Convergence of Gaussian-smoothed optimal transport distance with sub-gamma distributions and dependent samples

Yixing Zhang
Duke University

Xiuyuan Cheng
Duke University

Galen Reeves
Duke University

Abstract

The Gaussian-smoothed optimal transport (GOT) framework, recently proposed by Goldfeld et al., scales to high dimensions in estimation and provides an alternative to entropy regularization. This paper provides convergence guarantees for estimating the GOT distance under more general settings. For the Gaussian-smoothed p-Wasserstein distance in d dimensions, our results require only the existence of a moment greater than $d + 2p$. For the special case of sub-gamma distributions, we quantify the dependence on the dimension d and establish a phase transition with respect to the scale parameter. We also prove convergence for dependent samples, only requiring a condition on the pairwise dependence of the samples measured by the covariance of the feature map of a kernel space.

A key step in our analysis is to show that the GOT distance is dominated by a family of kernel maximum mean discrepancy (MMD) distances with a kernel that depends on the cost function as well as the amount of Gaussian smoothing. This insight provides further interpretability for the GOT framework and also introduces a class of kernel MMD distances with desirable properties. The theoretical results are supported by numerical experiments.

1 Introduction

There has been significant interest in optimal transport (OT) distances for data analysis, motivated by applications in statistics and machine learning ranging from computer graphics and imaging processing (Li et al., 2018; Ryu et al., 2018; Solomon et al., 2014) to deep learning (Bhushan Damodaran et al., 2018; Courty et al., 2016; Shen et al., 2017); see (Peyré and Cuturi, 2019). The OT cost between probability measures P and Q with cost function $c(x, y)$ is defined as

$$\mathcal{T}(P, Q) := \inf_{\pi \in \Pi(P,Q)} \int c(x, y) \, d\pi(x, y), \qquad (1)$$

where $\Pi(P, Q)$ is the set of all probability measures whose marginals are P and Q. Of central importance to applications in statistics and machine learning is the rate at which the empirical measure P_n of and iid sample approximates the true underlying distribution P. In this regard, one of the main challenges for OT distances is that rate convergence suffers from the curse of dimensionality: the number of samples n needs to grow exponentially with the dimension d of the data (Fournier and Guillin, 2015).

On a closely related note, OT also suffers from computational issues, particularly in the high-dimensional settings. To address both statistical and computation limitations, recent work has focused on regularized versions of OT including entropy regularization (Cuturi, 2013) and Gaussian-smoothed optimal transport (GOT) (Goldfeld et al., 2020b). The entropy-regularized OT has attracted intensive theoretical interest (Bigot et al., 2019; Feydy et al., 2019; Klatt et al., 2020), as well as an abundance of algorithm developments (Abid and Gower, 2018; Chakrabarty and Khanna, 2020; Gerber and Maggioni, 2017). In comparison, GOT is less understood both in theory and in computation. The goal of the current paper is thus to deepen the theoretical analysis of GOT under more general settings, so as to lay a theoretical foundation for computational study and potential applications.

In particular, we consider distributions that satisfy only a bounded moment condition and general settings involving dependent samples. For the special case of sub-gamma distributions, we show a phase transition depending on the dimension d and with respect to the scale parameter of the sub-gamma distribution. Going

beyond the case of iid samples, our convergence rate covers dependent samples as long as a condition on the pair-wise dependence quantified by the covariance of the kernel-space feature map is satisfied. A key step in our analysis is to establish a novel connection between the GOT distance and a family of kernel MMD distances, which can be of independent interest. In the kernel MMD upper bound, the kernel is neither bounded nor translation invariant, and is determined by both the cost function of OT and the amount of Gaussian smoothing. The theoretical findings are supported by numerical experiments.

To summarize our contribution, we provide an overview of the main theoretical results in the next subsection, and then close the introduction with a detailed review of related work. After introducing notations and needed preliminaries in Section 2, we derive upper bounds of GOT using kernel MMD of a new two-moment kernel in Section 3, which leads to the convergence rate results in Section 4 and numerical results in Section 5. All proofs are in the supplementary materials.

1.1 Overview of Main Results

In this paper, we focus on the OT cost $\mathcal{T}_p(P, Q)$ associated with the cost function $c_p(x, y) = \|x - y\|^p$ for $p > 0$ and $c_0(x, y) = 1_{\{x \neq y\}}$. The total variation distance is given by $\mathcal{T}_0(P, Q)$ and the p-Wasserstein distance is given by $\mathcal{T}_p(P, Q)$ if $p \in (0, 1]$ and $(\mathcal{T}_p(P, Q))^{1/p}$ if $p > 1$ (Villani, 2003).

The minimax convergence rate of $\mathcal{T}_p(P, P_n)$ was established by Fournier and Guillin (2015, Theorem 1) who showed that if P has a moment strictly greater than $2p$, then

$$\mathbb{E}[\mathcal{T}_p(P, P_n)] \asymp \begin{cases} n^{-\frac{1}{2}}, & p > d/2 \\ n^{-\frac{1}{2}} \log n, & p = d/2 \\ n^{-\frac{p}{d}}, & p \in (0, d/2) \end{cases} . \quad (2)$$

Unfortunately, this means that the sample complexity increases exponentially with the dimension for $d > 2p$.

Recently, Goldfeld et al. (2020b) showed that one way to overcome the curse of dimensionality is to consider the Gaussian-smoothed OT distance, defined as

$$\mathcal{T}_p^{(\sigma)}(P, Q) := \mathcal{T}_p(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma$ denotes the iid Gaussian measure with mean zero and variance σ^2. Under the assumption that P is sub-Gaussian with constant v, they proved an upper bound on the converge rate that is independent of the dimension:

$$\mathbb{E}[\mathcal{T}_p(P, P_n)] \leq \frac{C_{d,p,\sigma,v}}{\sqrt{n}}, \quad p \in \{0, 1, 2\}.$$

The precise the form of the constant $C_{d,p,\sigma,v}$ is provided for $p \in \{0, 1\}$ but not for the case $p = 2$ unless P is also assumed to have bounded support. Ensuing work by Goldfeld et al. (2020a) established the same convergence rate for $p = 1$ under the relaxed assumption that P has finite moment grater than $2d + 2$.

Metric properties of GOT were studied by Goldfeld and Greenewald (2020) who showed that $\mathcal{T}_1^{(\sigma)}(P, Q)$ is a metric on the space of probability measures with finite first moment and that the sequence of optimal couplings converges in the $\sigma \to 0$ limit to the optimal coupling for the unsmoothed Wasserstein distance. Their arguments depend only on the pointwise convergence of the characteristic functions under Gaussian smoothing, and thus also apply to the case of general p considered in this paper.

One of the main contributions of this paper is to prove an upper bound on the convergence rate for all orders of p and under more general assumptions on P. Specifically, we prove the following result:

Theorem 1. *Let P_n be the empirical measure of n iid samples from a probability measure P on $\mathbb{R}^d$ that satisfies $(\int \|x\|^s dP(x))^{1/s} \leq m$ for some $s > d + 2p$. There exists a positive constant $C_{d,p,s}$ such that for all $\sigma > 0$,*

$$\mathbb{E}\left[\mathcal{T}_p^{(\sigma)}(P, P_n)\right] \leq C_{d,p,s} \frac{\sigma^p}{\sqrt{n}} \left(1 + \frac{m}{\sigma}\right)^{\frac{d}{2}+p}. \quad (3)$$

This result brings the GOT framework in line with the general setting studied by Fournier and Guillin (2015, Theorem 1), and shows that the benefits obtained by smoothing extend beyond the special cases of small p and well-controlled tails. To help interpret this result it is important to keep in mind that for $p > 1$, the Wasserstein distance is given by the p-th root of the GOT. As for the tightness of the bound, there are two regimes worth considering, namely when $\sigma \to 0$ as $n \to \infty$ and when σ is fixed. In the former case, the dependence on σ seems to be nearly tight. In Section 4.1, we show that if $\sigma \asymp n^{-\frac{1}{d+2p}}$ then Theorem 1 implies an upper bound on the *unsmoothed* convergence rate

$$\mathbb{E}[\mathcal{T}_p(P, P_n)] \leq C_{d,p,s} m^p n^{-\frac{p}{d+2p}}. \quad (4)$$

Notice that for $d \ll 2p$ and $d \gg 2p$ this recovers the minimax convergence rate given in (2).

The main technical step in our approach is to establish a novel connection between GOT and a family of kernel MMD distances (Theorem 2). We then show how a particular member of this family, which we call the 'two-moment' kernel, defines a metric on the space of probability measures with finite moments strictly greater than $p + d/2$ (Theorem 3).

In addition to Theorem 1, we also provide further results that elucidate the role of the dimension as well the tail behavior of the underlying distribution (Theorem 6). Furthermore, we address the setting of dependent samples and provide an example illustrating how the connection with MMD can be used to go beyond the usual assumptions involving mixing conditions for stationary processes. Finally, we provide some numerical experiments that support our theory.

1.2 Comparison with Previous Work

The convergence of OT distances continues to be an active area of research (Lei et al., 2020; Niles-Weed and Berthet, 2019; Singh and Póczos, 2018). Building upon the the work of work of Cuturi (2013), a recent line of work has focused on entropy regularized OT defined by

$$S_\epsilon(P,Q) := \inf_{\pi \in \Pi(P,Q)} \int c(x,y)\, d\pi(x,y) + \epsilon D(\pi \| P \otimes Q)$$

where $D(\mu\|\nu) = \int \log \frac{d\mu}{d\nu} d\mu$ is the relative entropy between probability measures μ and ν. The addition of the regularization term facilitates the numerical approximation using the Sinkhorn algorithm. The amount of regularization interpolates between OT in the $\epsilon \to 0$ limit and the kernel MMD in $\epsilon \to \infty$ limit; see Feydy et al. (2019). In contrast to the Gaussian-smoothed Wasserstein distance, entropy regularized OT is not a metric since it does not satisfy the triangle inequality. Convergence rates for entropy regularized OT were obtained by Genevay et al. (2019) under the assumption of bounded support and more recently by Mena and Niles-Weed (2019) under the assumption of sub-Gaussian tails. Further properties have been studied by Luise et al. (2018) and Klatt et al. (2020).

There has also been work focusing on the sliced Wasserstein distance, which is obtained by averaging the one-dimensional Wasserstein distance over the unit sphere (Bonneel et al., 2015; Rabin et al., 2011). While the sliced Wasserstein distance is equivalent to the Wasserstein distance in the sense that convergence in one metric implies convergence in another, the rates of convergence need not be the same. See Section 1.2 of Goldfeld et al. (2020a) for further discussion.

Going beyond convergence rates for empirical measures, properties of smoothed empirical measures have been studied in a variety of contexts, including the high noise limit (Chen and Niles-Weed, 2020) and applications to the estimation of mutual information in deep networks (Goldfeld et al., 2018). Finally, we note that there has also been some work on convergence with dependent samples by Fournier and Guillin (2015), who focus on OT distance, and also by Young and Dunson (2019), who consider a closely related entropy estimation problem.

2 Preliminaries

Let $\mathcal{P}(\mathbb{R}^d)$ be the space of Borel probability measures on $\mathbb{R}^d$ and let $\mathcal{P}_s(\mathbb{R}^d)$ be the space of probability measures with finite s-th moment, i.e, $\int \|x\|^s dP(x) < \infty$. The Gaussian measure on $\mathbb{R}^d$ with mean-zero and covariance $\sigma^2 I_d$ is denoted by $\mathcal{N}_\sigma$. The convolution of probability measures P and Q is denoted by $P * Q$. The Gamma function is given by $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ for $z > 0$. We use C to denote a generic positive real number, and the value of C may change from place to place.

Kernel MMD. A symmetric function $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be a *positive-definite kernel* on $\mathbb{R}^d$ if and only if for every $x_1, \ldots, x_n \in \mathbb{R}^d$, the symmetrix matrix $(k(x_i, x_j))_{i,j=1}^n$ is positive semidefinite. A positive definite kernel k can be used to define a distance on probability measure known as RKHS MMD (Anderson et al., 1994; Gretton et al., 2005, 2012; Smola et al., 2007). Let $\mathcal{P}^k(\mathbb{R}^d)$ be the space of probability measures such that $\int \sqrt{k(x,x)} dP(x) < \infty$. For $P, Q \in \mathcal{P}^k(\mathbb{R}^d)$ the kernel MMD distance is defined as

$$\gamma_k^2(P,Q) = \iint k(x,y)\, d(P(x) - Q(x))\, d(P(y) - Q(y)).$$

The distance $\gamma_k(P,Q)$ is a pseudo-metric in general. A kernel k is said to be *characteristic* to a set $Q \subseteq \mathcal{P}$ of probability measures if and only if γ_k is a metric on $\mathcal{Q}$ (Gretton et al., 2012; Sriperumbudur et al., 2010). An alternative representation is given by

$$\gamma_k^2(P,Q) = \mathbb{E}[k(X,X')] + \mathbb{E}[k(Y,Y')] - 2\mathbb{E}[k(X,Y)] \tag{5}$$

where $X, X' \sim P$ iid, and $Y, Y' \sim Q$ iid. The kernel MMD distance was shown to be equivalent to energy distance in (Sejdinovic et al., 2013), and a variant form used in practice is the kernel mean embedding statistics (Chwialkowski et al., 2015; Jitkrittum et al., 2016; Muandet et al., 2017). Another appealing theoretical property of the kernel MMD distance is its representation via the spectral decomposition of the kernel (Epps and Singleton, 1986; Fernández et al., 2008), which gives rise to estimation consistency as well as practical algorithms of computing (Zhao and Meng, 2015). Kernel MMD has been widely applied in data analysis and machine learning, including independence testing (Fukumizu et al., 2008; Zhang et al., 2012) and generative modeling (Li et al., 2017, 2015).

Magnitude of multivariate Gaussian. Our results also depend on some properties of the noncentral chi distribution. Let $Z \sim \mathcal{N}(\mu, I_d)$ be a Gaussian vector with mean μ and identity covariance. The random variable $X = \|Z\|$ has chi-distribution with parameter

$u = \|\mu\|$. The density is given by

$$g_{d,u}(x) = \frac{e^{-(x^2+u^2)/2}x^{d-1}}{2^{\frac{d}{2}-1}} \sum_{k=0}^{\infty} \frac{(ux/2)^{2k}}{k!\Gamma\left(\frac{d+2k}{2}\right)}. \quad (6)$$

The s-th moment of this distribution is denoted by $M_{d,u}(s) = \int x^s \, g_{d,u}(x) \, dx$. This function is an even polynomial of degree s whenever s is an even integer (see Supplement C.1). The special case $u = 0$ corresponds to the (central) chi distribution and is given by $M_d(s) := 2^{\frac{s}{2}}\Gamma\left(\frac{d+s}{2}\right)/\Gamma\left(\frac{d}{2}\right)$.

3 Upper Bounds on GOT

In this section, we show that GOT is bounded from above by a family of kernel MMD distances. It is assumed throughout that d is a positive integer, $p \in (0, \infty)$, and $\sigma \in (0, \infty)$.

3.1 General Bound via Kernel MMD

Consider the feature map $\psi_x \colon \mathbb{R}^d \to [0, \infty)$ defined by

$$\psi_x(z) := \frac{\sqrt{\omega_d}}{2^{\frac{d+p}{2}}} \frac{\|z\|^{\frac{d-1+2p}{2}}}{\sqrt{f(\|z\|)}} \phi\left(\frac{z}{\sqrt{2}} - \frac{x}{\sigma}\right), \quad (7)$$

where $\omega_d = 2\pi^{d/2}/\Gamma(d/2)$ is the volume of the unit sphere in $\mathbb{R}^d$, $\phi(u) = (2\pi)^{-d/2}\exp(-\frac{1}{2}\|u\|^2)$ is the standard Gaussian density on $\mathbb{R}^d$, and f is a probability density function on $[0, \infty)$ that satisfies

$$f(x) \geq ax^{d+2p-1}\exp(-bx^2), \quad (8)$$

for some $a > 0$ and $b \in (0, 1/2)$. This feature map defines a positive semidefinite kernel k according to

$$k(x, y) = \langle \psi_x, \psi_y \rangle.$$

After some straightforward manipulations (see Supplement A), one finds that $k(x, y)$ is finite on $\mathbb{R}^d \times \mathbb{R}^d$ and can be expressed as the product of a Gaussian kernel and a term that depends only on $\|x + y\|$. Specifically,

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{4\sigma^2}\right)I_f\left(\frac{\|x+y\|}{\sqrt{2}\sigma}\right), \quad (9)$$

where

$$I_f(u) := \frac{\omega_d}{2^{d+p}(2\pi)^{\frac{d}{2}}} \int_0^{\infty} \frac{x^{d-1+2p}}{f(x)} g_{d,u}(x) \, dx, \quad (10)$$

and $g_{d,u}(x)$ is the density of the non-central chi-distribution given in (6). Note that this kernel is not shift invariant because of the term $I_f(u)$.

The next result shows that the MMD defined by this kernel provides an upper bound on the GOT.

Theorem 2. *Let k be defined as in* (9). *For any $P, Q \in \mathcal{P}(\mathbb{R}^d)$ such that $\int \sqrt{k(x,x)} \, dP(x)$ and $\int \sqrt{k(x,x)} \, dQ(x)$ are finite, the MMD defined by k provides and upper bound on the GOT:*

$$\mathcal{T}_p^{(\sigma)}(P, Q) \leq 2^{\max(p-1,0)}\sigma^p\gamma_k(P, Q).$$

The significance of Theorem 2 is twofold. From the perspective of GOT, it provides a natural connection between the role of Gaussian smoothing and normalization of the kernel. From the perspective of MMD, Theorem 2 describes a family of kernels that metrize convergence in distribution as well as convergence in p-th moments.

Similar to the analysis of convergence rates in previous work (Goldfeld and Greenewald, 2020; Goldfeld et al., 2020b), the proof of Theorem 2 builds upon the fact that $\mathcal{T}_p(P, Q)$ can be upper bounded by a weighted total variation distance (Villani, 2008, Theorem 6.13). The novelty of Theorem 2 is that it establishes an explicit relationship with the kernel MMD and also provides a much broader class of upper bounds parameterized by the density f.

3.2 A 'Two-moment' Kernel

One potential limitation of Theorem 2 is that for a particular choice of density f, the requirement that $\sqrt{k(x,x)}$ is integrable might not be satisfied for probability measures of interest. For example, the convergence rates in Goldfeld and Greenewald (2020) and Goldfeld et al. (2020b) can be obtained as a corollary of Theorem 2 by choosing f to be the density of the generalized gamma distribution. However, the inverse of this density grows faster than exponentially, and as a consequence, the resulting bound can be applied only to the case of sub-Gaussian distributions.

The main idea underlying the approach in this section is that choosing a density with heavier tails leads to an upper bound that holds for a larger class of probability measures. Motivated by the functional inequalities appearing in Reeves (2020), we consider the following density function, which belongs to the family of generalized beta-prime distributions:

$$f(x) = \frac{\epsilon}{2\pi x}\left(\left(\frac{x}{\lambda}\right)^{-\epsilon} + \left(\frac{x}{\lambda}\right)^{\epsilon}\right)^{-1}.$$

For this special choice, the function $I_f(u)$ can be expressed as the weighted sum of two moments of the non-central chi distribution. Starting with (9) and simplifying terms leads to the following:

Definition 1. The *two-moment* kernel $k \colon \mathbb{R}^d \times \mathbb{R}^d \to$

$\mathbb{R}$ is defined as

$$k(x, y) = \alpha_{d,p} \exp\left(-\frac{\|x-y\|^2}{4\sigma^2}\right) J\left(\frac{\|x+y\|}{\sqrt{2}\sigma}\right) \quad (11)$$

for all $\epsilon \in (0, d+2p]$ and $\lambda \in (0, \infty)$, where

$$\alpha_{d,p} := \frac{(2\pi)2^{-(p+d)}2^{-d/2}}{\Gamma(\frac{d}{2})}$$

$$J(u) := \frac{\lambda^{\epsilon} M_{d,u}(d+2p-\epsilon) + \lambda^{-\epsilon} M_{d,u}(d+2p+\epsilon)}{2\epsilon}.$$

In this expression, $M_{d,u}(s)$ denotes the s-th moment of the non-central chi distribution with d degrees of freedom and parameter u.

A useful property of the two-moment kernel is that it satisfies the upper bound

$$k(x, y) \leq C_{d,p,\epsilon,\lambda}\left(1 + \|x\|^{d+2p+\epsilon} + \|y\|^{d+2p+\epsilon}\right),$$

where the constant depends only on $(d, p, \epsilon, \lambda)$ (see Supplement A for details). As a consequence:

Theorem 3. *Fix any $s > p + d/2$. For all $0 < \epsilon < \min(d + 2p, 2s - 2p - d)$ and $\lambda \in (0, \infty)$ the MMD defined by the two-moment kernel is a metric on the space $\mathcal{P}_s(\mathbb{R}^d)$ of probability measures with finite s-th moment. Furthermore, for all $P, Q \in \mathcal{P}_s(\mathbb{R}^d)$,*

$$\mathcal{T}_p^{(\sigma)}(P, Q) \leq 2^{\max(p-1,0)}\sigma^p \gamma_k(P, Q).$$

Remark 1. If $d + 2p \pm \epsilon$ are even integers then $J(u)$ is an even polynomial of degree $s = d + 2p + \epsilon$ with non-negative coefficients. For example, if $d = 3$, $p = 1$, and $(\epsilon, \lambda) = (1, 1)$, then

$$J(u) = 60 + \frac{115}{2}u^2 + 11u^4 + \frac{1}{2}u^6. \quad (12)$$

Methods for efficient numerical approximation of $J(u)$ are provided in Supplement C.1.

4 Convergence Rate

We now turn our attention to the fundamental question of how well the empirical measure of iid samples approximates the true underlying distribution. Let $S_1, \ldots, S_n \in \mathbb{R}^n$ be a sequence of n independent samples with common distribution P. The empirical measure P_n is the (random) probability measure on $\mathbb{R}^d$ that places probability mass $1/n$ at each sample point:

$$P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{S_i}, \quad (13)$$

where δ_x denotes the pointmass distribution at x.

Distributional properties of the kernel MMD between P and P_n have been studied extensively (Gretton et al.,

2012). For the purposes of this paper, we will focus on the expected difference between these distributions. As a straightforward consequence of (5) one obtains an exact expression for the expectation of the squared MMD:

$$\mathbb{E}\left[\gamma_k^2(P, P_n)\right] = \frac{\mathbb{E}[k(X, X)] - \mathbb{E}[k(X, X')]}{n}, \quad (14)$$

where X, X' are independent draws from P. Note that the numerator can also be expressed as $\mathbb{E}\left[\gamma_1^2(P, P_1)\right]$, i.e., the squared MMD under $n = 1$ samples. By Jensen's inequality, the first moment satisfies

$$\mathbb{E}[\gamma_k(P, P_n)] \leq \frac{\sqrt{\mathbb{E}[k(X, X)]}}{\sqrt{n}}. \quad (15)$$

In the following, we focus on the two-moment kernel given in (11) and study how $\mathbb{E}[k(X, X)]$ depends on P and the parameters (p, σ).

We note that because γ_k satisfies the triangle inequality, all of the results provided here extend naturally to the two-sample settings where one the goal is to approximate the distance $\gamma_k(P, Q)$ based on the empirical measures P_n and Q_m.

4.1 Finite Moment Condition

We begin with an upper bound on $\mathbb{E}[k(X, X)]$ that holds whenever $\|X\|$ has a moment greater than $d + 2p$.

Theorem 4. *Let $X \in \mathbb{R}^d$ be a random vector that satisfies $(\mathbb{E}[\|X\|^s])^{1/s} \leq m$ for some $s > d + 2p$. If k is the two-moment kernel given in (11) with parameters $0 < \epsilon \leq \min(d + 2p, s - d - 2p)$ and $\lambda = \sqrt{d + 2p + \epsilon} + \sqrt{2}m/\sigma$, then*

$$\mathbb{E}[k(X, X)] \leq \frac{\alpha_{d,p}}{\epsilon}\left(\sqrt{2d + 2p + \epsilon} + \frac{\sqrt{2}m}{\sigma}\right)^{d+2p}.$$

In view of (15), Theorem 1 follows as an immediate corollary.

Small noise limit and unsmoothed OT. It is instructive to consider the implications of Theorem 1 in the limit as σ converges to zero. By two applications of the triangle inequality and the fact that $\mathcal{T}_p(Q, Q * \mathcal{N}_\sigma) \leq \sigma M_d(p)$ for every $Q \in \mathcal{P}$, one finds that, for any $\sigma \in [0, \infty)$, the (unsmoothed) OT distance can be upper bounded according to

$$\mathcal{T}_p(P, P_n) \leq C_{d,p}\left(\mathcal{T}_p^{(\sigma)}(P, P_n) + \sigma^p\right). \quad (16)$$

Combining (16) with Theorem 1 and then evaluating at $\sigma = mn^{-\frac{1}{d+2p}}$ leads to the (unsmoothed) convergence rate given in (4).

4.2 Sub-gamma Condition

Next, we provide a refined bound for distributions satisfying a sub-gamma tail condition.

Definition 2. A random vector $X \in \mathbb{R}^d$ is said to be *sub-gamma* with variance parameter $v > 0$ and scale parameter $b \geq 0$ if

$$\mathbb{E}\big[\exp(\alpha^\top X)\big] \leq \exp\left(\frac{v\|\alpha\|^2}{2(1 - b\|\alpha\|)}\right) \qquad (17)$$

for all $\alpha \in \mathbb{R}^d$ such that $\|\alpha\| \leq 1/b$. If this condition holds with $b = 0$ then X is said to be *sub-Gaussian* with variance parameter v.

Properties of sub-Gaussian and sub-gamma distributions have been studied extensively; see e.g, Boucheron et al. (2013). In particular, if X_1 and X_2 are independent sub-gamma random vectors with parameters (v_1, b_1) and (v_2, b_2), respectively, then $X_1 + X_2$ is sub-gamma with parameters $(v_1 + v_2, \max(b_1, b_2))$.

The next result provides an upper bound on the moments of the magnitude of a sub-gamma vector. Although there is a rich literature this topic, we were unable to find a previous statement of this result and so it may be of independent interest.

Theorem 5. *Let $X \in \mathbb{R}^d$ be a sub-gamma random vector with parameters (v, b). For all $s \in [0, \infty)$*

$$\mathbb{E}[\|X\|^s] \leq \sqrt{2e}\big(\sqrt{v} + \sqrt{s}\,b\big)^s M_d(s), \qquad (18)$$

where $M_d(s) := 2^{\frac{s}{2}}\Gamma(\frac{d+s}{2})/\Gamma(\frac{d}{2})$ is the s-th moment of the chi distribution with d degrees of freedom. Furthermore, $M_d(s) \leq \overline{M}_d(s)$ where

$$\overline{M}_d(s) := \left(\frac{d+s}{e}\right)^{\frac{s}{2}}\left(\frac{d+s}{d}\right)^{\frac{d-1}{2}}. \qquad (19)$$

Theorem 6. *Let $X \in \mathbb{R}^d$ be a sub-gamma random vector with parameters (v, b). Let $r = d + 2p$ and for $t \in (0, d + 2p]$ define*

$$m(t) := (\sqrt{\sigma^2 + 2v} + \sqrt{r + t}\,b)^{r+t}\overline{M}_d(r + t),$$

where $\overline{M}_d(s)$ is defined in (19). If k is the two-moment kernel defined in (11) with parameters $\epsilon = \sqrt{d}$ and $\lambda = (m(\epsilon)/m(-\epsilon))^{1/(2\epsilon)}$ then

$$\mathbb{E}[k(X, X)] \leq C(d + p)^p\Big(\sqrt{1 + \frac{2v}{\sigma^2}} + \frac{\sqrt{d + 2p}\,b}{\sigma}\Big)^{d+2p}, \qquad (20)$$

where $C = \sqrt{2\pi}e^{7/8} < 6.02$.

In view of (15), Theorem 6 gives an upper bound on the convergence rate of $\gamma_k(P, P_n)$ with an explicit

dependence on the sub-gamma parameters (d, p, σ, v, b). For the special case of a sub-Gaussian distribution ($b = 0$) and $p \in \{0, 1\}$, this bound recovers the results in Goldfeld and Greenewald (2020). Going beyond the setting of sub-Gaussian distributions (i.e., $b > 0$) this bound quantifies the dependence on the dimension d and the scale parameter b.

Phase transition in scale parameter. In the high-dimensional setting where d increases with n, Theorem 6 exhibits two distinct regimes depending on the behavior of the scale parameter. Suppose that (p, σ, v) are fixed while b scales with d. If $b = O(d^{-\delta})$ for some $\delta > 1/2$ then $\mathbb{E}[k(X, X)]$ grows at most exponentially with the dimension:

$$\frac{\log \mathbb{E}[k(X, X)]}{d} \leq \frac{1}{2}\log\Big(1 + \frac{2v}{\sigma^2}\Big) + o(1). \qquad (21)$$

Conversely if $b = \Omega(d^{-\delta})$ for $\delta < 1/2$ then the upper bound increases faster than exponentially:

$$\frac{\log \mathbb{E}[k(X, X)]}{d} \leq \frac{2(1 - \delta)\log d}{2} + O(1).$$

The following example provides a specific example of a sub-gamma distribution which shows that the upper bound in Theorem 6 is tight with respect to the scaling of the dimension d and the scale parameter b. Full details of this example are provided in Supplement C.2.

Example 1. Suppose that $X = \sqrt{U}Z$ where $Z \sim \mathcal{N}(0, I_d)$ is a standard Gaussian vector and U is an independent Gamma random variable with shape parameter $1/(2b^2)$ and scale parameter $2b^2$. Then, X satisfies the sub-gamma condition with parameters $(1, b)$ and so the upper bound in Theorem 6 applies. Moreover, for every pair (ϵ, λ) the expectation of the two-moment kernel satisfies the following lower bound

$$\mathbb{E}[k(X, X)] \geq \frac{\alpha_{d,p}}{\epsilon}\left(\frac{\sqrt{2}b}{\sigma}\right)^r (M_{b^{-2}}(r - \epsilon)M_{b^{-2}}(r + \epsilon))^{1/2}$$

$$\times (M_d(r - \epsilon)M_d(r + \epsilon))^{1/2}. \qquad (22)$$

The bounds on $\mathbb{E}[k(X, X)]$ given in (20) and (22) are shown in Figure 1 as a function of $\delta = (\log d)/(\log b)$ for various values of d. The plot demonstrates a phase transition at the critical value of $\delta = 0.5$. Further computational results on this example are given in Section 5.1.

4.3 Dependent Samples

Motivated by applications involving Markov chain Monte Carlo there is significant interest in understanding the rate of convergence when there is dependence
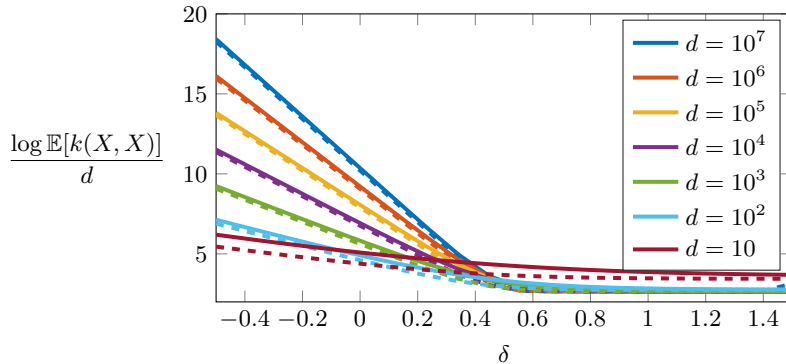
Figure 1: Bounds on $\frac{1}{d}\log\mathbb{E}[k(X,X)]$ for a distribution satisfying the sub-gamma condition with parameters $(1, d^{-\delta})$ as a function of δ for various d. In all cases, $p = 1$, $\sigma = 0.1$, and k is the 'two-moment' kernel described in Theorem 6. The upper bounds (solid line) are given by the right-hand side of (20). The lower bounds (dashed line) are given by the right-hand side of (22) evaluated at $\epsilon = \sqrt{d}$.

in the samples. Within the literature, this question is often address by focusing on stationary sequences satisfying certain mixing conditions (Peligrad, 1986). The basic idea is that the dependence between S_i and S_j decays rapidly as $|i - j|$ increases, then the effect of the dependence is negligible.

To go beyond the usual mixing conditions, a particularly useful property of the kernel MMD distance is that the second moment of $\gamma_k(P, P_n)$ depends only on the *pairwise* correlation in the samples. This perspective is useful for settings where there may not be a natural notion of time.

To make things precise, suppose that $S_1, \ldots, S_n \in \mathbb{R}^d$ is a collection of (possibly dependent) samples with identical distribution P. For each $i \neq j$ let Q_{ij} denote the law of (S_i, S_j). Starting with (5), the expectation of the squared MMD can now be decomposed as

$$\mathbb{E}\big[\gamma_k^2(P, P_n)\big] = \frac{1}{n}\mathbb{E}\big[\gamma_k^2(P, P_1)\big] + \frac{1}{n^2}\sum_{i\neq j} r_{ij}, \quad (23)$$

where $r_{ij} := \mathbb{E}_{Q_{ij}}[k(S_i, S_j)] - \mathbb{E}_{P\otimes P}[k(S_i, S'_j)]$. Notice that the first term in this decomposition is the second moment under *independent* samples. The second term is non-negative and depends only on the pairwise dependence, i.e., the difference between Q_{ij} and the probability measure obtained by the product of its marginals.

In some cases, it is natural to argue that only a small number of the terms r_{ij} are nonzero. More generally it is desirable to provide guarantees in terms of measures of dependence that do not rely on the particular choice of kernel. The next result provides such a bound in terms of the Hellinger distance.

Lemma 7. *If* $\mathbb{E}_P[k^2(X, X)]^{1/2} < C_{k,P}$, *then*

$$r_{ij} \leq \sqrt{2}\, C_{k,P}\, d_H(Q_{ij}, P \otimes P),$$

where d_H denotes the Hellinger distance.

The following example is inspired by random feature kernel interpretation of neural networks (Rahimi and Recht, 2008).

Example 2. Let $\{Z(\alpha) : \alpha \in \mathbb{R}^N\}$ be a $\mathbb{R}^d$-valued Gaussian processes with mean zero and covariance function $\mathsf{Cov}(Z(\alpha), Z(\beta)) = \langle\alpha, \beta\rangle I_d$. Suppose that samples from P are generated according to $S_i = T(Z(\alpha_i))$ where $\alpha_1, \ldots, \alpha_n$ are points on the unit sphere and $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is a function that maps a standard Gaussian vector into a vector with distribution P. Because Hellinger distance is non-increasing under the mapping given by T, it can be verified that

$$d_H(Q_{ij}, P \otimes P) \leq \sqrt{d}|\langle\alpha_i, \alpha_j\rangle|.$$

Thus, by (23) and Lemma 7, there exists a constant C_k depending only on k such that

$$\mathbb{E}\big[\gamma_k^2(P, P_n)\big] \leq C_{k,P}\Big(\frac{1}{n} + \frac{1}{n^2}\sum_{i\neq j}|\langle\alpha_i.\alpha_j\rangle|\Big). \quad (24)$$

This inequality holds for any set of points $\{\alpha_i\}$. To gain insight into the typical scaling behavior when the samples are nearly orthogonal on average, let us suppose that the $\{\alpha_i\}$ are drawn independently from the uniform distribution on the sphere. Then, $\mathbb{E}\big[|\langle a_i, a_j\rangle|^2\big] = 1/N$ and by, standard concentration arguments, one finds that the following upper bound holds with high probability when N is large:

$$\mathbb{E}\big[\gamma_k^2(P, P_n)\big] \leq C_{k,P}\Big(\frac{1}{n} + \sqrt{\frac{\log N}{N}}\Big). \quad (25)$$

5 Numerical Results

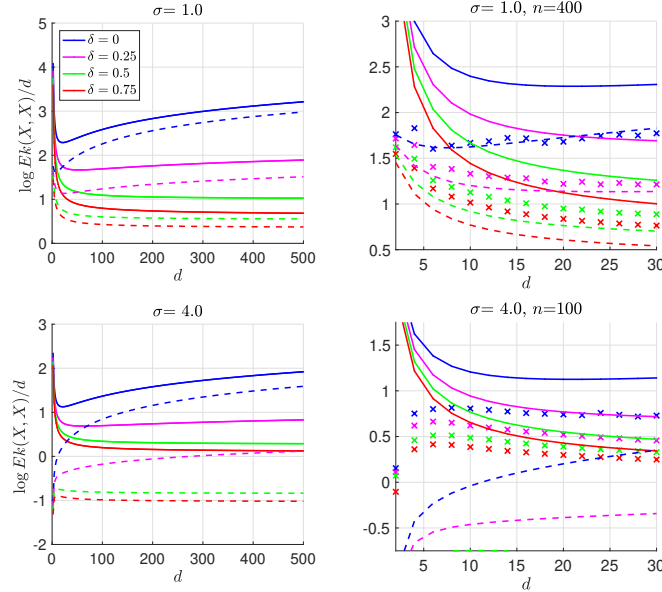In this section, we compute the upper bounds of GOT distance provided by the the empirical kernel MMD distance.

Figure 2: **Upper panel**: (Left) The UB (solid line) and LB (dashed line) of $\mathbb{E}[k(X,X)]$ given by (20) and (22), plotted as functions of d up to 500 and for various values of δ (shown in 4 colors), where X is sub-gamma random variable in $\mathbb{R}^d$ as in Example 1, $b = d^{-\delta}$, and the kernel function k has bandwidth $\sigma = 1$. (Right) Empirical estimates of $\mathbb{E}\hat{\Delta}^2_\gamma$ (cross markers), shown together with the UB and LB of $\mathbb{E}[k(X,X)]$ as in the left plot but over a range of d up to 30. The values are computed with $n = 400$ samples of X, and are averaged over 200 Monte-Carlo replicas. **Bottom panel**: Same plots where $\sigma = 4$, $n = 100$, and averaged over 100 Monte-Carlo replicas. In all the plots, the quantities are taken log and divide by d for better demonstration.

5.1 Bounds under Sub-gamma Condition

The sub-gamma condition allow us to address distributions that do not satisfy the sub-Gaussian condition. Upper bounds on the convergence rate for this class of distributions follow from Theorem 3 combined with Theorem 6.

For sub-gamma X as in Example 1, $\mathbb{E}[k(X,X)]$ has upper bound (UB) and lower bound (LB) as in (20) and (22) respectively. Here, in addition to the theoretical UB and LB as shown in Figure 1, we also compute the estimate of $\mathbb{E}\left[n\gamma_k^2(P,P_n)\right]$ approximated by the two-sample estimator. Let P_n and P'_n be the empirical measure defined as in (13) for independent iid copies of the sub-gamma random vector as in Example 1, $\{X_i\}_{i=1}^n$ and $\{X'_i\}_{i=1}^n$, respectively. Define $\hat{\Delta}^2_\gamma := \frac{n}{2}\gamma_k^2(P_n, P'_n)$, and then $\mathbb{E}[\hat{\Delta}^2_\gamma] = \mathbb{E}[k(X,X)] - \mathbb{E}[k(X,X')] = \mathbb{E}\left[n\gamma_k^2(P_n, P'_n)\right]$. Because the empirical kernel MMD (squared) distance $\gamma_k^2(P_n, P'_n)$ can be computed numerically from the two samples of sub-gamma random vectors, we can estimate the expectation of $\hat{\Delta}^2_\gamma$ by empirical average over Monte-Carlo replicas. Detailed numerical techniques to compute the two-moment kernel and experimental setup are provided in the supplementary materials.

The results are shown in Figure 2, where we set the parameter b controlling the shape and scale of X as

$b = d^{-\delta}$, $\delta = \{0, 0.25, 0.5, 0.75\}$. The data dimension d takes multiple values, and the kernel bandwidth σ takes values 1 and 4. In both cases, the left plot shows the same information as Figure 1 in view of increasing d, so as to be compared to the right plot. The right plot focuses on the case of small d, where the empirical estimates of $\mathbb{E}[\hat{\Delta}^2_\gamma]$ observe the UB. Notably, these values also lie between the UB and LB (recall that the LB applies to $\mathbb{E}[k(X,X)]$ but not $\mathbb{E}[\hat{\Delta}^2_\gamma]$) for both cases of σ, and approach the LB when $\sigma = 1$. This shows that our theoretical UB captures an important component of the kernel MMD distances for this example of X.

5.2 Dependent Samples via Gaussian Process

We generate dependent samples following a Gaussian process $\{Z(\alpha)\}_{\alpha \in S^N}$ as in Example 2 in Section 4.3, and numerically compute the values of $\gamma_k^2(P, P_n)$ by its two-sample version $\gamma_k^2(P_n, P'_n)$, using the same kernel as in the first experiment. Theoretically, when n is large, $\gamma_k^2(P, P_n)$ is expected to concentrate at its mean value as in (24), and then since α_i are uniformly sampled on the N-sphere, we also expect concentration at the value of (25).

We set $n = \{30, 50, 70, 100\}$, and vary N from 5 to 100. The data are in dimension $d = 5$, the kernel parameters are $\sigma = 0.5$, $c = 1$, $p = 1$, and for each value of n and N, $\gamma_k^2(P_n, P'_n)$ are computed for 100 realizations of the
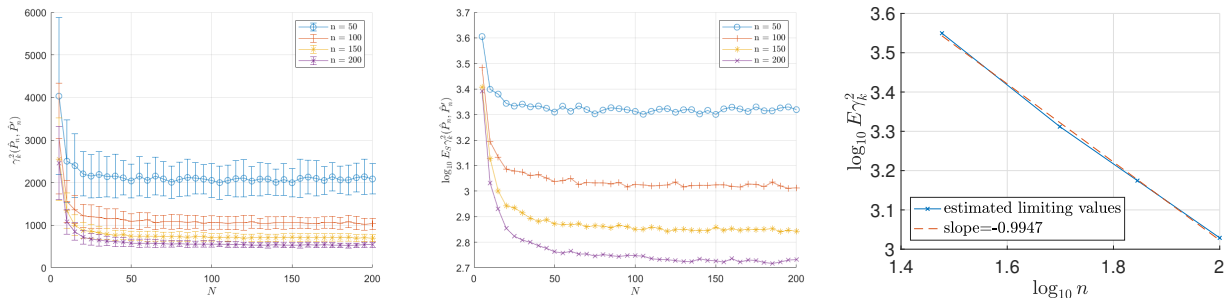
Figure 3: Empirical values of $\gamma_k^2(P_n, P'_n)$ for dependent samples in Example 2 in Section 4.3. (Left) Mean and standard deviation computed over 200 realizations of S_i's,P plotted against the increasing values of N and for different n. (Middle) The $\log_{10}$ of the mean value in the left plot. (Right) Log-log plot of the estimated limiting values of the mean in the left plot, by averaging over the range of $N \in [80, 100]$, v.s. n, showing a slope close to 1.

dependent random variables S_i, conditioned on one realization of the vectors α_i's. The results are shown in Figure 3, where for the small values of N, the computed approximate values of $\mathbb{E}[\gamma_k^2(P, P_n)]$ are decreasing because they are dominated by the contribution from the dependence in the samples, namely the part that is upper-bounded by the second term depending on α in (24). As N increases, these values converge to certain positive value which decrease as n increases.

Specifically, we take the average over the mean values of $\gamma_k^2(P_n, P'_n)$ for $N \in [80, 100]$ as an estimate of the limiting values, and Figure 3 (right) shows that these values decay as n^{-1}, as they correspond to the first term in (24). These numerical results show the competing two factors as predicted by the analysis in Section 4.3.

6   Conclusion and Discussion

The paper proves new convergence rates of GOT under general settings. Our results require only a finite moment condition and in the special case of sub-gamma distributions, we quantify the dependence on the dimension d and show a phase transition with respect to the scale parameter. Furthermore, our results cover the setting of dependent samples where convergence is proved only requiring a condition on pairwise sample dependence expressed by the kernel. Throughout our analysis, the main theoretical technique is to establish an upper bound using a kernel MMD where the kernel is called a "two-moment" kernel due to its special properties. The kernel depends on the cost function of the OT as well as the Gaussian smoothing used in GOT.

For the tightness of the kernel MMD upper bound, as has been pointed out in the comment beneath Theorem 1, our result shows that the convergence rate of $n^{-1/2}$ is tight in some regimes where $\sigma \to 0$ with $n \to \infty$ and the result matches the minimax rate in the unsmoothed OT setting. Alternatively, if σ is bounded

away from zero then it may be possible to obtain a better rate of convergence. For example, Proposition 6 in Goldfeld et al. (2020b) shows that if the pair (P, σ) satisfies an additional chi-square divergence condition, then $\mathcal{T}_2^{(\sigma)}(P, P_n)$ converges at rate $1/n$, which is faster than the general upper bound of $1/\sqrt{n}$ appearing in our paper. In this direction, pinning down the exact convergence rate in terms of regularity conditions on P remains an interesting open question for future work. In addition, it would be interesting to investigate the relationship between the Gaussian smoothing used in this paper and the multiscale representation of $\mathcal{T}_p$ in terms of partitions of the support of P, which was used in the analysis in Fournier and Guillin (2015) as well as the related work Weed et al. (2019).

In practice, the tightness of the kernel MMD upper bound also depends on the choice of kernel, which can be optimized for the data distribution and the level of smoothing in GOT. The question of whether the kernel MMD provides a useful alternative to OT distance in applications can be worthwhile of further investigation. Finally, another important direction of future work is to study computational methods and applications of the GOT approach, particularly in the high dimensional space. Currently, no specialized algorithm for GOP from finite samples has been developed, except for the direct method of applying any OT algorithm, e.g., entropy OT (Sinkhorn), to data with additive Gaussian noise (Goldfeld and Greenewald, 2020). Progress on the computational side will also enable various applications of GOT, e.g., the evaluation of generative models in machine learning.

References

Brahim Khalil Abid and Robert M Gower. Greedy stochastic algorithms for entropy-regularized optimal transport problems. *arXiv preprint arXiv:1803.01347*, 2018.

Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

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

Jérémie Bigot, Elsa Cazelles, Nicolas Papadakis, et al. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2): 5120–5150, 2019.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

Deeparnab Chakrabarty and Sanjeev Khanna. Better and simpler error analysis of the sinkhorn–knopp algorithm for matrix scaling. *Mathematical Programming*, pages 1–13, 2020.

Hong-Bin Chen and Jonathan Niles-Weed. Asymptotics of smoothed wasserstein distances. *arXiv preprint arXiv:2005.00738*, 2020.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

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

TW Epps and Kenneth J Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203, 1986.

V Alba Fernández, MD Jiménez Gamero, and J Muñoz García. A test for the two-sample problem based on empirical characteristic functions. *Computational statistics & data analysis*, 52(7):3730–3748, 2008.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162 (3):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

Samuel Gerber and Mauro Maggioni. Multiscale strategies for computing optimal transport. *The Journal of Machine Learning Research*, 18(1):2440–2471, 2017.

Ziv Goldfeld and Kristjan Greenewald. Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Italy, 2020.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *arXiv preprint arXiv:1810.05728*, 2018.

Ziv Goldfeld, Kristjan Greenewald, and Kengo Kato. Asymptotic guarantees for generative modeling based on the smooth wasserstein distance. *Advances in Neural Information Processing Systems*, 33, 2020a.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, July 2020b.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In

Advances in Neural Information Processing Systems, pages 181–189, 2016.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions*. John Wiley & Sons, Ltd, 1995.

Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

Jing Lei et al. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

Wuchen Li, Ernest K Ryu, Stanley Osher, Wotao Yin, and Wilfrid Gangbo. A parallel method for earth mover's distance. *Journal of Scientific Computing*, 75(1):182–197, 2018.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 5859–5870, 2018.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. *Kernel Mean Embedding of Distributions: A Review and Beyond*. 2017.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance, 2019.

Magda Peligrad. Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In *Dependence in probability and statistics*, pages 193–223. Springer, 1986.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*, volume 11 of *355-607*. Foundations and Trends in Machine Learning, 2019.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application

to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

Galen Reeves. A two-moment inequality with applications to rényi entropy and mutual information. *Entropy*, 22(11), 2020.

Ernest K Ryu, Yongxin Chen, Wuchen Li, and Stanley Osher. Vector and matrix optimal mass transport: theory, algorithm, and applications. *SIAM Journal on Scientific Computing*, 40(5):A3675–A3698, 2018.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*, 2017.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance, 2018. [Online]. Available `https://arxiv.org/abs/1802.08855`.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 33(4):67, 2014.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

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

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Alexander L Young and David B Dunson. Consistent entropy estimation for stationary time series. *arXiv preprint arXiv:1904.05850*, 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Ji Zhao and Deyu Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 2015.