40]–[42]).

Remark 4 (Convergence rates of Sinkhorn's algorithm)**.** *Neural estimation is proposed as a more scalable alternative to Sinkhorn's algorithm for computing the EOT cost/plan, e.g., by enabling the usage of mini-batches. We comment here on the rate of convergence that the Sinkhorn-based approach achieves. Denote the output of the Sinkhorn algorithm running on empirical measures, each over* n *samples, by* $\widetilde{\mathsf{OT}}^{\varepsilon}(X^n,Y^n)$*. The effective error can be decomposed as:*

$$|\widetilde{\mathsf{OT}}^{\varepsilon}(X^n,Y^n) - \mathsf{OT}(\mu,\nu)| \leq |\mathsf{OT}(\mu,\nu) - \mathsf{OT}(\mu_n,\nu_n)| + |\mathsf{OT}(\mu_n,\nu_n) - \widetilde{\mathsf{OT}}^{\varepsilon}(\mu_n,\nu_n)|,$$

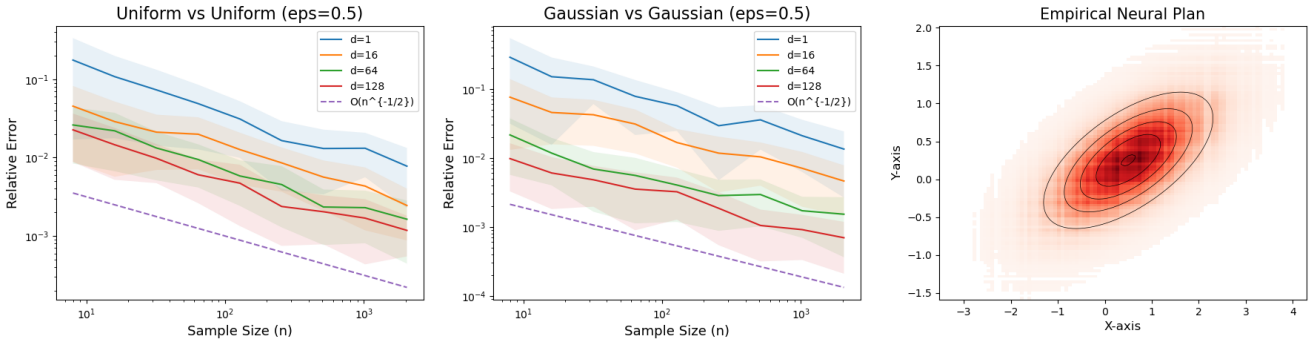
where first term decays as $O(n^{-\frac{1}{2}})$ *[14], while the second exhibits a convergence rate of* $o_P(n^{-1})$ *within* $o_P(\log(n\log(n)))$ *iterations [43].*

IV. Numerical Experiments

This section illustrates the performance of the EOT distance neural estimator via experiments with synthetic data. Specifically, we compute the estimate from (10) under various settings, allowing a to be unrestricted so as to enable optimization over the whole parameter space. We train the parameters of the ReLU network using the Adam algorithm [44]. We use an epoch number of 20, learning rate 10^{-3} and choose a best batch size from $\{2,4,8,16,32,64,128\}$. We test our EOT distance neural estimator by estimating the EOT cost and optimal plan between uniform and Gaussian distribution in different dimensions. We consider dimensions $d \in \{1,16,64,128\}$, and for each d, employ a ReLU network of size $k \in \{16,64,128,256\}$, respectively. Accuracy is measured using the relative error $\left|\widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n,Y^n) - \widetilde{\mathsf{OT}}^{\varepsilon}(\mu,\nu)\right| / \widetilde{\mathsf{OT}}^{\varepsilon}(\mu,\nu)$, where $\widetilde{\mathsf{OT}}^{\varepsilon}(\mu,\nu)$ is regarded as the ground truth, which we obtain by running Sinkhorn algorithm [12] with $n = 10,000$ samples (which we treat as $n\to\infty$ as it is $\times 5$ more than the largest sample set we use for our neural estimator). Each of the presented plots is averaged over 20 runs.

We first consider the EOT distance with $\varepsilon = 0.5$ between two uniform distribution over a hypercube, namely, $\mu = \nu = \mathrm{Unif}\big([-1/\sqrt{d}, 1/\sqrt{d}]^d\big)$. Figure 1a plots the EOT neural estimation error versus the sample size $n \in \{8,16,32,64,128,256,512,1024,2048\}$ in a log-log scale. The curves exhibit a slope of approximately $-1/2$ for all dimensions, which validates our theory. Notably, this rate is uniform across dimensions, like the bounds from Theorems 1 and 2 suggest.

Next, we test the EOT NE on unbounded measures. To that end, we set $\varepsilon = 0.5$ and take μ, ν as d-dimensional Gaussian



(a) (b) (c)

Fig. 1: Neural Estimation of EOT distance: (a) Relative error for the case where $\mu = \nu = \mathrm{Unif}\big([-1/\sqrt{d}, 1/\sqrt{d}]^d\big)$; (b) Relative error for μ, ν as Gaussian distributions with randomly generated mean vectors and covariance matrices; (c) Learned neural plan (in red) versus the true optimal EOT optimal plan (whose density is represented by the back contour lines).

distributions with randomly generated mean vectors and covariance matrices. Specifically, the mean vectors are randomly sampled from d-dimensional standard Gaussian, while the two covariance matrices are of the form $\mathbf{B}^\intercal\mathbf{B}+1/(3d)\mathbf{I}_d$, where $\mathbf{I}_d$ is a $d\times d$ identity and $\mathbf{B}$ is a matrix whose entries are randomly sampled from $\mathrm{Unif}([-1/d, 1/d])$. Note that the generated covariance matrix is positive semi-definite with eigenvalues set to lie in $[\frac{1}{3d}, \frac{1}{d}]$. Figure 1b plots the relative EOT neural estimation error for this Gaussian setting, again showing a parametric convergence rate for all considered dimensions.

Lastly, we assess the quality of the neural plan learned from our NE. Since doing so requires knowledge of the true (population) optimal plan $\pi^\varepsilon_\star$, we consider the EOT distance between Gaussians, for which a closed form expression for the optimal plan was derived in [45]. We take $\varepsilon = 0.5$, $\mu = \mathcal{N}(0.5, 1)$, and $\nu = \mathcal{N}(0.25, 0.25)$. By Theorem 3.1 of [46], the optimal EOT plan is given by

$$\pi^\varepsilon_\star \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ 0.25 \end{pmatrix}, \begin{pmatrix} 1 & \frac{\sqrt{5}-1}{4} \\ \frac{\sqrt{5}-1}{4} & 0.25 \end{pmatrix}\right)$$

Figure 1c compares the neural coupling learned from our algorithm, shown in red, to the optimal $\pi^\varepsilon_\star$ given above, whose density is represented by the black contour lines. The neural coupling is learned using $n = 10^4$ samples and is realized by a NN with $k = 32$ neurons. There is a clear correspondence between the two, which supports the result of Theorem 3.

V. Conclusion

This work proposed a novel neural estimation technique for the EOT distance with quadratic costs between Euclidean mm spaces. The estimator leveraged the semi-dual formulation of EOT. Our approach yielded estimates not only for the EOT distance value but also for the optimal plan. Non-asymptotic formal guarantees on the quality of the NE were provided, under the sole assumption of compactly supported population distributions, with no further regularity conditions imposed. Our bounds revealed optimal scaling laws for the NN and the dataset sizes that ensure parametric (and hence minimax-rate optimal) convergence. The proposed estimator was tested via numerical experiments on synthetic data, demonstrating its accuracy, scalability, and fast convergence rates that match the derived theory.

Future research directions stemming from this work are abundant. First, our theory currently accounts for NEs realized by shallow NNs, but deep nets are oftentimes preferable in practice. Extending our results to deep NNs should be possible by utilizing existing function approximation error bounds [47], although these bounds may not be sharp enough to yield the parametric rate of convergence. Another limitation of our analysis is that it requires compactly supported distributions. It is possible to extend our results to distributions with unbounded supports using the technique from [17] that considers a sequence of restrictions to balls of increasing radii. Unfortunately, as in [17], rate bounds obtained from this technique would be sub-optimal. Obtaining sharp rates for the unboundedly supported case would require new ideas and forms an interesting research direction. Lastly, while EOT serves as an important approximation of OT, neural estimation of the OT distance itself is a challenging and appealing research avenue. One may attempt to directly approximate this objective by NNs, but dual OT potential generally lack sufficient regularity to allow quantitative approximation bounds. Assuming smoothness of the population distributions, and employing estimators that adapt to this smoothness, e.g., based on kernel density estimators or wavelets [48], [49], may enable deriving sharp rates of convergence.

Acknowledgment

Z. Goldfeld is partially supported by NSF grants CCF-2046018, DMS-2210368, and CCF-2308446, and the IBM Academic Award.

References

[1] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, Jul. 2017, pp. 214–223.

[3] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," *arXiv preprint arXiv:1711.01558*, 2017.

[4] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," *Advances in neural information processing systems*, vol. 30, 2017.

[5] G. Carlier, V. Chernozhukov, and A. Galichon, "Vector quantile regression: an optimal transport approach," 2016.

[6] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry, "Monge–kantorovich depth, quantiles, ranks and signs," 2017.

[7] P. Ghosal and B. Sen, "Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing," *The Annals of Statistics*, vol. 50, no. 2, pp. 1012–1037, 2022.

[8] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the fokker–planck equation," *SIAM journal on mathematical analysis*, vol. 29, no. 1, pp. 1–17, 1998.

[9] F. Santambrogio, "{Euclidean, metric, and Wasserstein} gradient flows: an overview," *Bulletin of Mathematical Sciences*, vol. 7, pp. 87–154, 2017.

[10] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability theory and related fields*, vol. 162, no. 3-4, pp. 707–738, 2015.

[11] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," *Center for Research in Economics and Statistics Working Papers*, no. 2017-86, 2017.

[12] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[13] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, "Sample complexity of sinkhorn divergences," in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 1574–1583.

[14] G. Mena and J. Niles-Weed, "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[16] Z. Z. S. Sreekumar and Z. Goldfeld, "Non-asymptotic performance guarantees for neural estimation of f-divergences," in *International Conference on Artificial Intelligence and Statistics (AISTATS-2021)*, ser. Proceedings of Machine Learning Research, vol. 130, Virtual conference, April 2021, pp. 3322–3330.

[17] S. Sreekumar and Z. Goldfeld, "Neural estimation of statistical divergences," *Journal of Machine Learning Research*, vol. 23, no. 126, pp. 1–75, 2022.

[18] D. Tsur, Z. Goldfeld, and K. Greenewald, "Max-sliced mutual information," *arXiv preprint arXiv:2309.16200*, 2023.

[19] M. Groppe and S. Hundrieser, "Lower complexity adaptation for empirical entropic optimal transport," *arXiv preprint arXiv:2306.13580*, 2023.

[20] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine learning*, vol. 14, pp. 115–133, 1994.

[21] F. Bach, "Breaking the curse of dimensionality with convex neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 629–681, 2017.

[22] T. Suzuki, "Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality," *arXiv preprint arXiv:1810.08033*, 2018.

[23] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, pp. 1564–1599, 1999.

[24] A. Uppal, S. Singh, and B. Póczos, "Nonparametric density estimation & convergence rates for gans under besov ipm losses," *Advances in neural information processing systems*, vol. 32, 2019.

[25] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[26] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On variational lower bounds of mutual information," in *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[27] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," in *International Conference on Learning Representations*, 2019.

[28] C. Chan, A. Al-Bashabsheh, H. P. Huang, M. Lim, D. S. H. Tam, and C. Zhao, "Neural entropic estimation: A faster path to mutual information estimation," *arXiv preprint arXiv:1905.12957*, 2019.

[29] Y. Mroueh, I. Melnyk, P. Dognin, J. Ross, and T. Sercu, "Improved mutual information estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9009–9017.

[30] S. Molavipour, H. Ghourchian, G. Bassi, and M. Skoglund, "Neural estimator of information for time-series data with dependency," *Entropy*, vol. 23, no. 6, p. 641, 2021.

[31] D. Tsur, Z. Aharoni, Z. Goldfeld, and H. Permuter, "Neural estimation and optimization of directed information over continuous spaces," *IEEE Transactions on Information Theory*, 2023.

[32] ——, "Data-driven optimization of directed information over discrete alphabets," *Accepted to the IEEE Transactions on Information theory*, November 2023.

[33] M. Repasky, X. Cheng, and Y. Xie, "Neural Stein critics with staged l^2-regularization," *IEEE Transactions on Information Theory*, 2023.

[34] M. Diaz, P. Kairouz, and L. Sankar, "Lower bounds for the MMSE via neural network estimation and their applications to privacy," *arXiv preprint arXiv:2108.12851*, 2021.

[35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS-2017)*, Long Beach, CA, US, Dec. 2017, pp. 5767–5777.

[36] M. Daniels, T. Maunu, and P. Hand, "Score-based generative neural networks for large-scale optimal transport," *Advances in neural information processing systems*, vol. 34, pp. 12 955–12 965, 2021.

[37] P. Mokrov, A. Korotin, and E. Burnaev, "Energy-guided entropic neural optimal transport," *arXiv preprint arXiv:2304.06094*, 2023.

[38] Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu, "Limit theorems for entropic optimal transport maps and the sinkhorn divergence," *arXiv preprint arXiv:2207.08683*, 2022.

[39] T. Wang and Z. Goldfeld, "Neural estimation of entropic optimal transport," 2024. [Online]. Available: https://github.com/TaoWangPenn/Supplement-ISIT-2024/blob/main/supplement(ISIT_2024).pdf

[40] A. R. Barron, "Neural net approximation," in *Proc. 7th Yale workshop on adaptive and learning systems*, vol. 1, 1992, pp. 69–72.

[41] ——, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information theory*, vol. 39, no. 3, pp. 930–945, 1993.

[42] J. E. Yukich, M. B. Stinchcombe, and H. White, "Sup-norm approximation bounds for networks through probabilistic methods," *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 1021–1027, 1995.

[43] Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu, "Limit theorems for entropic optimal transport maps and sinkhorn divergence," *Electronic Journal of Statistics*, vol. 18, no. 1, pp. 980–1041, 2024.

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

[45] H. Janati, B. Muzellec, G. Peyré, and M. Cuturi, "Entropic optimal transport between unbalanced gaussian measures has a closed form," *Advances in neural information processing systems*, vol. 33, pp. 10 468–10 479, 2020.

[46] K. Le, D. Q. Le, H. Nguyen, D. Do, T. Pham, and N. Ho, "Entropic gromov-wasserstein between gaussian distributions," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 164–12 203.

[47] G. Bresler and D. Nagaraj, "Sharp representation theorems for relu networks with precise dependence on depth," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 697–10 706, 2020.

[48] N. Deb, P. Ghosal, and B. Sen, "Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 736–29 753, 2021.

[49] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman, "Plug-in estimation of smooth optimal transport maps," *arXiv preprint arXiv:2107.12364*, 2021.

[50] Z. Zhang, Z. Goldfeld, Y. Mroueh, and B. K. Sriperumbudur, "Gromov-wasserstein distances: Entropic regularization, duality, and sample complexity," *arXiv preprint arXiv:2212.12848*, 2022.

[51] A. W. van der Vaart and J. A. Wellner, "Springer series in statistics," *Weak convergence and empirical processesSpringer, New York*, 1996.

[52] G. Constantine and T. Savits, "A multivariate faa di bruno formula with applications," *Transactions of the American Mathematical Society*, vol. 348, no. 2, pp. 503–520, 1996.