Journal of Machine Learning Research 25 (2024) 1-55 Submitted 7/23; Revised 5/24; Published 10/24

Lower Complexity Adaptation for Empirical Entropic Optimal Transport

Michel Groppe MICHEL.GROPPE@UNI-GOETTINGEN.DE
Shayan Hundrieser S.HUNDRIESER@MATH.UNI-GOETTINGEN.DE
Institute for Mathematical Stochastics
University of Göttingen
Goldschmidtstraße 7, 37077 Göttingen, Germany

Editor: Marco Cuturi

Abstract

Entropic optimal transport (EOT) presents an effective and computationally viable alternative to unregularized optimal transport (OT), offering diverse applications for large-scale data analysis. In this work, we derive novel statistical bounds for empirical plug-in estimators of the EOT cost and show that their statistical performance in the entropy regularization parameter ε and the sample size n only depends on the simpler of the two probability measures. For instance, under sufficiently smooth costs this yields the parametric rate $n^{-1/2}$ with factor $\varepsilon^{-d/2}$, where d is the minimum dimension of the two population measures. This confirms that empirical EOT also adheres to the *lower complexity adaptation* principle, a hallmark feature only recently identified for unregularized OT. As a consequence of our theory, we show that the empirical entropic Gromov-Wasserstein distance and its unregularized version for measures on Euclidean spaces also obey this principle. Additionally, we comment on computational aspects and complement our findings with Monte Carlo simulations. Our technique employs empirical process theory and relies on a dual formulation of EOT over a single function class. Central to our analysis is the observation that the entropic cost-transformation of a function class does not increase its uniform metric entropy by much.

Keywords: Optimal transport, convergence rate, metric entropy, curse of dimensionality, Gromov-Wasserstein distance

1. Introduction

The mathematical theory of optimal transport (OT) offers versatile methods to compare complex objects that are modeled as probability distributions. From the problem of optimally moving soil as considered by the French mathematician Monge (1781), and its use in economics paved by the work of the Soviet mathematician Kantorovitch (1942, 1958) in the 20th century, it has evolved by now to a well-studied area of mathematics (Rachev and Rüschendorf, 1998a,b; Villani, 2003, 2009; Santambrogio, 2015; Panaretos and Zemel, 2020) and an active field of modern research. Its scope of applications spans across various disciplines such as machine learning (Courty et al., 2014, 2017; Flamary et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Ho et al., 2017; Grave et al., 2019), econometrics (Galichon, 2018; Hallin and Mordant, 2022; Hallin et al., 2022), computational biology (Evans and Matsen, 2012; Schiebinger et al., 2019; Tameling et al., 2021; Wang et al., 2021; Weitkamp et al., 2022), statistics (Munk and Czado, 1998; del Barrio et al., 1999; Sommerfeld and Munk,

2018; Panaretos and Zemel, 2019; Deb et al., 2021; Nies et al., 2021; Mordant and Segers, 2022), and computer vision (Solomon et al., 2015; Bonneel and Digne, 2023). Nevertheless, despite continuous progress, OT-based methodology is often burdened by computational and statistical limitations, restricting its utility for large-scale data analysis. This gave rise to the consideration of OT surrogates like entropic optimal transport (EOT) (Cuturi, 2013; Peyré and Cuturi, 2019; Nutz, 2021) in the hope of more favorable properties.

1.1 Optimal Transport

To formalize the OT problem, let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces equipped with Borel-σ-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$. Then, the OT cost between two probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ with respect to (w.r.t.) a Borel measurable cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is defined as

$$\mathrm{T}_c(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\pi. \tag{1}$$

Herein, $\Pi(\mu, \nu)$ denotes the set of all transport plans between μ and ν, i.e., every element $\pi \in \Pi(\mu, \nu)$ is a probability measure on $\mathcal{X} \times \mathcal{Y}$ such that $\pi(A \times \mathcal{Y}) = \mu(A)$ for all $A \in \mathcal{B}(\mathcal{X})$ and $\pi(\mathcal{X} \times B) = \nu(B)$ for all $B \in \mathcal{B}(\mathcal{Y})$. In case of a common Polish metric space $\mathcal{X} = \mathcal{Y}$ and costs chosen as the power of the metric, the OT cost quantifies the dissimilarity of μ and ν in a manner that is consistent with the ground space geometry (Villani, 2009, Chapter 6).

For most applied purposes the measures μ and ν are not available and instead one has only access to independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim \mu$ and independent thereof $Y_1, \ldots Y_n \sim \nu$. This gives rise to empirical measures

$$\hat{\mu}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}\,, \qquad \hat{\nu}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}\,,$$

which serve to define empirical plug-in estimators for the OT cost,

$$\widehat{\mathrm{T}}_{c,n} \in \{\mathrm{T}_c(\mu, \hat{\nu}_n), \mathrm{T}_c(\hat{\mu}_n, \nu), \mathrm{T}_c(\hat{\mu}_n, \hat{\nu}_n)\}. \tag{2}$$

The statistical performance of such estimators has been investigated extensively for various settings of which we only provide a selective overview, for a detailed recent account we refer to Staudt and Hundrieser (2024, Section 2). First contributions were made for metric costs with identical measures (Dudley, 1969; Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2019), but recently the analysis was refined to more general costs with possibly different measures (Chizat et al., 2020; Niles-Weed and Rigollet, 2022; Hundrieser et al., 2024b; Manole and Niles-Weed, 2024; Staudt and Hundrieser, 2024). A central quantity of interest in all these works is the mean absolute deviation of the empirical estimator to the true OT cost,

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,n} - \mathrm{T}_c(\mu, \nu)|]\,,$$

whose convergence behavior intricately depends on the regularity of the cost function, the intrinsic dimension of the measures μ and ν as well as their concentration. Elaborating on this, let us exemplarily consider Euclidean ground spaces $\mathcal{X} = \mathcal{Y} = [0, 1]^d$ with $d \neq 4$ and

squared Euclidean costs $c(x, y) = \|x - y\|_2^2$ for which it has been shown that (Chizat et al., 2020; Hundrieser et al., 2024b; Manole and Niles-Weed, 2024)[1]

$$\sup_{\mu,\nu\in\mathcal{P}([0,1]^d)} \mathbb{E}[|\widehat{\mathrm{T}}_{c,n} - \mathrm{T}_c(\mu,\nu)|] \asymp n^{-2/(d\vee 4)}. \tag{3}$$

Manole and Niles-Weed (2024) in fact also show that the empirical plug-in estimator is minimax rate optimal (up to logarithmic terms) among all measurable functions $\widehat{\mathrm{T}}$ of $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$. In particular, we observe that the statistical performance of the empirical plug-in estimator deteriorates exponentially with higher dimension d and that estimation of OT costs is affected by the curse of dimensionality.

Hence, the only way to hope for faster convergence rates is by imposing additional structural assumptions on the underlying population measures. For instance, if both μ and ν are supported on $[0,1]^s \times \{0\}^{d-s} \subseteq [0,1]^d$ for $s \neq 4$ the ground space is effectively only s-dimensional and the convergence rate improves to the better rate $n^{-2/s}$. A more surprising result by Hundrieser et al. (2024b) is that only one of the probability measures needs to be concentrated on a low-dimensional domain. If, say, μ is concentrated on an s-dimensional sub-manifold and $\nu \in \mathcal{P}([0,1]^d)$ is arbitrary, then the empirical OT cost estimator converges with rate $n^{-2/(s\vee 4)}$ and thus adapts to the lower complexity of μ. This phenomenon is termed *lower complexity adaptation* (LCA) principle. Nonetheless, note that the LCA principle only yields close to parametric rates in n if one of the measures has sufficiently low intrinsic dimension. Hence, unless $s < 4$, slower than parametric convergence rates for the estimation of OT costs can still occur.

In addition, OT is generally plagued by a high computational complexity (Peyré and Cuturi, 2019). For instance, the Auction algorithm (Bertsekas, 1981; Bertsekas and Castanon, 1989) or Orlin's algorithm (Orlin, 1988) have a worst-case computational complexity of $\mathcal{O}(n^3)$ (up to polylogarithmic terms) where n is the number of support points of the input measures. This effectively delimits the use of OT based methodology to measures with $n \sim 10^4$ support points.

1.2 Entropic Optimal Transport

To deal with the high computational complexity, Cuturi (2013) proposed to regularize the objective (1) by adding an entropic penalization term. For probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ the EOT cost w.r.t. to the cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is defined as

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \,\mathrm{d}\pi + \varepsilon \,\mathrm{KL}(\pi \mid \mu\otimes\nu)\,, \tag{4}$$

where $\varepsilon > 0$ is the entropic regularization parameter and $\mathrm{KL}(\pi \mid \mu \otimes \nu)$ denotes the Kullback-Leibler divergence of π relative the independent coupling of μ and ν, defined by $\int_{\mathcal{X}\times\mathcal{Y}} \log(\frac{\mathrm{d}\pi}{\mathrm{d}[\mu\otimes\nu]})\,\mathrm{d}\pi$ if $\pi \ll \mu \otimes \nu$ and $+\infty$ else. This regularization allows the use of an efficient and simple computational scheme called the Sinkhorn algorithm (Cuturi, 2013; Peyré and Cuturi, 2019; Schmitzer, 2019). Given a fixed precision, the Sinkhorn algorithm has computational complexity of order $\mathcal{O}(n^2)$ (up to polylogarithmic terms) (Altschuler et al.,

[1] In case of $d = 4$ the upper bound in (3) admits an additional logarithmic term while the lower bound does not. The sharp logarithmic order remains open.

2017; Dvurechensky et al., 2018; Luo et al., 2023) to approximate the EOT cost and is thus one order of magnitude faster than OT solvers. Owing to the computational appeal, EOT has therefore been employed to approximate quantities from unregularized OT by reducing the regularization parameter $\varepsilon \searrow 0$, which proves to be consistent under minimal regularity on the cost function (Pooladian and Niles-Weed, 2021; Altschuler et al., 2022; Bernton et al., 2022; Delalande, 2022; Nutz and Wiesel, 2022; Pal, 2024).

The EOT cost also admits a more desirable sample complexity compared to that of unregularized OT (Genevay et al., 2019; Mena and Niles-Weed, 2019; Chizat et al., 2020; Bayraktar et al., 2022; Rigollet and Stromme, 2022). Following an empirical plug-in approach as in (2),

$$\widehat{\mathrm{T}}_{c,\varepsilon,n} \in \{\mathrm{T}_{c,\varepsilon}(\mu,\hat{\nu}_n), \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\nu), \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)\}\,, \tag{5}$$

it was first shown by Genevay et al. (2019) under smooth costs and compactly supported $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ that for a fixed $\varepsilon > 0$ the estimator $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ based on i.i.d. observations $X_1,\dots,X_n \sim \mu$ and $Y_1,\dots,Y_n \sim \nu$ converges to the population quantity $\mathrm{T}_{c,\varepsilon}(\mu,\nu)$ at the parametric rate $n^{-1/2}$. In the setting where c is selected as the squared Euclidean norm their results imply that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim (1+\varepsilon^{-\lfloor d/2\rfloor})\exp(D^2/\varepsilon)n^{-1/2}\,, \tag{6a}$$

where D is the diameter of the bounded subset μ and ν are supported on, and the implicit constant only depends on d and D. Tailored to the squared Euclidean norm, Mena and Niles-Weed (2019) and Chizat et al. (2020) improved upon these results by reducing the exponential dependency in ε^{-1} to a polynomial one. Notably, the former also provide statistical bounds for σ^2-sub-Gaussian measures μ and ν,

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \varepsilon\left(1+\sigma^{\lceil 5d/2\rceil+6}\varepsilon^{-\lceil 5d/4\rceil-3}\right)n^{-1/2}\,, \tag{6b}$$

where the implicit constant only depends on d. For probability measures μ and ν that are supported on a joint set of diameter 1, the latter give for c being the squared Euclidean norm the bound

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim (1+\varepsilon^{-\lfloor d/2\rfloor})n^{-1/2}\,. \tag{6c}$$

More recently, Rigollet and Stromme (2022) showed for any bounded cost c, without imposing smoothness assumptions, that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \exp(\|c\|_\infty/\varepsilon)n^{-1/2}\,, \tag{6d}$$

where we again observe an exponential dependency in ε^{-1} and $\|c\|_\infty$. For all these results we see that for fixed $\varepsilon > 0$ the empirical EOT cost admits faster rates in n than the empirical unregularized OT cost. Such results are complemented by extensive research on distributional limits for the empirical OT cost at scaling rate $n^{1/2}$ which establish the parametric rate to be sharp (Bigot et al., 2019; Mena and Niles-Weed, 2019; Klatt et al., 2020; del Barrio et al., 2023; González-Sanz et al., 2022; González-Sanz and Hundrieser, 2023; Goldfeld et al., 2024a,b; Hundrieser et al., 2024a).

However, as the regularization parameter decreases to zero, the statistical error bound generally deteriorates in high dimensions polynomially or even exponentially in ε^{-1}. This

behavior is unavoidable, since for $\varepsilon \searrow 0$ the EOT cost tends to the unregularized OT cost which suffers from slower rates. Recalling the setting of $\mathcal{X} = \mathcal{Y} = [0,1]^d$ with squared Euclidean norm as c, it holds by Eckstein and Nutz (2024), see also Genevay et al. (2019), for some constant $K = K(d) > 0$ depending on d that

$$\sup_{\mu,\nu\in\mathcal{P}([0,1]^d)} |\,\mathrm{T}_{c,\varepsilon}(\mu,\nu) - \mathrm{T}_c(\mu,\nu)| \le d\varepsilon|\log\varepsilon| + K\varepsilon\,.$$

Combining this with the previously mentioned minimax result for the OT cost by Manole and Niles-Weed (2024) yields,

$$\inf_{\widehat{\mathrm{T}}} \sup_{\mu,\nu\in\mathcal{P}([0,1]^d)} \mathbb{E}[|\widehat{\mathrm{T}} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \gtrsim (n\log n)^{-2/d} - d\varepsilon|\log\varepsilon| - K\varepsilon\,,$$

where the infimum is taken over all measurable functions $\widehat{\mathrm{T}}$ of $X_1,\dots,X_n$ and $Y_1,\dots,Y_n$. Hence, we see that for $\varepsilon \searrow 0$ the statistical error in estimating the EOT cost in terms of ε must be affected by the ambient dimension d and that the empirical EOT cost is also burdened by the curse of dimensionality.

Nevertheless, in practical contexts it is often reasonable to expect that the data obeys some additional structure, e.g., that it is concentrated on a low-dimensional domain. The validity of this hypothesis is reflected by the performance of nonlinear dimension reduction techniques like manifold learning (Lin and Zha, 2008; Talwalkar et al., 2008; Zhu et al., 2018). Indeed, as for unregularized OT, if μ and ν are supported in $[0,1]^s \times \{0\}^{d-s}$, the bounds (6) hold for d replaced by s, indicating that empirical EOT depends on the intrinsic dimension. Qualitative results of this type were verified by Bayraktar et al. (2022) for settings where both measures μ and ν lie on s-dimensional domains. However, their results only assert slower than parametric convergence rates and do not shed light on the dependency of the constants in ε.

1.3 Contributions

The main primitive of this work is to provide a comprehensive statistical analysis of the empirical EOT cost with respect to n and ε in terms of the intrinsic dimension of the (possibly different) underlying measures. Fueled by findings of Hundrieser et al. (2024b), which assert that empirical (unregularized) OT adapts to lower complexity, we show that the empirical EOT cost estimator also obeys this principle and thus enjoys better constants in ε^{-1} if only one measure admits low intrinsic dimension.

To formulate our general LCA result, we first show for Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ with a bounded and measurable cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that the EOT cost admits a dual formulation over a single function class $\mathcal{F}_{c,\varepsilon}$, defined in (12) in Section 2.1, which depends on the cost function c and the regularization parameter ε (Proposition 3),

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = \max_{\phi\in\mathcal{F}_{c,\varepsilon}} \int_{\mathcal{X}} \phi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\mu)}\,\mathrm{d}\nu\,,$$

where $\phi^{(c,\varepsilon,\mu)}(\,\cdot\,) \coloneqq -\varepsilon\log\int_{\mathcal{X}} \exp[\varepsilon^{-1}(\phi(x) - c(x,\,\cdot\,))]\,\mu(\mathrm{d}x)$ is the entropic (c,ε)-transform of ϕ. Based on this formula, it follows from techniques of empirical process theory that the mean absolute deviation in (8) can be suitably bounded if the complexity of the function

classes $\mathcal{F}_{c,\varepsilon}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\mu)} = \{\phi^{(c,\varepsilon,\mu)} \mid \phi \in \mathcal{F}_{c,\varepsilon}\}$ as well as the empirical function class $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)} = \{\phi^{(c,\varepsilon,\hat{\mu}_n)} \mid \phi \in \mathcal{F}_{c,\varepsilon}\}$ can be suitably quantified. We characterize this via *covering numbers* w.r.t. to the uniform norm $\|\bullet\|_\infty$ for scale $\delta > 0$, defined by

$$N(\delta,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty) \coloneqq \inf\{n \in \mathbb{N} \mid \exists f_1, \dots, f_n : \mathcal{X} \to \mathbb{R} \text{ s.t. } \sup_{f \in \mathcal{F}_c} \min_{1 \le i \le n} \|f - f_i\|_\infty \le \delta\}\,.$$

Notably, the logarithm of the uniform covering number is called the *uniform metric entropy*.

A simple yet crucial observation for our approach is that the covering number of the function classes $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\mu)}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$ at any scale $\delta > 0$ is linked to that of $\mathcal{F}_{c,\varepsilon}$ (Lemma 5),

$$\max\Big(N(\delta,\, \mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\mu)},\, \|\bullet\|_\infty), N(\delta,\, \mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)},\, \|\bullet\|_\infty)\Big) \le N(\delta/2,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty)\,.$$

Then, assuming the existence of constants $K_\varepsilon, k > 0$ such that for all $\delta > 0$ sufficiently small,

$$\log N(\delta,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty) \le K_\varepsilon \delta^{-k}\,, \tag{7}$$

we show in Theorem 6 for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ in (5) based on independent samples that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \sqrt{1 + K_\varepsilon} \begin{cases} n^{-1/2} & k < 2\,, \\ n^{-1/2}\log(n+1) & k = 2\,, \\ n^{-1/k} & k > 2\,. \end{cases} \tag{8}$$

As we demonstrate in Section 3, suitable bounds as in (7) can be guaranteed if the space $\mathcal{X}$ and the partially evaluated cost $\{c(\bullet, y)\}_{y \in \mathcal{Y}}$ are sufficiently regular, while the space $\mathcal{Y}$ can be arbitrary, highlighting the adaptation to lower complexity.

For instance, in the semi-discrete setting, i.e., where $\mathcal{X}$ consists of finitely many points we infer the parametric rate $n^{-1/2}$ without ε-dependency (Theorem 13). Moreover, under Lipschitz continuous costs on metric spaces (Theorem 14) or semi-concave costs on Euclidean domains (Theorem 15), our theory implies ε-independent rates which are of order $n^{-1/2}$ in low dimensions and slower in higher dimensions, matching those of the empirical OT cost. Under sufficiently smooth costs on Euclidean domains, our approach is also capable in asserting parametric rate $n^{-1/2}$ with lead constant in $\varepsilon^{-d/2}$, where d corresponds to the minimum dimension of the two domains $\mathcal{X}$ and $\mathcal{Y}$, and can be arbitrarily large (Theorem 18). Tailored to the squared Euclidean norm, we also build on bounds by Mena and Niles-Weed (2019) and demonstrate that an LCA principle also remains valid for sub-Gaussian measures (Theorem 23).

In Section 4 we show that the Sinkhorn algorithm can be used to compute estimators that also fulfill the bound (8). In Section 5 we confirm, as an application of our theory, that the (entropic) Gromov-Wasserstein distance also obeys the LCA principle. In Section 6 we perform diverse Monte Carlo simulations which highlight that the implications of the LCA principle can be observed numerically. Lastly, in Section 7 we discuss possible directions for future research. For the sake of exposition most proofs are deferred to Section A. Section B contains auxiliary properties on uniform covering numbers, while Section C gives various rescaling properties used throughout this work.

1.4 Concurrent Work

Parallel to this work, Stromme (2023) derived statistical bounds for the empirical EOT cost which complement our work on the LCA principle. For compactly supported probability measures μ and ν on $\mathbb{R}^d$ and an L-Lipschitz continuous cost function c, it is shown that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim (1+\varepsilon)\sqrt{N(\varepsilon/L,\, \mu,\, \|\bullet\|_2) \wedge N(\varepsilon/L,\, \nu,\, \|\bullet\|_2)}\, n^{-1/2}\,, \tag{9}$$

where $N(\varepsilon/L,\, \mu,\, \|\bullet\|_2) \coloneqq N(\varepsilon/L,\, \mathrm{supp}(\mu),\, \|\bullet\|_2)$ is the covering number of the support of μ (and analogous for ν), a phenomenon Stromme refers to as *minimum intrinsic dimension scaling*. The proof employs concavity of the dual formulation and a suitable bound on the $L^2(\mu\otimes\nu)$-norm of the density of the EOT plan in terms of the covering number of the support. In contrast, our results are build on empirical process theory and covering numbers of $\mathcal{F}_{c,\varepsilon}$ at a scale that depends on the sample size n, see Remark 9 for a refined comparison of the notion of complexities. Notably, the two approaches are distinct and do not imply the results of one or the other, yet their implications are of comparable nature.

1.5 Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces that are equipped with their Borel-σ-fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$, respectively. Denote with $\mathcal{P}(\mathcal{X})$ the set of Borel-probability measures on $\mathcal{X}$. Random elements are always implicitly defined on an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\mathbb{E}[\bullet]$ is the expectation operator. For a function $f : \mathcal{X} \to \mathbb{R}$ denote the uniform norm $\|f\|_\infty \coloneqq \sup_{x\in\mathcal{X}}|f(x)|$. Provided that f is measurable, define for $p \geq 1$ and a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ the $L^p(\mu)$-norm $\|f\|_{L^p(\mu)} \coloneqq \left(\int_{\mathcal{X}} |f|^p \,\mathrm{d}\mu\right)^{1/p}$. With $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, denote the outer sum $f \oplus g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, $(x,y) \mapsto f(x) + g(y)$. For vectors $x, y \in \mathbb{R}^d$ and $p \geq 1$, we denote the p-norm by $\|x\|_p \coloneqq (\sum_{i=1}^d |x_i|^p)^{1/p}$ and the inner product by $\langle x, y\rangle \coloneqq \sum_{i=1}^d x_i y_i$. Denote with $I_d \in \mathbb{R}^{d\times d}$ the d-dimensional identity matrix and for a matrix A write $\|A\|_2$ for its Frobenius norm. Given a subset $\mathcal{X} \subseteq \mathbb{R}^d$ we write $\mathring{\mathcal{X}}$ to denote its interior. For $n \in \mathbb{N}$, we write $[\![n]\!] \coloneqq \{1, \ldots, n\}$. To denote the floor and ceiling function evaluated at $a \in \mathbb{R}$ we write $\lfloor a \rfloor$ and $\lceil a \rceil$, respectively. Further, the minimum and maximum of $a, b \in \mathbb{R}$ are denoted as $a \wedge b$ and $a \vee b$, respectively.

2. Entropic Optimal Transport

In this section, we first introduce preliminary facts about EOT (Subsection 2.1), and then derive the statistical rates for the EOT cost which reflect the LCA principle based on the dual formulation (Subsection 2.2) as well as a projection approach (Subsection 2.3). Unless stated otherwise proofs are deferred to Section A. Our general results are stated under the following assumption on the cost function $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$.

Assumption (C). *The cost function c is measurable and bounded in absolute value by* 1.

Note that Assumption **(C)** essentially demands that the cost function c is bounded. Indeed, if $\|c\|_\infty \in (1, \infty)$, we can rescale the cost function as detailed in Remark 44 in Appendix C and still infer quantitative convergence statements for the original cost.

2.1 Duality and Complexity

Central to our approach for the analysis of the empirical EOT cost is a suitable dual representation. To this end, we follow first the work by Marino and Gerolin (2020). For $\varepsilon > 0$ we define the function $\exp_\varepsilon := \exp(\bullet/\varepsilon)$ and the set

$$L_\varepsilon^{\exp}(\mu) := \left\{ \phi : \mathcal{X} \to [-\infty, \infty) \text{ measurable with } 0 < \int_{\mathcal{X}} \exp_\varepsilon(\phi) \, \mathrm{d}\mu < \infty \right\},$$

and analogously $L_\varepsilon^{\exp}(\nu)$. Further, for $\phi \in L_\varepsilon^{\exp}(\mu)$, $\psi \in L_\varepsilon^{\exp}(\nu)$ we define the ε-entropic cost-transform w.r.t. c, abbreviated by (c,ε)-transform, as

$$\begin{aligned} \phi^{(c,\varepsilon,\mu)}(y) &:= -\varepsilon \log \int_{\mathcal{X}} \exp_\varepsilon(\phi(x) - c(x,y)) \, \mu(\mathrm{d}x) \,, \quad y \in \mathcal{Y} \,, \\ \psi^{(c,\varepsilon,\nu)}(x) &:= -\varepsilon \log \int_{\mathcal{Y}} \exp_\varepsilon(\psi(y) - c(x,y)) \, \nu(\mathrm{d}y) \,, \quad x \in \mathcal{X} \,. \end{aligned}$$

Note that $\phi^{(c,\varepsilon,\mu)}$ and $\psi^{(c,\varepsilon,\nu)}$ are again measurable functions. Lastly, we state the entropy-Kantorovich functional

$$D_{c,\varepsilon}^{\mu,\nu}(\phi,\psi) := \int_{\mathcal{X}} \phi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \psi \, \mathrm{d}\nu - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp_\varepsilon(\phi \oplus \psi - c) \, \mathrm{d}[\mu \otimes \nu] + \varepsilon \,.$$

With this notation at our disposal, we can state a general dual formulation of EOT.

Theorem 1 (Marino and Gerolin 2020, Theorem 2.8 and Proposition 2.12)**.** *Let Assumption* **(C)** *hold. Then, it holds for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ that*

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = \max_{\substack{\phi \in L_\varepsilon^{\exp}(\mu) \\ \psi \in L_\varepsilon^{\exp}(\nu)}} D_{c,\varepsilon}^{\mu,\nu}(\phi,\psi) \,. \tag{10}$$

In particular, optimizers exist and a pair $(\phi,\psi) \in L_\varepsilon^{\exp}(\mu) \times L_\varepsilon^{\exp}(\nu)$ is a maximizer of the above if and only if

$$\phi = \psi^{(c,\varepsilon,\nu)} \ \mu\text{-a.s.} \quad \text{and} \quad \psi = \phi^{(c,\varepsilon,\mu)} \ \nu\text{-a.s.}, \tag{11}$$

and they can be chosen such that $\|\phi\|_\infty, \|\psi\|_\infty \leq 3/2$. Furthermore, given a maximizing pair (ϕ,ψ) in (10), the EOT plan in (4) is given by

$$\mathrm{d}\pi := \exp_\varepsilon(\phi \oplus \psi - c) \, \mathrm{d}[\mu \otimes \nu] \,.$$

Conversely, if there are potentials ϕ, ψ such that $\pi \in \Pi(\mu,\nu)$, they are optimal.

Remark 2 (Canonical extension)**.** *Theorem 1 only asserts that a maximizing pair ϕ, ψ satisfies the optimality conditions (11) μ- and ν-almost surely. By defining $\tilde{\phi} := \psi^{(c,\varepsilon,\nu)}$ and $\tilde{\psi} := (\psi^{(c,\varepsilon,\nu)})^{(c,\varepsilon,\mu)}$ we can uniquely extend the potentials beyond the support of the underlying measures. Further, by possibly shifting the potentials $(\tilde{\phi},\tilde{\psi})$ by $(a,-a)$ for suitable $a \in \mathbb{R}$ (see Marino and Gerolin, 2020, Lemma 2.7), it still holds that $\|\tilde{\phi}\|_\infty, \|\tilde{\psi}\|_\infty \leq 3/2$.*

The canonical extension of dual potentials allows us to further rewrite the dual formulation in (10) to reduce it to a supremum over a single function class.

Proposition 3 (Duality). *Let Assumption* **(C)** *hold and define the function class*

$$\mathcal{F}_{c,\varepsilon} \coloneqq \bigcup_{\xi \in \mathcal{P}(\mathcal{Y})} \left\{ \begin{array}{l} \phi : \mathcal{X} \to \mathbb{R} \text{ such that } \exists \psi : \mathcal{Y} \to \mathbb{R} \text{ with} \\ \phi = \psi^{(c,\varepsilon,\xi)} \text{ and } \|\phi\|_\infty, \|\psi\|_\infty \leq 3/2 \end{array} \right\}. \tag{12}$$

Then, for any $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ it holds that

$$\mathrm{T}_{c,\varepsilon}(\mu, \nu) = \max_{\phi \in \mathcal{F}_{c,\varepsilon}} \int_{\mathcal{X}} \phi \, \mathrm{d}\mu + \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\mu)} \, \mathrm{d}\nu \,. \tag{13}$$

The dual representation of the EOT cost in Proposition 3 implies the following stability bound for the EOT cost with respect to the underlying measures.

Lemma 4 (Stability bound). *Let Assumption* **(C)** *hold. Then, it holds for any pairs of probability measures $\mu, \tilde{\mu} \in \mathcal{P}(\mathcal{X})$ and $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{Y})$ that*

$$|\mathrm{T}_{c,\varepsilon}(\tilde{\mu}, \tilde{\nu}) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)| \leq 2 \sup_{\phi \in \mathcal{F}_{c,\varepsilon}} \left| \int_{\mathcal{X}} \phi \, \mathrm{d}[\tilde{\mu} - \mu] \right| + 2 \sup_{\phi \in \mathcal{F}_{c,\varepsilon}} \left| \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\tilde{\mu})} \, \mathrm{d}[\tilde{\nu} - \nu] \right|.$$

Plugging in the empirical probability measures $\tilde{\mu} = \hat{\mu}_n$ and $\tilde{\nu} = \hat{\nu}_n$ in Lemma 4, it follows that the absolute difference between empirical estimators of the EOT cost is dominated by the sum of two suprema of empirical processes. Expectations of such quantities can be bounded with methods from empirical process theory by controlling the uniform metric entropy of the indexing function classes $\mathcal{F}_{c,\varepsilon}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)} = \{\phi^{(c,\varepsilon,\hat{\mu}_n)} \mid \phi \in \mathcal{F}_{c,\varepsilon}\}$ (see, e.g., Wainwright, 2019, Chapter 4). An important observation in this context is that the uniform metric entropy of a function class never considerably increases under entropic (c, ε)-transforms.

Lemma 5. *Let $c \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a measurable cost function that is bounded from below and let $\tilde{\mu} \in \mathcal{P}(\mathcal{X})$. Consider a function class $\mathcal{F} \subseteq L_\varepsilon^{\exp}(\tilde{\mu})$ on $\mathcal{X}$. Then, it holds for the (c, ε)-transformed function class $\mathcal{F}^{(c,\varepsilon,\tilde{\mu})} \coloneqq \{\phi^{(c,\varepsilon,\tilde{\mu})} \mid \phi \in \mathcal{F}\}$ and any $\delta > 0$ that*

$$N(\delta, \mathcal{F}^{(c,\varepsilon,\tilde{\mu})}, \|\bullet\|_\infty) \leq N(\delta/2, \mathcal{F}, \|\bullet\|_\infty) \,.$$

A similar result as in Lemma 5 was recently established by Hundrieser et al. (2024b) in the context of unregularized OT for the unregularized c-transform, and serves as the basis of the LCA principle. In the following subsection we use this insight to establish results of similar nature for the empirical EOT cost.

2.2 Lower Complexity Adaptation: Dual Perspective

As described before, Lemma 4 together with tools from classical empirical process theory can be used to bound the statistical error in terms of the uniform metric entropy of the function classes $\mathcal{F}_{c,\varepsilon}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$. An application of Lemma 5 further reduces this to controlling the complexity of $\mathcal{F}_{c,\varepsilon}$.

Theorem 6 (General LCA). *Let Assumption* **(C)** *hold. Assume there exist constants $k > 0$, $K_\varepsilon > 0$ and $\delta_0 \in (0, 1]$ such that*

$$\log N(\delta, \mathcal{F}_{c,\varepsilon}, \|\bullet\|_\infty) \leq K_\varepsilon \delta^{-k} \qquad \text{for } 0 < \delta \leq \delta_0 \,. \tag{14}$$

Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ from (5) for i.i.d. random variables $X_1, \dots, X_n \sim \mu$ and (independent to that) $Y_1, \dots, Y_n \sim \nu$ it holds that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \sqrt{1+K_\varepsilon} \begin{cases} n^{-1/2} & k < 2\,, \\ n^{-1/2}\log(n+1) & k = 2\,, \\ n^{-1/k} & k > 2\,, \end{cases}$$

where the implicit constant only depends on k and δ_0.

This theorem establishes the LCA principle for the empirical EOT cost. More precisely, under regularity of the cost function c and constraints on the complexity of the space $\mathcal{X}$, we will see that suitable bounds on the covering numbers of $\mathcal{F}_{c,\varepsilon}$ are available, which do not depend on the complexity of $\mathcal{Y}$. Intuitively, the two measures μ and ν should be understood as being defined on high-dimensional ambient spaces, but where the measure μ is concentrated on a low-dimensional domain $\mathcal{X}$. As a consequence, the statistical error bound for the empirical EOT cost will be suitably bounded if one measure admits low intrinsic dimension. By interchanging the roles of μ and ν, we notice that the statistical error will depend on the minimum intrinsic dimension.

Proof of Theorem 6. We pursue a similar proof strategy as Hundrieser et al. (2024b, Theorem 2.2), which is inspired by that of Chizat et al. (2020), for $\widehat{\mathrm{T}}_{c,\varepsilon,n} = \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)$ and note that the remaining one-sample estimators from (5) can be handled similarly. Using Lemma 4, it holds that

$$\begin{aligned} \mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \leq 2\,\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{c,\varepsilon}}\left|\int_{\mathcal{X}} \phi \,\mathrm{d}[\hat{\mu}_n - \mu]\right|\right] \\ + 2\,\mathbb{E}\left[\sup_{\psi\in\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}}\left|\int_{\mathcal{Y}} \psi \,\mathrm{d}[\hat{\nu}_n - \nu]\right|\right]. \end{aligned} \tag{15}$$

An application of Proposition 4.11 from Wainwright (2019) yields

$$\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \leq 4[\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}) + \mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)})]\,,$$

where $\mathcal{R}_n(\mathcal{F}_{c,\varepsilon})$ and $\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)})$ are the Rademacher complexities of $\mathcal{F}_{c,\varepsilon}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$, respectively, i.e., set

$$\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}) \coloneqq \mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{c,\varepsilon}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(X_i)\right|\right],$$

where $\sigma_1, \dots, \sigma_n \sim \mathrm{Unif}\{\pm 1\}$ are i.i.d. Rademacher variables that are independent of $X_1, \dots, X_n \sim \mu$, and define $\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)})$ similarly. Note that $\mathcal{F}_{c,\varepsilon}$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$ are both bounded in uniform norm by 3. Using Theorem 16 from von Luxburg and Bousquet (2004), we have

$$\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}/3) \leq \inf_{\delta\in[0,1]}\left(2\delta + \sqrt{32}n^{-1/2}\int_{\delta/4}^1 \sqrt{\log N(3u,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty)}\,\mathrm{d}u\right).$$

With Lemma 5 we obtain by independence of the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ a similar upper bound for $\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}/3)$ with $3u$ in the covering number replaced by $\frac{3}{2}u$, i.e.,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}/3) &\leq \mathbb{E}_{X_1,\dots,X_n}\left[\mathbb{E}_{\sigma_1,\dots,\sigma_n,Y_1,\dots,Y_n}\left[\sup_{\phi\in\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}/3}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\phi(Y_i)\right| \,\middle|\, X_1,\dots,X_n\right]\right] \\
&\leq \mathbb{E}_{X_1,\dots,X_n}\left[\inf_{\delta\in[0,1]}\left(2\delta + \sqrt{32}n^{-1/2}\int_{\delta/4}^1 \sqrt{\log N(3u,\, \mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)},\, \|\bullet\|_\infty)}\,\mathrm{d}u\right)\right] \\
&\leq \inf_{\delta\in[0,1]}\left(2\delta + \sqrt{32}n^{-1/2}\int_{\delta/4}^1 \sqrt{\log N(\tfrac{3}{2}u,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty)}\,\mathrm{d}u\right).
\end{aligned}$$

From the covering number bound (14) in conjunction with the fact that covering numbers are non-decreasing for decreasing scale we conclude

$$\begin{aligned}
&\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \\
&\quad\leq 12[\mathcal{R}_n(\mathcal{F}_{c,\varepsilon}/3) + \mathcal{R}_n(\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}/3)] \\
&\quad\leq 24\inf_{\delta\in[0,1]}\left(2\delta + \sqrt{32}n^{-1/2}\int_{\delta/4}^1 \sqrt{\log N(\tfrac{3}{2}u,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty)}\,\mathrm{d}u\right) \\
&\quad\leq 24\inf_{\delta\in[0,1]}\left(2\delta + \sqrt{32K_\varepsilon}n^{-1/2}\int_{\delta/4}^1 ([\tfrac{3}{2}u]\wedge\delta_0)^{-k/2}\,\mathrm{d}u\right) \qquad (16)
\end{aligned}$$

and the choices of $\delta \coloneqq 4n^{-1/(k\vee 2)}$ yield the desired result. ∎

In Theorem 6, we see that the statistical error bound may depend on the entropic regularization parameter ε through the term K_ε. Later in Subsection 3.4, we observe a trade-off between K_ε and k in leveraging the smoothness of the cost function. Namely, the rates in n improve with higher degree of smoothness while the dependency on small ε gets worse. As a consequence, depending on how fast ε tends to 0 with increasing sample size, an error bound leveraging less smoothness can assert a smaller mean absolute deviation of the empirical plug-in estimator.

Corollary 7 (Comparison of rates)**.** *Let Assumption* **(C)** *hold. Suppose that there exist constants k_1, k_2, $a > 0$ such that for sufficiently small $\delta > 0$ it holds that*

$$\log N(\delta,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty) \lesssim \min\{\varepsilon^{-a}\delta^{-k_1},\, \delta^{-k_2}\}.$$

Choose $\varepsilon = n^{-\gamma}$ for some $\gamma > 0$. Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ from (5) it holds (up to $\log(n+1)$-factors) that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \min\{n^{a\gamma/2 - 1/(k_1\vee 2)},\, n^{-1/(k_2\vee 2)}\}.$$

In particular, it follows that the second term in the minimum is smaller if and only if

$$\gamma > \left[\frac{1}{k_1\vee 2} - \frac{1}{k_2\vee 2}\right]\frac{2}{a}.$$

Remark 8 (Unbounded costs). *For the formulation of Theorem 6 we impose boundedness of the cost function c via Assumption* **(C)** *which ensures that dual potentials are appropriately bounded without assuming concentration properties of the underlying probability measures. Later, in Subsection 3.5 we employ a similar approach to show the validity of the LCA principle for squared Euclidean costs in the setting where one measure is sub-Gaussian while the other is compactly supported and concentrated on a low-dimensional domain.*

Remark 9 (Comparison with complexity scales of Stromme 2023). *The proof of Theorem 6, see Inequality* (16)*, illustrates that condition* (14) *can be relaxed to an upper bound on the metric entropy of $\mathcal{F}_{c,\varepsilon}$ at scales $\delta \geq 4n^{-1/(k\vee 2)}$, independent of ε, i.e.,*

$$\log N(\delta, \mathcal{F}_{c,\varepsilon}, \|\bullet\|_\infty) \leq K_\varepsilon \delta^{-k} \qquad \text{for } 4n^{-1/(k\vee 2)} \leq \delta \leq \delta_0\,.$$

This highlights that the empirical EOT cost serves as an effective estimator for the population EOT cost, similar to unregularized OT (Weed and Bach, 2019; Hundrieser et al., 2024b; Manole and Niles-Weed, 2024), if the uniform metric entropy of $\mathcal{F}_{c,\varepsilon}$ is small at scales of order $n^{-1/(k\vee 2)}$. Moreover, it complements the concurrently derived bound (9) *by Stromme (2023) which asserts that the empirical EOT cost also performs well if the minimum covering number of the supports of the underlying measures at an ε-dependent scale is sufficiently small. Collectively, our results and those of Stromme (2023) show that empirical EOT benefits from two* distinct *notions of covering number bounds operating at different scales. We like to stress that these two notions differ in their implications and allow for a trade-off in terms of the dependency in n and ε: While the bound* (9) *by Stromme (2023) achieves the parametric rate $n^{-1/2}$ at the cost of a strong dependency in ε, our results allow for potentially improved dependency in ε with possibly slower than parametric rates in n.*

Remark 10 (Comparison with unregularized OT). *As opposed to the entropic (c,ε)-transform, for unregularized OT the corresponding c-transform does not dependent on the probability measures. This allows for an LCA principle based on two function class $\mathcal{F}_c$, $\mathcal{F}_c^c$ that are independent of the underlying probability measures (Hundrieser et al., 2024b). In our setting, it hinges on the two classes $\mathcal{F}_{c,\varepsilon}$, $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$, where one is dependent on an empirical measure and thus causes additional technical difficulties. In particular, due to the complicated dependency of $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$ on $\hat{\mu}_n$, we are limit ourselves to imposing independence between the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The smooth nature of the entropic (c,ε)-transform permits however (see Subsection 3.4) a refinement in terms of the constants: The LCA principle for unregularized OT can only leverage smoothness up to order 2 of the underlying cost function, whereas here we benefit from arbitrarily large degrees of smoothness.*

2.3 Lower Complexity Adaptation: Primal Perspective

Our main result for the validity of the LCA principle (Theorem 6) employs metric entropy bounds, however does not provide much intuition for why the LCA principle is reasonable to expect. We therefore explore in the following a simple setting where the EOT plan entails a projective action in order to foster some additional insight about the LCA principle.

Proposition 11 (Nutz 2021, Theorem 4.2 and Remark 4.4). *Let $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$ be the product of two Polish spaces and suppose that the measurable cost function c decomposes for $x \in \mathcal{X}$*

and $y = (y_1, y_2) \in \mathcal{Y}$ *into*

$$c(x, y) = c_1(x, y_1) + c_2(y_2)\,, \tag{17}$$

where $c_1 : \mathcal{X} \times \mathcal{Y}_1 \to \mathbb{R}$ *and* $c_2 : \mathcal{Y}_2 \to \mathbb{R}$. *Then, it follows for all probability measures* $\mu \in \mathcal{P}(\mathcal{X})$ *and* $\nu \in \mathcal{P}(\mathcal{Y})$ *such that* $c_1 \in L^1(\mu \otimes \nu_1)$ *and* $c_2 \in L^1(\nu_2)$ *that*

$$\mathrm{T}_{c,\varepsilon}(\mu, \nu) = \mathrm{T}_{c_1,\varepsilon}(\mu, \nu_1) + \int_{\mathcal{Y}_2} c_2 \,\mathrm{d}\nu_2\,,$$

where ν_1 *and* ν_2 *are the marginals of* ν *on* $\mathcal{Y}_1$ *and* $\mathcal{Y}_2$, *respectively.*

The representation of the EOT cost under cost functions of the form (17) asserts that the EOT plan can be decomposed into a projective action $\nu \mapsto \nu_2$, which contributes a cost of the magnitude $\int_{\mathcal{Y}_2} c_2 \,\mathrm{d}\nu_2$, and the EOT plan between μ and ν_1, which causes the term $\mathrm{T}_{c_1,\varepsilon}(\mu, \nu_1)$.

Since for the squared Euclidean norm Proposition 11 remains valid for affine subspaces, we obtain the following representation. To ease notation, for two random variables X and Y on $\mathbb{R}^d$ we define $\mathrm{T}_{c,\varepsilon}(X, Y)$ as the EOT cost between their laws.

Corollary 12 (Squared Euclidean costs)**.** *Let* $c(x, y) = \|x - y\|_2^2$ *be the squared Euclidean norm on* $\mathbb{R}^d$. *Let* $s \leq d$, $U \in \mathbb{R}^{d\times s}$ *be orthogonal*[2] *and* $v \in \mathbb{R}^d$. *Then, it holds for random variables* X *and* Y *with finite second moments on* $\mathbb{R}^s$ *and* $\mathbb{R}^d$, *respectively, that*

$$\mathrm{T}_{c,\varepsilon}(UX + v, Y) = \mathrm{T}_{c,\varepsilon}(X, U^\intercal(Y - v)) + \mathbb{E}[\|(I_d - UU^\intercal)[Y - v]\|_2^2]\,.$$

Example 1 (Unit cubes)**.** *Consider* $\mathcal{X} = \mathcal{Y}_1 = [0, 1]^{d_1}$ *and* $\mathcal{Y}_2 = [0, 1]^{d_2}$ *and let* c *be the squared Euclidean norm where we embed* $\mathcal{X}$ *into* $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$ *by appending zeros (or embed via affine transformation). Then, it holds for* $x \in \mathcal{X}$ *and* $y = (y_1, y_2) \in \mathcal{Y}$ *that*

$$c(x, y) = \|x - y_1\|_2^2 + \|y_2\|_2^2\,,$$

and Proposition 11 (or Corollary 12) reduces the $(d_1 + d_2)$*-dimensional EOT problem to only* d_1 *dimensions.*

The representation of the EOT cost in Proposition 11 yields the following additional interpretation of the LCA principle. Let $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ with empirical versions $\hat{\mu}_n$, $\hat{\nu}_n$, then under the assumptions of Proposition 11 we see that

$$\begin{aligned}\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] \leq\ & \mathbb{E}[|\mathrm{T}_{c_1,\varepsilon}(\hat{\mu}_n, \hat{\nu}_{n,1}) - \mathrm{T}_{c_1,\varepsilon}(\mu, \nu_1)|] \\ & + \mathbb{E}\left[\left|\int_{\mathcal{Y}_2} c_2 \,\mathrm{d}[\hat{\nu}_{n,2} - \nu_2]\right|\right].\end{aligned}$$

The second term admits under a finite second moment $\int_{\mathcal{Y}_2} c_2^2 \,\mathrm{d}\nu_2 < \infty$ a statistical error of order $n^{-1/2}$ that is independent of ε. Hence, the ε-dependence of the statistical error for $\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)$ reduces to that of $\mathrm{T}_{c_1,\varepsilon}(\hat{\mu}_n, \hat{\nu}_{n,1})$ which depends on the simpler measure $\hat{\nu}_{n,1}$. In particular, the convergence rate for $\mathrm{T}_{c_1,\varepsilon}(\hat{\mu}_n, \hat{\nu}_{n,1})$ only depends on the regularity of c_1 on the smaller space $\mathcal{X} \times \mathcal{Y}_1$ (compared to $\mathcal{X} \times \mathcal{Y}$).

[2]By this we mean that the matrix $U \in \mathbb{R}^{d\times s}$ fulfills the relation $U^\intercal U = I_s$.

Moreover, note that Proposition 11 can also be employed to obtain lower bounds for the statistical error of empirical EOT. More specially, the reverse triangle yields that

$$\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \geq \mathbb{E}\bigg[\bigg|\int_{\mathcal{Y}_2} c_2 \,\mathrm{d}[\hat{\nu}_{n,2}-\nu_2]\bigg|\bigg] - \mathbb{E}[|\mathrm{T}_{c_1,\varepsilon}(\hat{\mu}_n,\hat{\nu}_{n,1})-\mathrm{T}_{c_1,\varepsilon}(\mu,\nu_1)|]\,.$$

Hence, if the second term on the right-hand side contributes a comparably small statistical error, the statistical error of $\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)$ approximately matches that of $\int_{\mathcal{Y}_2} c_2 \,\mathrm{d}\hat{\nu}_{n,2}$. This implies that, although the ε-dependency is only determined by the less complex space, the underlying (ε-independent) constant can be affected by the more complex space.

This is a scaling effect and arises if c_2 admits a large variance w.r.t. ν; we will revisit this observation later in our simulations for squared Euclidean costs. To suppress this effect we will mainly work in the subsequent section under Assumption **(C)**, i.e., a cost function that is uniformly bounded by one. However, for our result on sub-Gaussian measures in Subsection 3.5 we cannot simply rescale the cost function and arrive at underlying constants which do depend on the ambient dimension (Theorem 23).

3. Sample Complexity

In this section, we focus on concrete statistical implications of the LCA principle for the empirical EOT cost and employ Theorem 6 in various scenarios to derive upper bounds for the statistical error. To apply Theorem 6, we need to bound the uniform metric entropy of the function class $\mathcal{F}_{c,\varepsilon}$. Recall that each $\phi \in \mathcal{F}_{c,\varepsilon}$ can be written as

$$\phi(x) = -\varepsilon \log \int_{\mathcal{Y}} \exp_\varepsilon(\psi(y)-c(x,y))\,\xi(\mathrm{d}y)\,,$$

for some function $\psi : \mathcal{Y} \to \mathbb{R}$ and probability measure $\xi \in \mathcal{P}(\mathcal{Y})$. Based on this integral representation we see that suitable regularity properties of the partially evaluated cost $\{c(\,\cdot\,,y)\}_{y\in\mathcal{Y}}$ are transmitted to ϕ, which allows us to bound the uniform metric entropy of $\mathcal{F}_{c,\varepsilon}$. Another important aspect for these bounds is that the domain $\mathcal{X}$ is not too complex. We model this by assuming that $\mathcal{X}$ can be represented as a finite union $\bigcup_{i=1}^I g(\mathcal{U}_i)$ of $I \in \mathbb{N}$ spaces $\mathcal{U}_i$ embedded via $g_i : \mathcal{U}_i \to \mathcal{X}$ where $c(g_i(\,\cdot\,),y)$ needs to satisfy certain assumptions. This kind of structural assumption encompasses the setting where $\mathcal{X}$ consists of finitely many points, but also covers the scenario that $\mathcal{X}$ is a smooth compact sub-manifold embedded in a high-dimensional Euclidean space and where $(g_i,\mathcal{U}_i)_{i\in[\![I]\!]}$ corresponds to the collection of charts (see Lee 2013 for comprehensive treatment).

The following subsections explore various settings which are based on this approach: they all rely on imposing regularity on the cost c and the space $\mathcal{X}$ to establish suitable uniform metric entropy bounds for $\mathcal{F}_{c,\varepsilon}$. For exposition, we relegate all proofs to Appendix A.1 while technical insights on the uniform metric entropy are detailed in Section B.

3.1 Semi-Discrete

First, we consider the semi-discrete case, i.e., when $\mathcal{X}$ consists of finitely many points, while $\mathcal{Y}$ is a general Polish space. Then, under uniformly bounded costs, the function class $\mathcal{F}_{c,\varepsilon}$

can be understood as a set of uniformly bounded vectors and its uniform metric entropy is thus particularly simple to control.

Theorem 13 (Semi-discrete LCA)**.** *Let Assumption* **(C)** *hold and suppose that* $\mathcal{X} = \{x_1, \dots, x_I\}$*. Then, for all probability measures* $\mu \in \mathcal{P}(\mathcal{X})$*,* $\nu \in \mathcal{P}(\mathcal{Y})$ *and any of the empirical estimators* $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ *from* (5) *it holds for an implicit universal constant that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \sqrt{I} n^{-1/2}\,.$$

Notably, the obtained bound is independent of the regularization parameter ε and the parametric rate $n^{-1/2}$ is attained. This is in line with convergence rates for semi-discrete unregularized OT, see Theorem 3.2 from Hundrieser et al. (2024b), which is the limiting behavior $\varepsilon \searrow 0$. Theorem 13 also slightly improves upon the bound given by Stromme (2023, Example 4), asserting no dependence on ε and leaving out the condition on Lipschitz continuity of the cost.

3.2 Lipschitz Cost

Next, we assume a more general representation for the space $\mathcal{X}$ and impose a Lipschitz continuity condition on the cost function.

Assumption (Lip). *It holds that* $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ *for* $I \in \mathbb{N}$ *connected metric spaces* $(\mathcal{U}_i, d_i)$ *and maps* $g_i : \mathcal{U}_i \to \mathcal{X}$ *such that* $c(g_i(\,\cdot\,), y)$ *is* 1*-Lipschitz w.r.t.* d_i *for all* $y \in \mathcal{Y}$*.*

Using the scaling property from Remark 44, note that other uniform Lipschitz constants than 1 can be reduced to the above case. A central consequence of this condition is that for any element $\phi \in \mathcal{F}_{c,\varepsilon}$ the composition $\phi \circ g_i \colon \mathcal{U}_i \to \mathbb{R}$ is again 1-Lipschitz. The uniform metric entropy of the class of uniformly bounded, 1-Lipschitz functions on a metric space is well-studied (Kolmogorov and Tikhomirov, 1961) and implies the following result.

Theorem 14 (Lipschitz LCA)**.** *Let Assumption* **(C)** *and Assumption* **(Lip)** *hold and suppose that there exists some* $k > 0$ *such that for all* $i = 1, \dots, I$ *it holds that*

$$N(\delta, \mathcal{U}_i, d_i) \lesssim \delta^{-k} \qquad \text{for } \delta > 0 \text{ sufficiently small.}$$

Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$*,* $\nu \in \mathcal{P}(\mathcal{Y})$ *and any of the empirical estimators* $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ *from* (5) *it holds for an implicit constant which only depends on* $\mathcal{U}_1, \dots, \mathcal{U}_I$ *that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \begin{cases} n^{-1/2} & k < 2\,, \\ n^{-1/2}\log(n+1) & k = 2\,, \\ n^{-1/k} & k > 2\,. \end{cases}$$

The rate for $k \geq 2$ may appear suboptimal in terms of n when compared to the results by Rigollet and Stromme (2022); Stromme (2023). However, we note that the obtained bound does not depend on the regularization parameter ε. This is due to the fact that it does not affect the Lipschitz constant of the function class $\mathcal{F}_{c,\varepsilon}$. Similar to Corollary 7, we see that the above bound might be superior if ε decreases sufficiently fast. As for the semi-discrete

case, the upper bounds are identical to that of the empirical OT under Lipschitz costs, see Theorem 3.3 from Hundrieser et al. (2024b).

Furthermore, note that Assumption **(Lip)** requires connected metric spaces. For general metric spaces, slightly worse uniform metric entropy bounds are available, see Lemma A.2 from Hundrieser et al. (2024b). In this setting, the assertion of Theorem 14 remains valid at the price of an additional $\log(n+1)$-term for $k \geq 2$ (Staudt and Hundrieser, 2024, Remark 3.3 and Appendix B).

3.3 Semi-Concave Cost

In addition to Lipschitz continuity of the cost function c, we now assume semi-concavity. More specifically, we suppose that c is Lipschitz continuous and semi-concave in the first component with a uniform modulus over the second component. A function $f : \mathcal{U} \to \mathbb{R}$ on a bounded, convex subset $\mathcal{U} \subseteq \mathbb{R}^s$ is called Λ-semi-concave with modulus $\Lambda \geq 0$ if the function

$$u \mapsto f(u) - \Lambda \|u\|_2^2$$

is concave. With this definition at our disposal we state the following set of assumptions.

Assumption (SC). *It holds that $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ bounded, convex subsets $\mathcal{U}_i \subseteq \mathbb{R}^s$ and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ such that $c(g_i(\,\cdot\,), y)$ is 1-Lipschitz w.r.t. $\|\,\cdot\,\|_2$ and 1-semi-concave for all $y \in \mathcal{Y}$.*

The additional assumption of semi-concavity improves the available uniform metric entropy bounds for the function class $\mathcal{F}_{c,\varepsilon}$ (Bronshtein, 1976; Guntuboyina and Sen, 2013). Again, invoking the scaling property from Remark 44, the Lipschitz constants and semi-concavity moduli can be assumed to be 1.

Theorem 15 (Semi-concave LCA)**.** *Let Assumption* **(C)** *and Assumption* **(SC)** *hold. Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ from* (5) *it holds for an implicit constant which only depends on $\mathcal{U}_1, \dots, \mathcal{U}_I$ that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \begin{cases} n^{-1/2} & s < 4\,, \\ n^{-1/2}\log(n+1) & s = 4\,, \\ n^{-2/s} & s > 4\,. \end{cases}$$

Note that for $s \geq 4$ the rates are strictly slower than $n^{-1/2}$ but do not depend on ε, recall also Corollary 7. In particular, we see in the proof that the regularization parameter ε does not affect the semi-concavity modulus of the function class $\mathcal{F}_{c,\varepsilon}$. As for the semi-discrete and Lipschitz case, the rates are identical for OT under semi-concave costs (Hundrieser et al., 2024b, Theorem 3.8).

3.4 Hölder Cost

Overall, unregularized OT is not capable of leveraging higher degree of smoothness of the underlying cost function for faster convergence rates (Manole and Niles-Weed, 2024). In stark contrast, the entropic (c,ε)-transform transmits smoothness of the cost to the EOT potentials. Smoothness of the cost function has indeed been employed by several works (Genevay et al.,

2019; Mena and Niles-Weed, 2019; Chizat et al., 2020) for upper bounds on the statistical error of the empirical EOT cost with a polynomial dependency in ε^{-1} determined by the ambient dimension. In what follows, we show that this dimensional dependency obeys the LCA principle.

Assumption (Hol). *It holds that $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ bounded, convex subsets $\mathcal{U}_i \subseteq \mathbb{R}^s$ with nonempty interior and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ such that $c(g_i(\,\cdot\,), y)$ is α-times continuously differentiable with $\alpha \in \mathbb{N}$ and bounded partial derivatives uniformly in $y \in \mathcal{Y}$.*

Under this assumption, we show that the function classes $\mathcal{F}_{c,\varepsilon} \circ g_i$ are α-Hölder, for which suitable uniform metric entropy bounds are available (van der Vaart and Wellner, 1996, Theorem 2.7.1). To formalize this, we introduce some additional notation. Let $\mathcal{U} \subseteq \mathbb{R}^s$ be bounded and convex with nonempty interior. For $k \in [\![s]\!]^\kappa$ with $\kappa \in \mathbb{N}$, define $|k| \coloneqq \kappa$ and the differential operator

$$\mathrm{D}^k \coloneqq \frac{\partial^\kappa}{\partial u_{k_\kappa} \cdots \partial u_{k_1}} .$$

Furthermore, for $\alpha > 0$ let $\underline{\alpha} = \max\{m \in \mathbb{N}_0 \mid m < \alpha\}$ and for a function $f : \mathcal{U} \to \mathbb{R}$ set

$$\|f\|_\alpha \coloneqq \max_{|k| \le \underline{\alpha}} \sup_u \|\mathrm{D}^k f(u)\| + \max_{|k| = \underline{\alpha}} \sup_{u,v} \frac{|\mathrm{D}^k f(u) - \mathrm{D}^k f(v)|}{\|u - v\|^{\alpha - \underline{\alpha}}} ,$$

where the supremum is taken over $u, v \in \mathring{\mathcal{U}} : u \neq v$. In addition, we consider for $M > 0$ the class of α-Hölder functions on $\mathcal{U}$ with norm bounded by M,

$$\mathcal{C}_M^\alpha(\mathcal{U}) \coloneqq \{f : \mathcal{U} \to \mathbb{R} \text{ continuous with } \|f\|_\alpha \le M\} .$$

Under Assumption **(C)** and Assumption **(Hol)**, a simple consequence of the dominated convergence theorem is that the first α partial derivatives of the functions in $\mathcal{F}_{c,\varepsilon} \circ g_i$ exist. By definition of the entropic (c,ε)-transform, it follows that they adhere to a certain recursive structure, see Lemma 36. Using this, Genevay et al. (2019) show that the κ-th partial derivatives are bounded in uniform norm by a polynomial in ε^{-1} of order $\kappa - 1$. We adapt their proof to make the dependence on the cost function more explicit.

Lemma 16 (Bounds for derivatives)**.** *Let Assumption* **(C)** *and Assumption* **(Hol)** *hold. Denote the quantities*

$$C_{i,m} \coloneqq \sup_{|j|=m} \|\mathrm{D}^j[c \circ (g_i, \mathrm{id}_\mathcal{Y})]\|_\infty$$

and define

$$C^{(i,1)} \coloneqq C_{i,1} , \qquad C^{(i,\kappa+1)} \coloneqq \max\left(C_{i,\kappa+1}, C^{(i,\kappa)}, \max_{m=1,\dots,\kappa} C_{i,m} C^{(i,\kappa)} \right) .$$

Then, it holds for all $\phi \in \mathcal{F}_{c,\varepsilon}$, $k \in [\![s]\!]^\kappa$, $\kappa \le \alpha$ that

$$\|\mathrm{D}^k[\phi \circ g_i]\|_\infty \lesssim (\varepsilon \wedge 1)^{-(\kappa-1)} C^{(i,\kappa)} ,$$

where the implicit constant only depends on κ.

The above bounds can be used to uniformly bound the α-Hölder norm of functions in $\mathcal{F}_{c,\varepsilon} \circ g_i$. As a consequence, we obtain the following uniform metric entropy bound.

Proposition 17. *Let Assumption* **(C)** *and Assumption* **(Hol)** *hold. Then, it follows for* $\delta > 0$ *that*

$$\log N(\delta,\, \mathcal{F}_{c,\varepsilon},\, \|\bullet\|_\infty) \lesssim \left(\sum_{i=1}^{I} [C^{(i,\alpha)}]^{s/\alpha} \right) (\varepsilon \wedge 1)^{-s\frac{\alpha-1}{\alpha}} \delta^{-s/\alpha}\,,$$

where the implicit constant only depends on s, α *and* $\mathcal{U}_1, \dots \mathcal{U}_I$.

An application of Theorem 6 directly yields the following LCA result.

Theorem 18 (Hölder LCA)**.** *Let Assumption* **(C)** *and Assumption* **(Hol)** *hold. Then, for all probability measures* $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ *and any of the empirical estimators* $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ *from* (5) *it holds that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \left(1 + \sum_{i=1}^{I} [C^{(i,\alpha)}]^{\frac{s}{2\alpha}} \right) (\varepsilon \wedge 1)^{-\frac{s}{2}\frac{\alpha-1}{\alpha}} \begin{cases} n^{-1/2} & s/\alpha < 2\,, \\ n^{-1/2}\log(n+1) & s/\alpha = 2\,, \\ n^{-\alpha/s} & s/\alpha > 2\,, \end{cases}$$

where the implicit constant only depends on s, α *and* $\mathcal{U}_1, \dots, \mathcal{U}_I$.

Note that in the bound of Theorem 18 we have a trade-off between the exponent of ε and n. Namely, with increasing smoothness α the rate in ε or n get worse or better, respectively. In particular, in the case that Assumption **(Hol)** is satisfied for $\alpha > s/2$, we get the parametric rate $n^{-1/2}$. This condition is met by costs that are smooth in the first component with bounded derivatives uniformly over the second.

Example 2 (Squared Euclidean norm)**.** *Let* $B_r(0) \coloneqq \{x \in \mathbb{R}^d \mid \|x\|_2 \le r\}$ *be the centered ball of radius* $r \ge 1$ *and assume that we have* $\mathcal{X} = B_r(0) = \mathcal{Y}$ *with the scaled and squared Euclidean norm* $c(x,y) = \frac{1}{4}\|x-y\|_2^2$. *As* $\|c\|_\infty = r^2 \ge 1$, *we rescale via Remark 44 and Remark 45. This way, we obtain pushforwards* $\mu^{r^2}, \nu^{r^2}, \hat{\mu}_n^{r^2}, \hat{\nu}_n^{r^2}$ *that are all supported on* $B_1(0)$. *Hence, for* $d < 4$ *we can apply Theorem 15 and for* $d > 4$ *Theorem 18 with* $\alpha \coloneqq \lceil d/2 \rceil + 1$ *to obtain the* $n^{-1/2}$*-rate. Putting everything together, it follows for* $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] &= r^2\, \mathbb{E}[|\mathrm{T}_{c/r^2,\varepsilon/r^2}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{T}_{c/r^2,\varepsilon/r^2}(\mu,\nu)|] \\ &= r^2\, \mathbb{E}[|\mathrm{T}_{c,\varepsilon/r^2}(\hat{\mu}_n^{r^2},\hat{\nu}_n^{r^2}) - \mathrm{T}_{c,\varepsilon/r^2}(\mu^{r^2},\nu^{r^2})|] \\ &\lesssim n^{-1/2} \begin{cases} r^2 & d < 4\,, \\ r^{d\frac{\lceil d/2 \rceil}{\lceil d/2 \rceil + 1} + 2} (\varepsilon \wedge r^2)^{-\frac{d}{2}\frac{\lceil d/2 \rceil}{\lceil d/2 \rceil + 1}} & d \ge 4\,. \end{cases} \end{aligned}$$

This is in line with bounds obtained by Stromme (2023, Example 3) and Chizat et al. (2020, Lemma 5). Notably, this bound can also be obtained by combining Theorem 18 with the rescaling approach as described in Remark 46, which treats more general cost functions.

Remark 19 (Comparison of bounds)**.** *Let Assumption* **(C)** *and Assumption* **(Hol)** *hold, and let* $0 < \varepsilon \le 1$ *be arbitrary. We consider the following three cases:*

1. *For $\alpha = 1$, we are also in the setting of Assumption* **(Lip)** *w.r.t. $\|\bullet\|_2$. Notably, Theorem 14 and Theorem 18 yield the same rates for the statistical error in n and ε.*

2. *Let $\alpha = 2$. Then, Assumption* **(SC)** *holds. Theorem 15 and Theorem 18 again yield the same rates in n. However, under semi-concavity we have no dependence on ε whereas the Hölder condition has the factor $\varepsilon^{-s/4}$.*

3. *If $\alpha \geq 3$, as before we are in the setting of Assumption* **(SC)**. *In this case, the Hölder condition yields better rates in n whereas under semi-concavity we have no dependence on ε. Hence, loosely speaking, for a fixed or slowly decreasing ε the statistical error obtained by Theorem 18 is better. More specially, choosing $\varepsilon = n^{-\gamma}$ for some $\gamma > 0$, Corollary 7 yields that the bound obtained under semi-concavity yields a smaller error if and only if*

$$\gamma > \left[\frac{1}{s/\alpha \vee 2} - \frac{1}{s/2 \vee 2}\right] \frac{\alpha}{\alpha - 1} \frac{2}{s}\,.$$

In particular, for $s > 4$ and $s/\alpha < 2$ the above condition reduces to

$$\gamma > \left[\frac{1}{2} - \frac{2}{s}\right] \frac{\alpha}{\alpha - 1} \frac{2}{s} > \frac{s-4}{s^2}\,.$$

Notably, for increasing dimension s it follows that the regime $(\frac{s-4}{s^2}, \infty)$, where the bounds induced by semi-concavity yield a smaller bound, gets larger.

Example 3. *Let c be the squared Euclidean norm and choose $\varepsilon = n^{-\gamma}$ for some $\gamma > 0$.*

For $d \leq 4$, Theorem 15 yields the parametric rate $n^{-1/2}$ (up to $\log(n+1)$-term for $d = 4$) without any ε-dependency. This rate is faster than (or for $d = 1$ equal to) the one obtained from Theorem 18, independently of γ.

For $d > 4$, we have by Remark 19.3 that the rate obtained under semi-concavity is faster to the Hölder bound with $\alpha \in \mathbb{N}$ if and only if

$$\gamma > \left[\frac{1}{d/\alpha \vee 2} - \frac{2}{d}\right] \frac{\alpha}{\alpha - 1} \frac{2}{d}\,.$$

Choosing $\alpha \coloneqq \lceil d/2 \rceil + 1$ such that $d/\alpha < 2$, the above inequality reduces to

$$\gamma > \left[\frac{1}{2} - \frac{2}{d}\right] \cdot \frac{2\lceil d/2 \rceil + 2}{\lceil d/2 \rceil d} > \frac{d-4}{d^2}\,.$$

3.5 Squared Euclidean Norm with Sub-Gaussian Measures

All the previously given statistical error bounds are derived under Assumption **(C)** which requires the cost function to be bounded. In this subsection, we show a version of the LCA principle for a (partially) unbounded setting. Namely, we build on Mena and Niles-Weed (2019) and consider the squared Euclidean norm as cost and require the measures to be sub-Gaussian (Vershynin, 2018). A probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is called σ^2-sub-Gaussian for $\sigma > 0$ if

$$\int_{\mathbb{R}^d} \exp\left(\frac{\|x\|_2^2}{2d\sigma^2}\right) \mu(\mathrm{d}x) \leq 2\,.$$

We denote with $\mathrm{SG}_d(\sigma^2)$ the set of all σ^2-sub-Gaussian probability measures on $\mathbb{R}^d$. Mena and Niles-Weed (2019) derive statistical error bounds for the empirical EOT cost under sub-Gaussian measures by controlling suprema over empirical processes using L^2-metric entropy bounds (and not the uniform norm as we do). However, this approach does not seem to directly produce a result that shows the LCA principle as it is not compatible with Lemma 5 which requires the use of the uniform norm. To circumvent this technical limitation, we impose the condition that one of the measures has bounded support. Note that while the general proof strategy stays the same, the unboundedness requires different arguments than the ones used in the proof of Theorem 6. The proofs can be found in Section A.

Assumption (SG). *It holds that $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i) \subseteq \mathbb{R}^d$ for $I \in \mathbb{N}$ bounded, convex subsets $\mathcal{U}_i \subseteq \mathbb{R}^s$ with nonempty interior and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ that are α-times continuously differentiable with bounded partial derivatives where $\alpha \in \mathbb{N}$ with $s/\alpha < 2$. Furthermore, let $\mathcal{Y} = \mathbb{R}^d$, $c(x,y) = \frac{1}{2}\|x-y\|_2^2$ be the squared Euclidean norm, $\sigma^2 \geq 1$ and $\sup_{x\in\mathcal{X}}\|x\|_2 \leq r$ with $r \geq 1$.*

Note that $\mu \in \mathcal{P}(\mathcal{X})$ is always sub-Gaussian because of its bounded support. To apply results from Mena and Niles-Weed (2019) directly, we assume that $\mu \in \mathcal{P}(\mathcal{X}) \cap \mathrm{SG}_d(\sigma^2)$ and $\nu \in \mathrm{SG}_d(\sigma^2)$, i.e., that σ^2 is large enough to include the sub-Gaussianity of both measures. Further, because of Remark 45 we can fix $\varepsilon = 1$ without loss of generality.

We proceed similar to the approach in Section 2. First, we formulate a version of duality (compare with Proposition 3).

Proposition 20 (Duality, Mena and Niles-Weed 2019, Proposition 6)**.** *Let Assumption* **(SG)** *hold, fix $\varepsilon = 1$ and define the function class*

$$\mathcal{F}_\sigma := \left\{ \begin{array}{l} \phi : \mathcal{X} \to \mathbb{R} \text{ such that } \exists \xi \in \mathrm{SG}_d(\sigma^2), \psi : \mathcal{Y} \to \mathbb{R} \\ \text{with } \phi = \psi^{(c,\varepsilon,\xi)}, \|\phi\|_\infty \leq 6d^2r^2\sigma^4 \\ \text{and } \psi(y) - \dfrac{1}{2}\|y\|_2^2 \leq d\sigma^2 + \sqrt{2d}\sigma\|y\|_2 \ \forall y \in \mathcal{Y} \end{array} \right\}.$$

Then, for any $\mu \in \mathcal{P}(\mathcal{X}) \cap \mathrm{SG}_d(\sigma^2)$ and $\nu \in \mathrm{SG}_d(\sigma^2)$ it holds that

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = \max_{\phi\in\mathcal{F}_\sigma} \int_{\mathcal{X}} \phi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\mu)} \,\mathrm{d}\nu\,.$$

As the squared Euclidean norm is smooth, it follows in conjunction with the strong concentration assumption via the Leibniz integral rule that elements of $\mathcal{F}_\sigma \circ g_i$ are α-times differentiable. Note that the dominated convergence theorem is still applicable because of sub-Gaussianity. To employ the uniform metric entropy bounds for Hölder classes, we need uniform bounds for the partial derivatives as in Lemma 16.

Lemma 21 (Bounds for derivatives)**.** *Let Assumption* **(SG)** *hold and fix $\varepsilon = 1$. Define*

$$G_{i,m} := \sup_{|j|=m} \|\mathrm{D}^j g_i\|_\infty$$

and

$$G^{(i,0)} := 1\,, \qquad G^{(i,\kappa+1)} := \max(G_{i,\kappa+1}, G^{(i,\kappa)}, \max_{m=1,\dots,\kappa} G_{i,m}G^{(i,\kappa)})\,.$$

Then, it holds for all $\phi \in \mathcal{F}_\sigma$ and indices tuples $k \in [\![s]\!]^\kappa$, $1 \leq \kappa \leq \alpha$, that

$$\|\mathrm{D}^k[\phi \circ g_i - \frac{1}{2}\|\bullet\|_2^2 \circ g_i]\|_\infty \lesssim G^{(i,\kappa)}\sigma^{3\kappa}\,,$$

where the implicit constant only depends on α and d.

Having uniform bounds on the partial derivatives, we arrive by combining Lemma 39, Lemma 40, and Lemma 43 at the following novel uniform metric entropy estimate.

Proposition 22. *Let Assumption* **(SG)** *hold and fix $\varepsilon = 1$. Then, it follows for $\delta > 0$,*

$$\log N(\delta,\, \mathcal{F}_\sigma,\, \|\bullet\|_\infty) \lesssim \left(\sum_{i=1}^I [G^{(i,\alpha)}]^{s/\alpha}\right)\sigma^{3s}\delta^{-s/\alpha}\,,$$

where the implicit constant only depends on α, d, s and $\mathcal{U}_1, \ldots, \mathcal{U}_I$.

This result enables us to derive an LCA result for the setting where one measure is compactly supported while the other is sub-Gaussian.

Theorem 23 (Partially unbounded LCA)**.** *Let Assumption* **(SG)** *hold and fix $\varepsilon = 1$. Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X}) \cap \mathrm{SG}_d(\sigma^2)$, $\nu \in \mathrm{SG}_d(\sigma^2)$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ from* (5) *it holds that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \left(\sum_{i=1}^I 1 + [G^{(i,\alpha)}]^{s/\alpha}\right) r^2\sigma^{4\alpha\vee(4+3s/2)}n^{-1/2}\,,$$

where the implicit constant only depends on α, d, s and $\mathcal{U}_1, \ldots, \mathcal{U}_I$. In particular, for $s \geq 8$ by Assumption **(SG)** *we must have $\alpha > 4$ and the exponent of σ equals 4α.*

Using Remark 45, we obtain the following result for more general $\varepsilon \neq 1$.

Corollary 24. *Let Assumption* **(SG)** *hold and let $\varepsilon > 0$. Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X}) \cap \mathrm{SG}_d(\sigma^2)$, $\nu \in \mathrm{SG}_d(\sigma^2)$ and any of the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ from* (5) *it holds that*

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim \left(\sum_{i=1}^I 1 + [G^{(i,\alpha)}]^{s/\alpha}\right) r^2\sigma^{4\alpha\vee(4+3s/2)}(\varepsilon\wedge 1)^{-[2\alpha\vee(2+3s/4)]-s/2}n^{-1/2}\,,$$

where the implicit constant only depends on α, d, s and $\mathcal{U}_1, \ldots, \mathcal{U}_I$. In particular, for $s \geq 8$ by Assumption **(SG)** *we must have $\alpha > 4$ and the exponent of $\varepsilon \wedge 1$ equals $-2\alpha - s/2$.*

Note that the implicit constant in the above bound still depends on the dimension d of the ground space $\mathbb{R}^d$. Nevertheless, only the dimension s of the (lower-dimensional) $\mathcal{U}_i$ influences the dependency of the bound on ε, σ^2 and the g_i. Hence, it also shows the validity of the LCA principle.

Further, observe that if the second measure ν is compactly supported, the setting of Theorem 18 is also satisfied (up to scaling of the cost function). However, as a trade-off, by staying under Assumption **(SG)** and not making use of the compactness of the support, we get a worse dependency in ε.

4. Computational Complexity

As seen in the previous section, there are several settings where the empirical EOT cost benefits from the LCA principle. However, in practice the plug-in estimator $\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)$ is in turn only *approximated* using the Sinkhorn algorithm. For this reason, we now investigate if we can construct a computable estimator that reflects the LCA principle. For the considerations to follow, we rely on the analysis of the Sinkhorn algorithm by Marino and Gerolin (2020) as well as Dvurechensky et al. (2018).

First, we briefly recall the Sinkhorn algorithm. Let $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures. Given a bounded start potential $\psi_0 : \mathcal{Y} \to \mathbb{R}$, the Sinkhorn algorithm can be viewed as approximating dual optimizers by alternatingly applying the entropic (c, ε)-transform to it (Marino and Gerolin, 2020). More precisely, for $m \in \mathbb{N}$ we recursively define

$$\phi_m \coloneqq \begin{cases} \psi_{m-1}^{(c,\varepsilon,\nu)} & m \text{ odd}, \\ \phi_{m-1} & m \text{ even}, \end{cases} \qquad \psi_m \coloneqq \begin{cases} \psi_{m-1} & m \text{ odd}, \\ \phi_{m-1}^{(c,\varepsilon,\mu)} & m \text{ even}. \end{cases}$$

Then, for $m \to \infty$ the pair (ϕ_m, ψ_m) converges to a pair of dual optimizers of (10) for μ and ν (Marino and Gerolin, 2020). To employ these potentials for a EOT cost estimator we define the transport plan

$$\mathrm{d}\pi^{(m)} \coloneqq \exp_\varepsilon(\phi_m \oplus \psi_m - c)\, \mathrm{d}[\mu \otimes \nu]\,,$$

whose marginal measures we denote by $\mu^{(m)}$ and $\nu^{(m)}$. Then, by definition of the (c, ε)-transform it follows that $\mu^{(m)} = \mu$ for odd m and $\nu^{(m)} = \nu$ for even m, asserting that $\mu^{(m)}$ and $\nu^{(m)}$ are always probability measures. In fact, this construction also yields that (ϕ_{m+1}, ψ_{m+1}) are for any $m \in \mathbb{N}$ optimal potentials for the measures $(\mu^{(m)}, \nu^{(m)})$ (see Lemma 38), which implies the representation

$$\mathrm{T}_{c,\varepsilon}(\mu^{(m)}, \nu^{(m)}) = \int \phi_{m+1}\, \mathrm{d}\mu^{(m)} + \int \psi_{m+1}\, \mathrm{d}\nu^{(m)}\,.$$

Further, by convergence of the potentials (ϕ_m, ψ_m) to dual optimizers for μ and ν as $m \to \infty$, the measures $(\mu^{(m)}, \nu^{(m)})$ also tend towards (μ, ν). A suitable termination criterion for the Sinkhorn algorithm is therefore to stop once the difference between $(\mu^{(m)}, \nu^{(m)})$ and (μ, ν) drops below a certain prespecified threshold. This difference can be quantified for odd m by the TV-norm $\|\mu^{(m)} - \mu\|_1$ and for even m by $\|\nu^{(m)} - \nu\|_1$.

Theorem 25 (Computational complexity)**.** *Let Assumption* **(C)** *hold and consider $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$. Denote with $\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} = \int \phi_{m+1}\, \mathrm{d}\hat{\mu}_n^{(m)} + \int \psi_{m+1}\, \mathrm{d}\hat{\nu}_n^{(m)}$ the empirical EOT cost estimator based on the Sinkhorn algorithm after $m+1$ iterations with empirical measures $\hat{\mu}_n$ and $\hat{\nu}_n$ as input. Furthermore, suppose for some $K_{\varepsilon,n} > 0$ that*

$$\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] \leq K_{\varepsilon,n}\,.$$

Then, for some deterministic $L_{\varepsilon,n} > 0$ we have after $m = \lfloor 2 + 20 L_{\varepsilon,n}^{-1}(3 \log n + \varepsilon^{-1}) \rfloor$ iterations,

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] \leq L_{\varepsilon,n} + K_{\varepsilon,n}\,.$$

Proof Invoking Lemma 38 from Appendix A.3 we know for any $m \in \mathbb{N}$ that the potentials (ϕ_{m+1}, ψ_{m+1}) are optimal for $(\hat{\mu}_n^{(m)}, \hat{\nu}_n^{(m)})$, asserting that $\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} = \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n^{(m)}, \hat{\nu}_n^{(m)})$. Further, recall that either $\hat{\mu}_n^{(m)} = \hat{\mu}_n$ or $\hat{\nu}_n^{(m)} = \hat{\nu}_n$, depending on the parity of m. We therefore consider w.l.o.g. the first case, i.e., $\hat{\mu}_n^{(m)} = \hat{\mu}_n$, the argument for the latter case is analogous. Using Lemma 4, we find that

$$\begin{aligned}
\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} - \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)|] &= \mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n^{(m)}) - \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)|] \\
&\leq 2\,\mathbb{E}\left[\sup_{\phi \in \mathcal{F}_{c,\varepsilon}} \left| \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\hat{\mu}_n)} \,\mathrm{d}[\hat{\nu}_n^{(m)} - \hat{\nu}_n] \right| \right] \\
&\leq 5\,\mathbb{E}[\|\hat{\nu}_n^{(m)} - \hat{\nu}_n\|_1]\,,
\end{aligned}$$

where the last step uses that $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\hat{\mu}_n)}$ is bounded in uniform norm by $5/2$. Choosing $\delta \coloneqq L_{\varepsilon,n}/5$, according to Theorem 1 from Dvurechensky et al. (2018) we can achieve

$$\|\hat{\nu}_n^{(m)} - \hat{\nu}_n\|_1 \leq \delta$$

after $m = \lfloor 2 + 4\delta^{-1}(\log n - \log \ell) \rfloor$ iterations, where we set

$$\ell \coloneqq \min_{i,j=1,\dots,n} \frac{1}{n^2} \exp_\varepsilon(-c(X_i, Y_j))\,.$$

As $\ell \geq n^{-2} \exp_\varepsilon(-\|c\|_\infty)$, it holds that $-\log \ell \leq 2 \log n + \varepsilon^{-1}$. In particular, this implies that

$$5\,\mathbb{E}[\|\hat{\nu}_n^{(m)} - \hat{\nu}_n\|_1] \leq L_{\varepsilon,n}$$

after $m = \lfloor 2 + 20 L_{\varepsilon,n}^{-1}(3 \log n + \varepsilon^{-1}) \rfloor$ iterations. Finally, by assumption and the triangle inequality, we conclude that

$$\begin{aligned}
\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] &\leq \mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} - \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)|] + \mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] \\
&\leq L_{\varepsilon,n} + K_{\varepsilon,n}\,.
\end{aligned}$$

■

Remark 26. *In the above proof, the inequality*

$$|\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)} - \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)| \leq L_{\varepsilon,n}$$

is even met deterministically *after $m = \lfloor 2 + 20 L_{\varepsilon,n}^{-1}(3 \log n + \varepsilon^{-1}) \rfloor$ iterations. We emphasize that this does not depend on the input measures $\hat{\mu}_n$ and $\hat{\nu}_n$.*

Remark 27. *For the dual perspective of the LCA principle (Theorem 6), we obtain with Theorem 25 that the Sinkhorn estimator $\widehat{\mathrm{T}}_{c,\varepsilon,n}^{(m)}$ after $m = \Omega(K_\varepsilon^{-1/2} n^{1/(k \vee 2)}[\log n + \varepsilon^{-1}])$ steps satisfies the statistical error rate $\mathcal{O}(K_\varepsilon^{1/2} n^{-1/(k \vee 2)})$ (up to log-terms). Since every Sinkhorn iteration encompasses $\mathcal{O}(n^2)$ arithmetic operations we conclude that the proposed estimator has a computational complexity of order $\mathcal{O}(K_\varepsilon^{-1/2} n^{2+1/(k \vee 2)}[\log n + \varepsilon^{-1}])$. Herein, we observe a trade-off between statistical accuracy and computational effort. Lastly, let us point out that Theorem 25 is also applicable to the upper bound (9) by Stromme (2023).*

Remark 28. *In Theorem 25, computation accuracy $L_{\varepsilon,n}$ can always be chosen to be of order $n^{-1/2}$. As the bound on the mean absolute deviation $K_{\varepsilon,n}$ is typically also at least of this order, we see that an estimator with similar statistical efficiency can be calculated at computational cost of order $\mathcal{O}(n^{2.5})$ (up to log-factors).*

Example 4. *Consider the setting of Example 2, i.e., c is the squared Euclidean norm. Then, for $d < 4$ we can obtain the statistical error rate given in the aforementioned example with computational complexity $\mathcal{O}(\varepsilon^{-1}n^{2.5})$ and else $\mathcal{O}(\varepsilon^{d/2-1}n^{2.5})$, if ε is small enough. Hence, in high dimensions d and for small regularization parameter ε we notice a reduced computational complexity due to the worse statistical accuracy.*

Overall, we conclude that the Sinkhorn algorithm for empirical measures outputs an estimator for the empirical EOT cost that also adheres to the LCA principle. Provided that the statistical error $K_{\varepsilon,n}$ is independent of the ambient dimension of the ground spaces, e.g., in settings where one measure is of low intrinsic dimension, Theorem 25 yields a well-computable estimator with good statistical accuracy.

5. Implications to the Entropic Gromov-Wasserstein Distance

The Gromov-Wasserstein distance provides an OT based tool to quantify the dissimilarity between two metric measure spaces (a metric space equipped with a probability measure) up to isometry[3] (Mémoli, 2011; Sturm, 2012). Hence, by modeling heterogeneous data as metric measure spaces the Gromov-Wasserstein distance serves as a conceptually appealing discrimination measure for registration invariant comparison, e.g., in the context of protein matching (Weitkamp et al., 2022). Unfortunately, practical usage of the Gromov-Wasserstein distance faces severe obstacles due to significant computational challenges. Indeed, for finitely supported measures computation of the Gromov-Wasserstein distance reduces to a non-convex quadratic assignment program, which are known to be NP-complete in general (Commander, 2005).

Motivated by the computational benefits of entropy regularization for the OT cost, the entropic Gromov-Wasserstein distance was introduced by Peyré et al. (2016); Solomon et al. (2016). To formalize it w.r.t. the Euclidean norm, let $\mathcal{X} \subseteq \mathbb{R}^s$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ be Polish subsets and consider $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ which admit finite fourth moments. Then, their entropic $(2,2)$-Gromov-Wasserstein distance for regularization parameter $\varepsilon > 0$ is defined as

$$\mathrm{GW}_\varepsilon(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \Bigg[\int_{\mathcal{X}\times\mathcal{Y}} \int_{\mathcal{X}\times\mathcal{Y}} |\|x - x'\|_2^2 - \|y - y'\|_2^2|^2 \, \pi(\mathrm{d}x, \mathrm{d}y) \, \pi(\mathrm{d}x', \mathrm{d}y') + \varepsilon \, \mathrm{KL}(\pi \mid \mu \otimes \nu) \Bigg].$$

The corresponding unregularized Gromov-Wasserstein distance is defined by omitting the entropy penalization term and denoted by GW_0.

Remarkably, despite $\mathrm{GW}_\varepsilon(\mu,\nu)$ also encompassing a non-convex optimization problem, the recent work by Rioux et al. (2023) show that it can be computed up to arbitrary precision

[3] Two metric measure spaces $(\mathcal{X}, \rho_\mathcal{X}, \mu)$ and $(\mathcal{Y}, \rho_\mathcal{Y}, \nu)$ are said to be isometric, if there exists a bijective isometry $\Gamma\colon (\mathcal{X}, \rho_\mathcal{X}) \mapsto (\mathcal{Y}, \rho_\mathcal{Y})$ such that $\Gamma_\# \mu = \nu$.

using a gradient method that employs the Sinkhorn algorithm. In particular, for input measures with n support points it admits a computational complexity of order $\mathcal{O}(n^2)$ (up to polylogarithmic terms). This makes it a viable tool for statistical data analysis. To analyze its sample complexity consider the empirical plug-in estimators

$$\widehat{\mathrm{GW}}_{\varepsilon,n} \in \{\mathrm{GW}_\varepsilon(\mu, \hat{\nu}_n), \mathrm{GW}_\varepsilon(\hat{\mu}_n, \nu), \mathrm{GW}_\varepsilon(\hat{\mu}_n, \hat{\nu}_n)\}\,. \tag{18}$$

In this context Zhang et al. (2022, Theorem 2) show for 4-sub-Weibull probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ with concentration parameter $\sigma > 0$ that

$$\mathbb{E}[|\widehat{\mathrm{GW}}_{\varepsilon,n} - \mathrm{GW}_\varepsilon(\mu,\nu)|] \lesssim (1+\sigma^4)n^{-1/2} + \varepsilon\left(1 + \left[\frac{\sigma}{\sqrt{\varepsilon}}\right]^{9\left\lceil \frac{s\vee d}{2}\right\rceil + 11}\right) n^{-1/2}\,. \tag{19}$$

In this upper bound, we see that the parametric rate is attained but the polynomial scaling in ε^{-1} depends on the maximum dimension of s and d. The analysis of the authors hinges on the following representation of the entropic Gromov-Wasserstein distance for centered probability measures μ, ν,

$$\mathrm{GW}_\varepsilon(\mu,\nu) = \mathrm{GW}_{1,1}(\mu,\nu) + \mathrm{GW}_{2,\varepsilon}(\mu,\nu)\,,$$

where the two terms on the right-hand side are defined as

$$\begin{aligned}
\mathrm{GW}_{1,1}(\mu,\nu) &\coloneqq \int_{\mathcal{X}\times\mathcal{X}} \|x-x'\|_2^4\,\mu(\mathrm{d}x)\,\mu(\mathrm{d}x') + \int_{\mathcal{Y}\times\mathcal{Y}} \|y-y'\|_2^4\,\nu(\mathrm{d}y)\,\nu(\mathrm{d}y') \\
&\qquad - 4\int_{\mathcal{X}\times\mathcal{Y}} \|x\|_2^2\|y\|_2^2\,\mu(\mathrm{d}x)\,\nu(\mathrm{d}y)\,, \\
\mathrm{GW}_{2,\varepsilon}(\mu,\nu) &\coloneqq \inf_{\pi\in\Pi(\mu,\nu)}\Bigg[-4\int_{\mathcal{X}\times\mathcal{Y}} \|x\|_2^2\|y\|_2^2\,\pi(\mathrm{d}x,\mathrm{d}y) - 8\sum_{i,j=1}^{s,d}\left(\int_{\mathcal{X}\times\mathcal{Y}} x_i y_j\,\pi(\mathrm{d}x,\mathrm{d}y)\right)^2 \\
&\qquad + \varepsilon\,\mathrm{KL}(\pi \mid \mu\otimes\nu)\Bigg]\,.
\end{aligned}$$

Note that this decomposition does not directly hold for the plug-in estimators (18) as the empirical distributions are in general not centered. To this end, Zhang et al. (2022) debias the empirical measures which contributes an additional $\sigma^2 n^{-1/2}$-term (see their Lemma 2). Then, the empirical plug-in estimator for the first term $\mathrm{GW}_{1,1}$ is a Monte-Carlo estimator and can be analyzed via V-statistics, contributing the first term on the right-hand side of (19). For the empirical estimator to the second term they link $\mathrm{GW}_{2,\varepsilon}$ to the EOT cost w.r.t. a class of cost functions and obtain statistical error bounds by controlling empirical processes uniformly over the class of cost functions. Applying an adjusted version of Lemma 5, we can confirm an LCA principle for the entropic $(2,2)$-Gromov-Wasserstein distance. As in Subsection 2.2, we require the cost(s) to be bounded and therefore impose the following compactness assumption.

Assumption (GW). *Let $\mathcal{X} \subseteq \mathbb{R}^s$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ be compact with $s \le d$ as well as* $\mathrm{diam}(\mathcal{X}) \le r$ *and* $\mathrm{diam}(\mathcal{Y}) \le r$ *for some $r \ge 1$.*

First, we give the aforementioned link between $\mathrm{GW}_{2,\varepsilon}$ and a collection of EOT costs.

Theorem 29 (Duality, Zhang et al. 2022, Theorem 1)**.** *Let Assumption* **(GW)** *hold. Denote the set of matrices* $\mathcal{D} \coloneqq \big[-r^2/2, r^2/2\big]^{s\times d}$ *and for* $A \in \mathcal{D}$ *define the cost function*

$$c_A : \mathcal{X} \times \mathcal{Y}\,, \qquad (x,y) \mapsto -4\|x\|_2^2\|y\|_2^2 - 32x^\mathsf{T} A y\,.$$

Then, it holds for all $\mu \in \mathcal{P}(\mathcal{X})$ *and* $\nu \in \mathcal{P}(\mathcal{Y})$ *that*

$$\mathrm{GW}_{2,\varepsilon}(\mu,\nu) = \min_{A\in\mathcal{D}} 32\|A\|_2^2 + \mathrm{T}_{c_A,\varepsilon}(\mu,\nu)\,.$$

Upon defining for arbitrary $\tilde{\mu} \in \mathcal{P}(\mathcal{X})$ the two function classes

$$\mathcal{F}_{\mathcal{D},\varepsilon} \coloneqq \bigcup_{A\in\mathcal{D}} \mathcal{F}_{c_A,\varepsilon}\,, \qquad \mathcal{F}_{\mathcal{D},\varepsilon}^{(\mathcal{D},\varepsilon,\tilde{\mu})} \coloneqq \bigcup_{A\in\mathcal{D}} \mathcal{F}_{\mathcal{D},\varepsilon}^{(c_A,\varepsilon,\tilde{\mu})}\,,$$

an application of Theorem 29 in combination with Lemma 4 yields the following stability bound for $\mathrm{GW}_{2,\varepsilon}$. The proof of this result and subsequent assertion of this section are detailed in Appendix A.4

Lemma 30 (Stability bound)**.** *Let Assumption* **(GW)** *hold. Then, it holds for any pairs of probability measures* $\mu, \tilde{\mu} \in \mathcal{P}(\mathcal{X})$ *and* $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{Y})$ *that*

$$\begin{aligned}|\mathrm{GW}_{2,\varepsilon}(\tilde{\mu},\tilde{\nu}) - \mathrm{GW}_{2,\varepsilon}(\mu,\nu)| &\le 2\sup_{A\in\mathcal{D}} |\mathrm{T}_{c_A,\varepsilon}(\tilde{\mu},\tilde{\nu}) - \mathrm{T}_{c_A,\varepsilon}(\mu,\nu)| \\ &\le 4\sup_{\phi\in\mathcal{F}_{\mathcal{D},\varepsilon}} \left|\int_{\mathcal{X}} \phi \,\mathrm{d}[\tilde{\mu}-\mu]\right| + 4\sup_{\psi\in\mathcal{F}_{\mathcal{D},\varepsilon}^{(\mathcal{D},\varepsilon,\tilde{\mu})}} \left|\int_{\mathcal{Y}} \psi \,\mathrm{d}[\tilde{\nu}-\nu]\right|.\end{aligned}$$

Hence, bounding $\mathbb{E}[|\mathrm{GW}_\varepsilon(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{GW}_\varepsilon(\mu,\nu)|]$ reduces to controlling two empirical processes over the function classes $\mathcal{F}_{\mathcal{D},\varepsilon}$ and $\mathcal{F}_{\mathcal{D},\varepsilon}^{(\mathcal{D},\varepsilon,\hat{\mu}_n)}$ (i.e., by selecting $\tilde{\mu} \coloneqq \hat{\mu}_n$). As the class of cost functions $\{c_A\}_{A\in\mathcal{D}}$ is Lipschitz continuous in A w.r.t. to the uniform norm $\|\bullet\|_\infty$, we observe that the union over entropic cost transforms do not increase the uniform metric entropy of a function class by much.

Lemma 31. *Let Assumption* **(GW)** *hold and let* $\tilde{\mu} \in \mathcal{P}(\mathcal{X})$*. Consider a function class* $\mathcal{F} \subseteq L_\varepsilon^{\exp}(\tilde{\mu})$ *on* $\mathcal{X}$*. Then, it holds for the union over* (c_A,ε)*-transformed function classes* $\mathcal{F}^{(\mathcal{D},\varepsilon,\tilde{\mu})} \coloneqq \bigcup_{A\in\mathcal{D}} \mathcal{F}^{(c_A,\varepsilon,\tilde{\mu})}$ *and any* $\delta > 0$ *that*

$$N(\delta,\, \mathcal{F}^{(\mathcal{D},\varepsilon,\tilde{\mu})},\, \|\bullet\|_\infty) \le N(\delta/4,\, \mathcal{F},\, \|\bullet\|_\infty) N(\delta/[64r^2],\, \mathcal{D},\, \|\bullet\|_\infty)\,. \tag{20}$$

A modified version of Theorem 6, adjusted to Lemma 31, yields the following LCA result.

Theorem 32 (Entropic Gromov-Wasserstein LCA)**.** *Let Assumption* **(GW)** *hold. Then, for all probability measures* $\mu \in \mathcal{P}(\mathcal{X})$*,* $\nu \in \mathcal{P}(\mathcal{Y})$ *and any of the empirical estimators* $\widehat{\mathrm{GW}}_{\varepsilon,n}$ *from* (18) *it holds that*

$$\mathbb{E}[|\widehat{\mathrm{GW}}_{\varepsilon,n} - \mathrm{GW}_\varepsilon(\mu,\nu)|] \lesssim r^4 n^{-1/2} + r^{4(s\wedge d)+4}(\varepsilon\wedge r^4)^{-(s\wedge d)/2} n^{-1/2}\,, \tag{21}$$

where the implicit constant only depends on s*,* d *and* $\mathcal{X}$*.*

Note that the implicit constant in (21) still depends on $s \vee d$. However, the dependency in ε^{-1} is determined by $s \wedge d$ and thus obeys the LCA principle.

Remark 33. *Following our arguments from Subsection 3.4, we would like to point out that Assumption* **(GW)** *could be refined to assuming that $\mathcal{X}$ is a union $\bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ for $I \in \mathbb{N}$ bounded, convex subsets $\mathcal{U}_i \subseteq \mathbb{R}^s$ with nonempty interior and maps $g_i : \mathcal{U}_i \to \mathcal{X}$ that are α-times continuously differentiable with bounded partial derivatives where $\alpha \in \mathbb{N}$. Then, for all probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ and any of the empirical estimators $\widehat{\mathrm{GW}}_{\varepsilon,n}$ from* (18) *it holds for sufficiently small ε that*

$$\mathbb{E}[|\widehat{\mathrm{GW}}_{\varepsilon,n} - \mathrm{GW}_{\varepsilon}(\mu,\nu)|] \lesssim \varepsilon^{-s/2} \begin{cases} n^{-1/2} & s/\alpha < 2\,, \\ n^{-1/2}\log(n+1) & s/\alpha = 2\,, \\ n^{-\alpha/s} & s/\alpha > 2\,. \end{cases}$$

An explicit dependency on the bounds for the partial derivatives can be obtained along the lines of Lemma 16 and is omitted here.

Remark 34 (Unregularized Gromov-Wasserstein LCA)**.** *Zhang et al. (2022) also derive a representation of the unregularized $(2,2)$-Gromov-Wasserstein distance as an infimum of unregularized OT costs over a suitable class of cost functions. Hence, invoking methods for statistical error bounds on the empirical unregularized OT cost by Hundrieser et al. (2024b, Section 3.3), the validity of the LCA principle can also be confirmed for the unregularized $(2,2)$-Gromov-Wasserstein distance. More specifically, under Assumption* **(GW)** *it follows,*

$$\mathbb{E}[|\widehat{\mathrm{GW}}_{0,n} - \mathrm{GW}_{0}(\mu,\nu)|] \lesssim \begin{cases} n^{-1/2} & s \wedge d < 4\,, \\ n^{-1/2}\log(n+1) & s \wedge d = 4\,, \\ n^{-2/(s\wedge d)} & s \wedge d > 4\,, \end{cases}$$

where the implicit constant only depends on s, d, r and $\mathcal{X}$. A proof is detailed in Section A.

6. Simulations

In the previous sections, we analyzed various settings where the LCA principle holds for empirical EOT. We now investigate whether the LCA principle can also be observed numerically. More specifically, for probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, various ε and one of the empirical estimators from (5) we approximate the mean absolute deviation

$$\Delta_n \coloneqq \mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|]$$

for different sample sizes n by Monte Carlo simulations with 1000 repetitions. Due to a lack of explicit formulas, the population quantity $\mathrm{T}_{c,\varepsilon}(\mu,\nu)$ is also approximated using Monte Carlo simulations with 1000 repetitions and a large sample size specified below for each simulation setting. The empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ are approximated via the Sinkhorn algorithm such that the error of the first marginal is less than 10^{-8} w.r.t. the norm $\|\bullet\|_1$.[4]

[4]The code used for our simulations can be found under `https://gitlab.gwdg.de/michel.groppe/eot-lca-simulations`.

We also consider the Sinkhorn divergence in our simulations. Assuming that $\mathcal{X} = \mathcal{Y}$ and thus $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the Sinkhorn divergence between μ and ν is defined as

$$S_{c,\varepsilon}(\mu, \nu) \coloneqq \mathrm{T}_{c,\varepsilon}(\mu, \nu) - \frac{1}{2}\,\mathrm{T}_{c,\varepsilon}(\mu, \mu) - \frac{1}{2}\,\mathrm{T}_{c,\varepsilon}(\nu, \nu)\,.$$

The last two terms have a debiasing effect such that $S_{c,\varepsilon}(\mu, \nu) = 0$ for $\mu = \nu$. Under certain assumptions on the cost function, the Sinkhorn divergence is even positive definite (Feydy et al., 2019). Moreover, by its definition it is not clear if it also benefits from the LCA principle. Indeed, all available bounds for the statistical error of $\mathrm{T}_{c,\varepsilon}(\hat{\nu}_n, \hat{\nu}_n)$ depend at least on the intrinsic dimension of ν and do not suggest that the LCA principle holds.

Overall, we examine the following simulation settings:

1. Cube: For $d_1 \in [\![10]\!]$ and $d_2 = 5$ take

   $$\mu = \mathrm{Unif}([0,1]^{d_1} \times \{0\}^{d_1 \vee d_2 - d_1})\,, \qquad \nu = \mathrm{Unif}([0,1]^{d_2} \times \{0\}^{d_1 \vee d_2 - d_2})$$

   and as the cost the by $d_1 \vee d_2$ normalized squared Euclidean norm $\|\bullet\|_2^2$ or 1-norm $\|\bullet\|_1$. For $n \in \{100k \mid k \in [\![10]\!]\}$ we compute the two-sample estimator $\widehat{\mathrm{T}}_{c,\varepsilon,n} = \mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)$. The true value $\mathrm{T}_{c,\varepsilon}(\mu, \nu)$ is approximated using $n = 6000$ samples.

2. Semi-discrete: For each $I \in \{5, 10, 50\}$ and $d \in \{10, 100, 1000\}$, we define

   $$\mu = \frac{1}{I}\sum_{i=1}^{I} \delta_{x_i^{(I,d)}}\,, \qquad \nu = \mathrm{Unif}[0,1]^d\,,$$

   where $x_1^{(I,d)}, \dots, x_I^{(I,d)}$ are fixed and drawn i.i.d. from $\mathrm{Unif}[0,1]^d$, and as the cost the uniform norm $\|\bullet\|_\infty$. Now, for $n \in \{100, \dots, 5000\}$ we calculate the one-sample estimator $\mathrm{T}_{c,\varepsilon}(\mu, \hat{\nu}_n)$ and use $n = 20000$ samples to approximate $\mathrm{T}_{c,\varepsilon}(\mu, \nu)$.

3. Sinkhorn divergence: We again consider the cube setting with cost $\|\bullet\|_2^2$ but instead of the EOT cost $\mathrm{T}_{c,\varepsilon}$ we employ the Sinkhorn divergence $S_{c,\varepsilon}$.

Figure 1 and Figure 2 show the results for the cube setting with smooth cost $\|\bullet\|_2^2$ and non-smooth $\|\bullet\|_1$, respectively. As they are very similar we therefore focus on the former. In particular, we see for fixed $\varepsilon > 0$ that Δ_n roughly has a convergence rate of order $n^{-1/2}$, which is in line with Example 2 for $\|\bullet\|_2^2$, and the combination of Rigollet and Stromme (2022) (recall (6d)) with our projective perspective on the LCA principle (Subsection 2.3) for $\|\bullet\|_1$. Furthermore, we observe that the underlying constant decreases from $d_1 = 1$ to $d_1 = 5$, and from $d_1 = 6$ to $d_1 = 10$ approximately stays the same. In the latter case, it seems as if the constant only depends on the smaller dimension $d_2 = 5$, thus corroborating the LCA principle.

At first glance the behavior of the constant for $d_1 = 1$ to $d_1 = 5$ appears surprising as it decreases with growing dimension. This behavior can be explained with Proposition 11. Here (and also similarly with $\|\bullet\|_1$ as cost), we have $\mathcal{X} = [0,1]^{d_1}$, $\mathcal{Y}_1 = \mathcal{X}$ and $\mathcal{Y}_2 = [0,1]^{d_2 - d_1}$ with $c_1(x, y_1) = \|x - y_1\|_2^2$ and $c_2(y_2) = \|y_2\|_2^2$. Note, that $\mathcal{Y}_2$ has the highest dimension for $d_1 = 1$ and is empty for $d_1 = 5 = d_2$. Hence, the dependence on the additional part c_2 decreases from $d_1 = 1$ to $d_1 = 5$. This suggests that the constant for the statistical

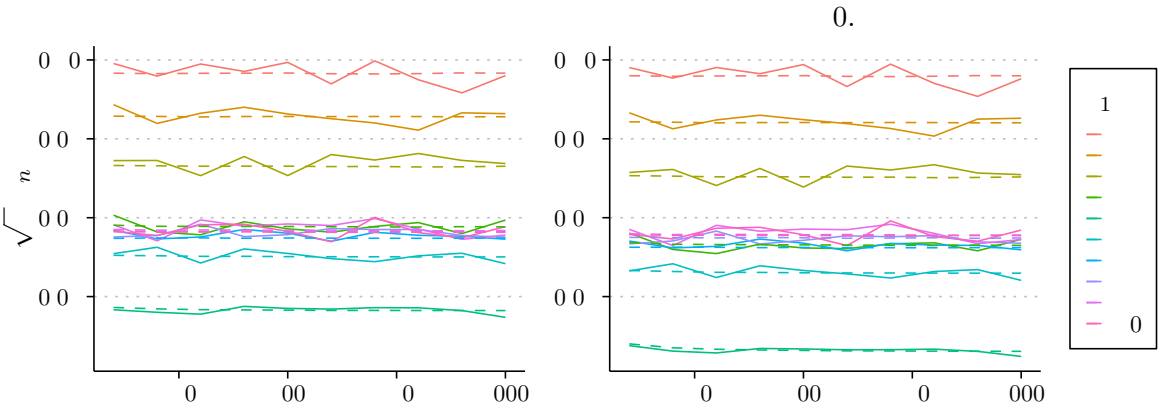

Figure 1: Simulations of the mean absolute deviation Δ_n (solid) and the by $\sqrt{2/\pi}$ scaled asymptotic standard deviation of the fluctuations $\sqrt{n}[\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)]$ (dashed) in the cube setting with cost $\|\cdot\|_2^2$.

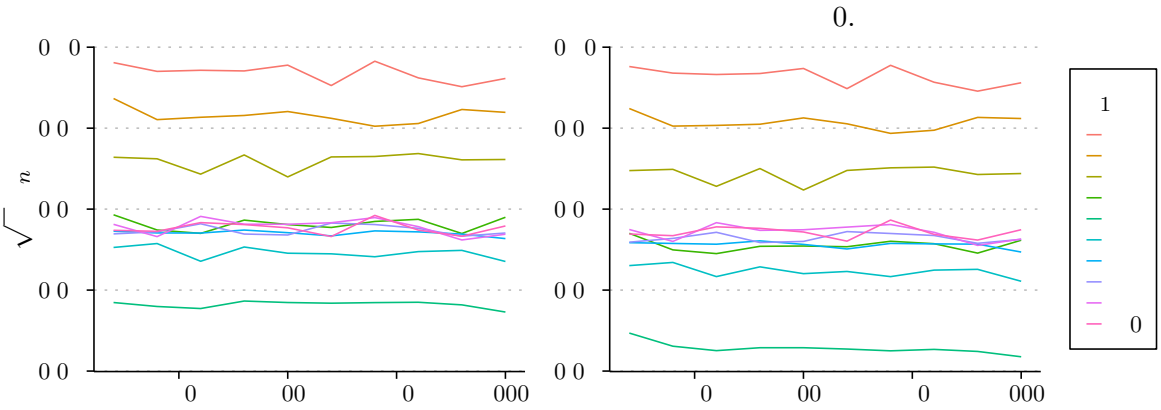

Figure 2: Simulations of the mean absolute deviation Δ_n in the cube setting with cost $\|\cdot\|_1$.

error of the integral term $\int_{\mathcal{Y}_2} c_2 \, \mathrm{d}\hat{\nu}_{n,2}$ dominates the one for $\mathrm{T}_{c_1,\varepsilon}(\hat{\mu}_n, \hat{\nu}_{n,1})$. Further, note that in this setting the fluctuations $\sqrt{n}[\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)]$ asymptotically admit a zero-mean normal distribution with variance σ^2 equal to the sum of the variances of the optimal potentials (González-Sanz and Hundrieser, 2023, Theorem 1.1). Hence, we roughly have that $\sqrt{n}\Delta_n \approx \sqrt{2/\pi}\sigma$. Approximating said variance within our Monte Carlo simulation, we see in Figure 1 that the scaled standard deviation seems to behave similar to the statistical

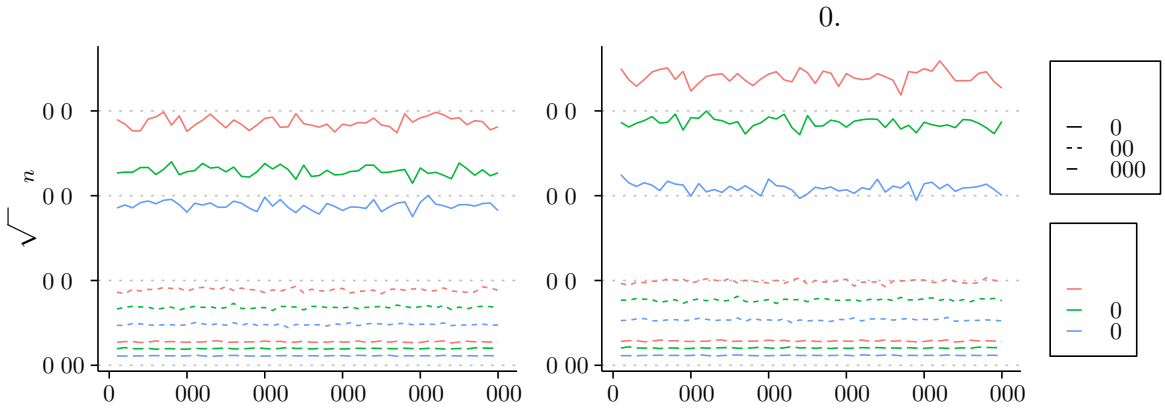

Figure 3: Simulations of the mean absolute deviation Δ_n in the semi-discrete setting.

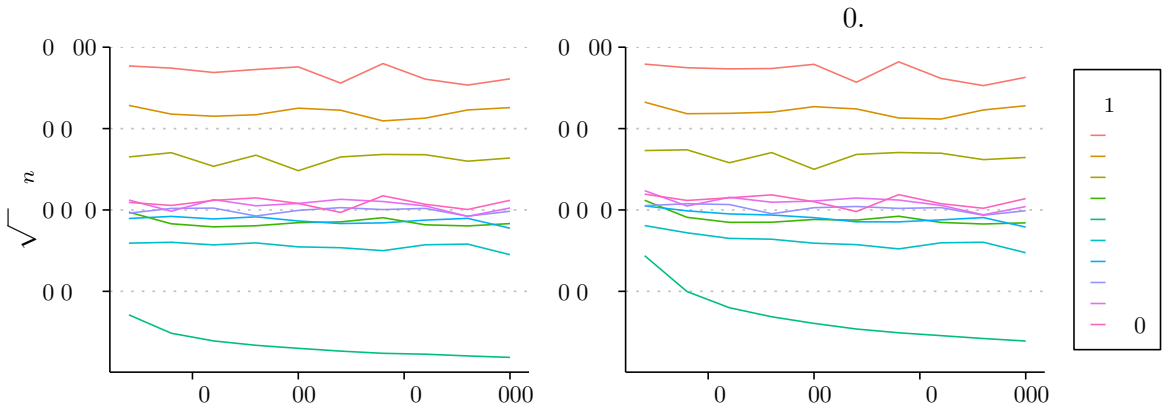

Figure 4: Simulations of the mean absolute deviation Δ_n for the Sinkhorn divergence.

error Δ_n in Figure 1 (which also explains the latter). In particular, the variance obeys the LCA principle.

Figure 3 showcases the simulation results for the semi-discrete setting. We observe that the mean absolute deviation Δ_n decreases with higher d and I. This indicates that the normalization by $d_1 \vee d_2$ is only proper in the sense that it achieves $\|c\|_\infty \leq 1$, but does not capture the sharp dependency in the underlying constant for the convergence. A possible explanation for the behavior in I is due to the fact that with more points (I) the expected distance at which mass is assigned with respect to the uniform norm $\|\bullet\|_\infty$ decreases. Meanwhile, for increasing dimension (d) the average distance between two random points in the unit cube increases to one, and thus the cost function tends to be more homogenous with increasing dimension. Note in particular that for a constant cost function $c(x, y) \equiv a$, the entropic OT cost fulfills $T_{c,\varepsilon}(\mu, \nu) = a$ for every $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\varepsilon > 0$. This indicates

that statistical problem of estimating the population EOT cost in the semi-discrete setting rather simplifies as the dimension increases. In particular, this behavior is remarkable as it suggests for the semi-discrete setting that the empirical EOT cost might benefit from additional structure.

Lastly, Figure 4 shows the results for the Sinkhorn divergence. We see a behavior which is in line with the cube setting, except for $d_1 = 5$ where the error decreases with increasing sample size. This is a consequence of the fact that the Sinkhorn divergence converges at the rate n^{-1} under identical population measures (González-Sanz et al., 2022; Goldfeld et al., 2024b). Moreover, the Sinkhorn divergence also seems to be affected by the LCA principle. However, recalling the discussion of the cube setting, note that in this case for the debiasing terms $\mathrm{T}_{c,\varepsilon}(\mu,\mu)$ and $\mathrm{T}_{c,\varepsilon}(\nu,\nu)$ in the setting of Proposition 11 the additional part c_2 is zero. As a consequence, the estimation error caused by the debasing terms appears for these medium dimensional settings rather negligible in comparison to the estimation error for the estimation of $\mathrm{T}_{c,\varepsilon}(\mu,\nu)$.

7. Discussion

In this work, we showed that the empirical EOT cost, just like the empirical OT cost, adheres to the LCA principle. More precisely, for suitably bounded costs, we derived an upper bound for the statistical estimation error of the empirical EOT cost in terms of n and ε which only depends on the simpler probability measure. We stress that this holds for the empirical EOT estimator and no additional knowledge of the underlying space is necessary, i.e., the estimator automatically adapts. Further, we observed that the empirical EOT cost can be approximated using the Sinkhorn algorithm in at most $\mathcal{O}(n^{2.5})$ arithmetic operations such that the resulting quantity still obeys the LCA principle. Most of our results are derived under boundedness of the cost function. This allows us to leverage bounds on the uniform metric entropy of the function class $\mathcal{F}_{c,\varepsilon}$ which in turn is needed for the crucial Lemma 5. Nevertheless, Theorem 23 shows that the LCA principle can also hold in a (partially) unbounded setting. As of now, this setting is limited to the squared Euclidean norm. We leave a more extensive analysis in the case of unbounded costs to future work.

Such an analysis will most likely entail the use of concentration constraints for the probability measures that are tailored to the cost function (like sub-Gaussianity for the squared Euclidean norm). More specifically, based on the convergence behavior for empirical OT in unbounded settings (Fournier and Guillin, 2015; Staudt and Hundrieser, 2024), we conjecture for non-negative costs dominated by $c(x,y) \leq \kappa\left(\|x\|^p + \|y\|^p\right)$ for some $\kappa > 0$ that finite moments of order $2p + \delta$ for $\delta > 0$ for the two population measures are sufficient to infer parametric convergence rates in n. We also conjecture that under sufficient moment concentration the entropic LCA principle manifests for totally unbounded settings, i.e., that the dependency in ε only depends on the smoothness of the cost function and the minimum intrinsic dimension of the two population measures.

Moreover, for a complete statistical analysis of the empirical EOT cost it is critical to also obtain complementing lower bounds on the convergence rates. Such lower bounds will likely depend on the dual optimizers for empirical and population measures. Indeed, recently derived distributional limits by González-Sanz and Hundrieser (2023) assert that the empirical estimator $\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)$ asymptotically fluctuates around its population counterpart $\mathrm{T}_{c,\varepsilon}(\mu,\nu)$

at variance $\mathrm{Var}_{X\sim\mu}[\phi(X)] + \mathrm{Var}_{Y\sim\nu}[\psi(Y)]$, where (ϕ, ψ) are optimal EOT potentials for μ and ν. For fixed $\varepsilon > 0$ this implies the parametric rate $n^{-1/2}$ to be sharp in n, nevertheless, the sharp dependency of the mean absolute deviation in terms of ε still remains open.

Acknowledgments

The research of M. Groppe and S. Hundrieser is supported by the Research Training Group 2088 "*Discovering structure in complex data: Statistics meets Optimization and Inverse Problems*", which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The authors thank Axel Munk and Marcel Klatt for fruitful discussions. Further, the authors acknowledge helpful comments by three anonymous referees, and Gonzalo Mena for spotting a mistake in a previous version of Theorem 23.

Appendix A. Omitted Proofs

A.1 Duality and Complexity

Proof of Proposition 3. First, note by Assumption **(C)** that $\mathcal{F}_{c,\varepsilon} \subseteq L_\varepsilon^{\exp}(\mu)$ and $\mathcal{F}_{c,\varepsilon}^{(c,\varepsilon,\mu)} = \{\phi^{(c,\varepsilon,\mu)} \mid \phi \in \mathcal{F}_{c,\varepsilon}\} \subseteq L_\varepsilon^{\exp}(\nu)$. Furthermore, from Remark 2 we know that there exists a maximizing pair ϕ, ψ of (10) such that

$$\phi = \psi^{(c,\varepsilon,\nu)} \quad \text{and} \quad \psi = \phi^{(c,\varepsilon,\mu)}$$

as well as $\|\phi\|_\infty, \|\psi\|_\infty \leq 3/2$. This implies that $\phi \in \mathcal{F}_{c,\varepsilon}$ and thus

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = D_{c,\varepsilon}^{\mu,\nu}(\phi,\psi) = \max_{\phi\in\mathcal{F}_{c,\varepsilon}} D_{c,\varepsilon}^{\mu,\nu}(\phi, \phi^{(c,\varepsilon,\mu)})\,.$$

Moreover, by the Tonelli-Fubini theorem we have for $\phi \in \mathcal{F}_{c,\varepsilon}$ that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{Y}} \exp_\varepsilon(\phi(x) + \phi^{(c,\varepsilon,\mu)}(y) - c(x,y))\,\mu(\mathrm{d}x)\,\nu(\mathrm{d}y) \\
&\qquad = \int_{\mathcal{Y}} \exp_\varepsilon(\phi^{(c,\varepsilon,\mu)}(y)) \left[\int_{\mathcal{X}} \exp_\varepsilon(\phi(x) - c(x,y))\,\mu(\mathrm{d}x)\right] \nu(\mathrm{d}y) \\
&\qquad = \int_{\mathcal{Y}} \exp_\varepsilon(\phi^{(c,\varepsilon,\mu)}(y)) \exp_\varepsilon(-\phi^{(c,\varepsilon,\mu)}(y))\,\nu(\mathrm{d}y) = 1\,,
\end{aligned}$$

which yields

$$D_{c,\varepsilon}^{\mu,\nu}(\phi, \phi^{(c,\varepsilon,\mu)}) = \int_{\mathcal{X}} \phi\,\mathrm{d}\mu + \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\mu)}\,\mathrm{d}\nu\,,$$

and we conclude (13). ∎

Proof of Lemma 4. We follow the proof of Proposition 2 from Mena and Niles-Weed (2019). First, according to Remark 2 there are $\phi, \tilde{\phi} \in \mathcal{F}_{c,\varepsilon}$ with

$$\begin{aligned}
\phi &= \psi^{(c,\varepsilon,\nu)}\,, & \psi &= \phi^{(c,\varepsilon,\mu)}\,, \\
\tilde{\phi} &= \tilde{\psi}^{(c,\varepsilon,\nu)}\,, & \tilde{\psi} &= \tilde{\phi}^{(c,\varepsilon,\tilde{\mu})}\,,
\end{aligned}$$

such that

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi)\,, \qquad \mathrm{T}_{c,\varepsilon}(\tilde{\mu},\nu) = D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi})\,.$$

By optimality, it holds that

$$\begin{aligned} D^{\mu,\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi}) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi}) &\leq D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi}) \\ &\leq D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\phi,\psi)\,, \end{aligned}$$

which implies

$$\begin{aligned} |\mathrm{T}_{c,\varepsilon}(\mu,\nu) - \mathrm{T}_{c,\varepsilon}(\tilde{\mu},\nu)| &= |D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi})| \\ &\leq |D^{\mu,\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi}) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi})| \\ &\quad + |D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\phi,\psi)|\,. \end{aligned}$$

As $\tilde{\phi} = \tilde{\psi}^{(c,\varepsilon,\nu)}$, we obtain using the Tonelli-Fubini theorem that

$$\begin{aligned} &\int_{\mathcal{X}\times\mathcal{Y}} \exp_\varepsilon(\tilde{\phi}\oplus\tilde{\psi} - c)\,\mathrm{d}[(\mu-\tilde{\mu})\otimes\nu] \\ &\qquad = \int_{\mathcal{X}} \exp_\varepsilon(\tilde{\phi}(x)) \int_{\mathcal{Y}} \exp_\varepsilon(\tilde{\psi}(y) - c(x,y))\,\nu(\mathrm{d}y)\,[\mu-\tilde{\mu}](\mathrm{d}x) \\ &\qquad = \int_{\mathcal{X}} \exp_\varepsilon(\tilde{\phi}(x)) \exp_\varepsilon(-\tilde{\phi}(x))\,[\mu-\tilde{\mu}](\mathrm{d}x) = 0\,, \end{aligned}$$

which yields for the first term

$$|D^{\mu,\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi}) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\tilde{\phi},\tilde{\psi})| = \left|\int_{\mathcal{X}} \tilde{\phi}\,\mathrm{d}[\mu-\tilde{\mu}]\right| \leq \sup_{\phi\in\mathcal{F}_{c,\varepsilon}} \left|\int_{\mathcal{X}} \phi\,\mathrm{d}[\mu-\tilde{\mu}]\right|.$$

Analogously, we get the same bound for the second term $|D^{\mu,\nu}_{c,\varepsilon}(\phi,\psi) - D^{\tilde{\mu},\nu}_{c,\varepsilon}(\phi,\psi)|$ and thus

$$|\mathrm{T}_{c,\varepsilon}(\mu,\nu) - \mathrm{T}_{c,\varepsilon}(\tilde{\mu},\nu)| \leq 2 \sup_{\phi\in\mathcal{F}_{c,\varepsilon}} \left|\int_{\mathcal{X}} \phi\,\mathrm{d}[\mu-\tilde{\mu}]\right|.$$

Similarly, we obtain that

$$|\mathrm{T}_{c,\varepsilon}(\tilde{\mu},\nu) - \mathrm{T}_{c,\varepsilon}(\tilde{\mu},\tilde{\nu})| \leq 2 \sup_{\phi\in\mathcal{F}_{c,\varepsilon}} \left|\int_{\mathcal{Y}} \phi^{(c,\varepsilon,\tilde{\mu})}\,\mathrm{d}[\nu-\tilde{\nu}]\right|.$$

Using the triangle inequality and combining the two bounds, we obtain the assertion. ∎

Proof of Lemma 5. Note by assumption on $\mathcal{F}$ that its (c,ε)-transform is well-defined. W.l.o.g. we can assume that $N \coloneqq N(\delta/2, \mathcal{F}, \|\bullet\|_\infty) < \infty$, else the asserted inequality is vacuous. Let $\{\tilde{\phi}_1,\dots,\tilde{\phi}_N\}$ be a $\delta/2$-covering for $\mathcal{F}$ w.r.t $\|\bullet\|_\infty$. For each $i=1,\dots,N$, we can pick a $\phi_i\in\mathcal{F}$ such that $\|\phi_i - \tilde{\phi}_i\|_\infty \leq \delta/2$. Using the triangle inequality, we see that $\{\phi_1,\dots,\phi_N\}$ is a δ-covering for $\mathcal{F}$ w.r.t. $\|\bullet\|_\infty$. By construction, note that every function of the δ-covering is measurable and its entropic transform well-defined. Let $\phi\in\mathcal{F}$, then there exists a ϕ_i such that $\|\phi-\phi_i\|_\infty\leq\delta$ or equivalently $\phi_i-\delta\leq\phi\leq\phi_i+\delta$. The monotonicity of the entropic (c,ε)-transform yields that

$$\phi_i^{(c,\varepsilon,\tilde{\mu})} - \delta = (\phi_i+\delta)^{(c,\varepsilon,\tilde{\mu})} \leq \phi^{(c,\varepsilon,\tilde{\mu})} \leq (\phi_i-\delta)^{(c,\varepsilon,\tilde{\mu})} = \phi_i^{(c,\varepsilon,\tilde{\mu})} + \delta\,.$$

Hence, we see that $\|\phi^{(c,\varepsilon,\tilde{\mu})} - \phi_i^{(c,\varepsilon,\tilde{\mu})}\|_\infty \leq \delta$ and $\{\phi_1^{(c,\varepsilon,\tilde{\mu})},\dots,\phi_N^{(c,\varepsilon,\tilde{\mu})}\}$ is a δ-covering for $\mathcal{F}^{(c,\varepsilon,\tilde{\mu})}$ w.r.t. $\|\bullet\|_\infty$. ∎

Remark 35. *In Lemma 5 and its proof the scale $\delta/2$ appears in $N(\delta/2, \mathcal{F}, \|\cdot\|_\infty)$ to sidestep potential measurability issues. Indeed, if we knew that the optimal covering of $\mathcal{F}$ for every δ consisted only of measurable functions, then the factor $1/2$ would not be needed. However, this is not the case, in general. The measurability is crucial since the entropic (c,ε)-transform requires the integration of the function.*

A.2 Sample Complexity

W.l.o.g. we assume in the following proofs that $I = 1$ and $g_1 = \mathrm{id}_{\mathcal{X}}$, i.e., $\mathcal{U}_1 = \mathcal{X}$. Based on Lemma 40 and Lemma 39 this is not a genuine restriction.

Proof of Theorem 13. By definition, it holds for $\phi \in \mathcal{F}_{c,\varepsilon}$ that $\|\phi\|_\infty \leq 3/2$. Hence, it follows for all $i = 1, \dots, I$ that

$$N(\delta, \mathcal{F}_{c,\varepsilon}\big|_{\{x_i\}}, \|\cdot\|_\infty) \leq \lceil 3/\delta \rceil$$

and using the union bound Lemma 39 we obtain

$$\log N(\delta, \mathcal{F}_{c,\varepsilon}, \|\cdot\|_\infty) \lesssim I\delta^{-1}\,.$$

An application of Theorem 6 yields the assertion. ∎

Proof of Theorem 14. Note that by assumption c is 1-Lipschitz in the first component. By Proposition 2.4 from Marino and Gerolin (2020), it follows that $\psi^{(c,\varepsilon,\nu)} \in \mathcal{F}_{c,\varepsilon,\nu}$ is also 1-Lipschitz. Hence, the class $\mathcal{F}_{c,\varepsilon}$ is contained in the set of uniformly bounded 1-Lipschitz functions. An application of Lemma 41 combined with Theorem 6 yields the assertion. ∎

Proof of Theorem 15. Upon defining

$$\tilde{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\,, \qquad (x,y) \mapsto c(x,y) - \|x\|_2^2\,,$$

1-semi-concavity of c means that for all $y \in \mathcal{Y}$, $t \in (0,1)$ and $x_1, x_2 \in \mathcal{X}$ it holds that

$$\tilde{c}(tx_1 + (1-t)x_2, y) \geq t\tilde{c}(x_1, y) + (1-t)\tilde{c}(x_2, y)\,.$$

This and the Hölder inequality with $p \coloneqq 1/t$, $q \coloneqq 1/(1-t)$ yield for $\psi^{(c,\varepsilon,\xi)} \in \mathcal{F}_{c,\varepsilon}$ that

$$\begin{aligned}
&\int_{\mathcal{Y}} \exp_\varepsilon(\psi(y) - \tilde{c}(tx_1 + (1-t)x_2, y))\,\xi(\mathrm{d}y) \\
&\qquad \leq \int_{\mathcal{Y}} \exp_\varepsilon(\psi(y) - \tilde{c}(x_1, y))^t \exp_\varepsilon(\psi(y) - \tilde{c}(x_2, y))^{1-t}\,\xi(\mathrm{d}y) \\
&\qquad \leq \left[\int_{\mathcal{Y}} \exp_\varepsilon(\psi(y) - \tilde{c}(x_1, y))\,\xi(\mathrm{d}y)\right]^t \left[\int_{\mathcal{Y}} \exp_\varepsilon(\psi(y) - \tilde{c}(x_2, y))\,\xi(\mathrm{d}y)\right]^{1-t}\,,
\end{aligned}$$

and therefore

$$\psi^{(\tilde{c},\varepsilon,\xi)}(tx_1 + (1-t)x_2) \geq t\psi^{(\tilde{c},\varepsilon,\xi)}(x_1) + (1-t)\psi^{(\tilde{c},\varepsilon,\xi)}(x_2)\,.$$

Since for any $x \in \mathcal{X}$ it holds that

$$\psi^{(\tilde{c},\varepsilon,\xi)}(x) = (\psi + \|\cdot\|_2^2)^{(c,\varepsilon,\xi)}(x) = \psi^{(c,\varepsilon,\xi)}(x) - \|x\|_2^2\,,$$

we conclude that $\psi^{(c,\varepsilon,\xi)}$ is 1-semi-concave. Further, note that Assumption **(SC)** includes Assumption **(Lip)**. Hence, $\mathcal{F}_{c,\varepsilon}$ is contained in the set of bounded, 1-Lipschitz and 1-semi-concave functions. Lemma 42 combined with Theorem 6 thus yield the assertion. ∎

Lemma 36 (Structure of derivatives, Genevay et al. 2019, Lemma 1)**.** *Let Assumption* **(C)** *and Assumption* **(Hol)** *(w.l.o.g. $I = 1$ and $g_1 = \mathrm{id}_{\mathcal{X}}$) hold. Then, for $\phi = \psi^{(c,\varepsilon,\xi)} \in \mathcal{F}_{c,\varepsilon}$, $k \in [\![s]\!]^\kappa$, $\kappa \le \alpha$, and $x \in \mathring{\mathcal{X}}$ it follows that*

$$\mathrm{D}^k \phi(x) = \int_{\mathcal{Y}} \Phi_\kappa^k(x,y) \gamma_\varepsilon(x,y)\, \xi(\mathrm{d}y)\,,$$

where

$$\gamma_\varepsilon \coloneqq \exp_\varepsilon(\phi \oplus \psi - c)\,, \qquad \Phi_1^k \coloneqq \frac{\partial}{\partial x_{k_1}} c\,,$$

and for $m = 2, \dots, \kappa$ we set

$$\Phi_m^k \coloneqq \frac{\partial}{\partial x_{k_m}} \Phi_{m-1}^k + \frac{1}{\varepsilon}\left[\frac{\partial}{\partial x_{k_m}}\phi - \frac{\partial}{\partial x_{k_m}} c\right] \Phi_{m-1}^k\,.$$

Proof of Lemma 16. W.l.o.g. we can assume that $g_i = \mathrm{id}_{\mathcal{X}}$ and write $C_m = C_{i,m}$ and $C^{(m)} = C^{(i,m)}$. Recall Lemma 36. We show that for all $k \in [\![d]\!]^\kappa$, $\kappa \le \alpha$, $m = 0, \dots, \kappa - 2$ and indices tuple $|j| = 0, \dots, m$ it holds that

$$\|\mathrm{D}^j \Phi_{\kappa-m}^k\|_\infty \lesssim (\varepsilon \wedge 1)^{-(\kappa - m + |j| - 1)} C^{(\kappa - m + |j|)}\,,$$

where the implicit constant only depends on κ. Note that the above bound also holds for $\kappa = 1$ by definition of Φ_1^k. Then, we can conclude by noting that

$$\|\mathrm{D}^k \phi\|_\infty \le \|\Phi_\kappa^k\|_\infty\,.$$

We follow the proof of Lemma 2 from Genevay et al. (2019) and prove this by induction over $\kappa = |k|$. W.l.o.g. we can assume that $0 < \varepsilon \le 1$. Indeed, for $\varepsilon > 1$ in all the following manipulations the term $1/\varepsilon$ can be bounded by 1.

We have the base case $\kappa = 2$ which implies $m = 0$ and $|j| = 0$. Hence,

$$\begin{aligned} |\mathrm{D}^j \Phi_{\kappa-m}^k| = |\Phi_2^k| &= \left| \frac{\partial}{\partial x_{k_2}} \frac{\partial}{\partial x_{k_1}} c + \frac{1}{\varepsilon}\left[\frac{\partial}{\partial x_{k_2}}\phi - \frac{\partial}{\partial x_{k_2}} c\right] \frac{\partial}{\partial x_{k_1}} c \right| \\ &\le C_2 + \frac{1}{\varepsilon}[C_1 + C_1] C_1 \lesssim \varepsilon^{-1} C^{(2)} = \varepsilon^{-(\kappa - m + |j| - 1)} C^{(\kappa - m + |j|)}\,. \end{aligned}$$

We now suppose the validity of the assertion for given κ and verify it for $\kappa + 1$. To this end, assume that we have

$$\|\mathrm{D}^j \Phi_{\kappa-m}^k\|_\infty \lesssim \varepsilon^{-(\kappa - m + |j| - 1)} C^{(\kappa - m + |j|)}$$

for all $|k| = \kappa$, $m = 0, \dots, \kappa - 2$ and $|j| = 0, \dots, m$, and we want to extend this to $\|\mathrm{D}^j \Phi_{\kappa+1-m}^k\|_\infty$ for $|k| = \kappa + 1$, $m = 0, \dots, \kappa - 1$ and $|j| = 0, \dots, m$.

Denote with k' the first κ components of k. Then, as Φ_m^k only depends on the first m components of k, note that $\Phi_m^k = \Phi_m^{k'}$ for all $m = 0, \dots, \kappa$. Hence, it follows for $m = 1, \dots, \kappa - 1$ and $|j| = 0, \dots, m - 1$ by the induction assumption that

$$\|\mathrm{D}^j \Phi_{\kappa+1-m}^k\|_\infty = \|\mathrm{D}^j \Phi_{\kappa-(m-1)}^{k'}\|_\infty \lesssim \varepsilon^{-(\kappa + 1 - m + |j| - 1)} C^{(\kappa + 1 - m + |j|)}\,.$$

Consequently, it remains to show the validity of the bounds for $m = 0 \dots, \kappa - 1$ and $|j| = m$. To this end, we employ reverse induction on m.

Now, the base case is $m = \kappa - 1$ and thus $\kappa + 1 - m = 2$, $|j| = \kappa - 1$. The multivariate Leibniz rule yields

$$\begin{aligned}\mathrm{D}^j \Phi^k_{\kappa+1-m} = \mathrm{D}^j \Phi^k_2 &= \mathrm{D}^j \left(\frac{\partial}{\partial x_{k_2}} \frac{\partial}{\partial x_{k_1}} c \right) + \frac{1}{\varepsilon} \mathrm{D}^j \left(\left[\frac{\partial}{\partial x_{k_2}} \phi - \frac{\partial}{\partial x_{k_2}} c \right] \frac{\partial}{\partial x_{k_1}} c \right) \\ &= \mathrm{D}^{(j,k_2,k_1)} c + \frac{1}{\varepsilon} \sum_{i \subseteq j} \binom{j}{i} \left[\mathrm{D}^{(i,k_2)} \phi - \mathrm{D}^{(i,k_2)} c \right] \mathrm{D}^{(j-i,k_1)} c \,.\end{aligned}$$

As $\int_{\mathcal{Y}} \gamma_\varepsilon(x,y) \, \xi(\mathrm{d}y) = 1$, note that

$$\| \mathrm{D}^{(i,k_2)} \phi \|_\infty \leq \| \Phi^{(i,k_2)}_{|i|+1} \|_\infty \lesssim \varepsilon^{-(|i|+1-1)} C^{(|i|+1)}$$

by the first induction assumption as $|(i,k_2)| = |i| + 1 \leq |j| + 1 = \kappa$. Consequently,

$$\begin{aligned}|\mathrm{D}^j \Phi^k_{\kappa+1-m}| &\lesssim C_{|j|+2} + \varepsilon^{-1} \sum_{i \subseteq j} \binom{j}{i} \left[\varepsilon^{-|i|} C^{(|i|+1)} + C_{|i|+1} \right] C_{|j|-|i|+1} \\ &\lesssim C^{(\kappa+1)} + \varepsilon^{-1} \varepsilon^{-|j|} C^{(|j|+2)} \sum_{i \subseteq j} \binom{j}{i} \\ &\lesssim \varepsilon^{-(\kappa+1-m+|j|-1)} C^{(\kappa+1-m+|j|)} \,.\end{aligned}$$

Assume that we have the bounds for $\| \mathrm{D}^i \Phi^k_{\kappa+1-m} \|_\infty$ where $m \leq |i| \leq \kappa - 1$ and extend them to $m - 1$. For $|j| = m - 1$, we get similarly to before

$$\begin{aligned}&|\mathrm{D}^j \Phi^k_{\kappa+1-(m-1)}| = |\mathrm{D}^j \Phi^k_{\kappa+2-m}| \\ &= \left| \mathrm{D}^{(j,k_{\kappa+2-m})} \Phi^k_{\kappa+1-m} + \varepsilon^{-1} \sum_{i \subseteq j} \binom{j}{i} \left[\mathrm{D}^{(i,k_{\kappa+2-m})} \phi - \mathrm{D}^{(i,k_{\kappa+2-m})} c \right] \mathrm{D}^{j-i} \Phi^k_{\kappa+1-m} \right| \\ &\lesssim \varepsilon^{-(\kappa+1-m+|j|+1-1)} C^{(\kappa+1-m+|j|+1)} \\ &\qquad + \varepsilon^{-1} \sum_{i \subseteq j} \binom{j}{i} [\varepsilon^{-|i|} C^{(|i|+1)} + C_{|i|+1}] \varepsilon^{-(\kappa+1-m+|j|-|i|-1)} C^{(\kappa+1-m+|j|-|i|)} \\ &\lesssim \varepsilon^{-\kappa} C^{(\kappa+1)} + \varepsilon^{-(\kappa-1)} C^{(\kappa+1)} \sum_{i \subseteq j} \binom{j}{i} \\ &\lesssim \varepsilon^{-\kappa} C^{(\kappa+1)} = \varepsilon^{-(\kappa+1-(m-1)+|j|-1)} C^{(\kappa+1-(m-1)+|j|)} \,.\end{aligned}$$

■

Proof of Proposition 17. W.l.o.g. assume that $0 < \varepsilon \leq 1$ and let $\phi \in \mathcal{F}_{c,\varepsilon}$. Using Lemma 16, we get for $|k| = \alpha - 1$ that

$$\| \mathrm{D}^k \phi \|_\infty \lesssim \varepsilon^{-(\alpha-2)} C^{(\alpha-1)} \,.$$

Similarly, we have for $x_1, x_2 \in \mathring{\mathcal{X}}$ that

$$\| \mathrm{D}^k \phi(x_1) - \mathrm{D}^k \phi(x_2) \|_2 \leq \| \nabla \, \mathrm{D}^k \phi \|_2 \| x_1 - x_2 \|_2 \lesssim \varepsilon^{-(\alpha-1)} C^{(\alpha)} \| x_1 - x_2 \|_2 \,.$$

Since $\|x_1 - x_2\|_2 \leq \operatorname{diam}(\mathcal{X})$, this yields

$$\|\phi\|_\alpha \lesssim \varepsilon^{-(\alpha-2)} C^{(\alpha-1)} + \varepsilon^{-(\alpha-1)} C^{(\alpha)} \operatorname{diam}(\mathcal{X}) \lesssim \varepsilon^{-(\alpha-1)} C^{(\alpha)} .$$

Hence, we conclude that $\mathcal{F}_{c,\varepsilon} \subseteq \mathcal{C}^\alpha_M(\mathcal{X})$ for $M \coloneqq \varepsilon^{-(\alpha-1)} C^{(\alpha)} K$ with some $K > 0$ that only depends on $\mathcal{X}$ and α. Now, Lemma 43 yields

$$\begin{aligned} \log N(\delta, \mathcal{F}_{c,\varepsilon}, \|\bullet\|_\infty) &\leq \log N(\delta, \mathcal{C}^\alpha_M(\mathcal{X}), \|\bullet\|_\infty) \\ &\lesssim M^{s/\alpha} \delta^{-s/\alpha} \lesssim [C^{(\alpha)}]^{s/\alpha} \varepsilon^{-s(\alpha-1)/\alpha} \delta^{-s/\alpha} . \end{aligned}$$

■

Proof of Lemma 21. Write $g = g_i$ and $G_m = G_{i,m}$, $G^{(m)} = G^{(i,m)}$. We extend the bounds provided by Mena and Niles-Weed (2019, Proposition 1) to compositions $\phi \circ g$ with $\phi = \psi^{(c,\varepsilon,\xi)} \in \mathcal{F}_\sigma$ for some measure $\xi \in \mathrm{SG}_d(\sigma^2)$. Denoting $\bar\phi \coloneqq \phi \circ g - \frac{1}{2}\|\bullet\|_2^2 \circ g$, we want to bound for $k \in [\![s]\!]^\kappa$, $\kappa \leq \alpha$, $u \in \mathring{\mathcal{U}}$ the partial derivative

$$\mathrm{D}^k \bar\phi(u) = -\mathrm{D}^k \log(\exp[-\bar\phi(u)])$$

via the multivariate Faá di Bruno formula (Constantine and Savits, 1996). First, note that for all $x > 0$ it holds with some constant λ_κ that

$$\frac{\partial^\kappa}{\partial x} \log(x) = \lambda_k \frac{1}{x^\kappa}$$

Furthermore, we obtain similarly to Lemma 36 that

$$\mathrm{D}^k \exp(-\bar\phi(u)) = \int_{\mathcal{Y}} \Phi^k_\kappa(u,y) \exp(\psi(y) - \frac{1}{2}\|y\|_2^2 + \langle g(u), y\rangle)\, \xi(\mathrm{d}y) ,$$

where $\Phi^k_0(u,y) \coloneqq 1$ and for $m = 1, \ldots, \kappa$ we set

$$\Phi^k_m(u,y) \coloneqq \frac{\partial}{\partial u_{k_m}} \Phi^k_{m-1}(u,y) + \Phi^k_{m-1}(u,y) \sum_{q=1}^d \frac{\partial}{\partial u_{k_m}} g(u)_q y_q . \tag{22}$$

Hence, upon defining

$$\Psi_k(u) \coloneqq \frac{\int_{\mathcal{Y}} \Phi^k_\kappa(u,y) \exp(\psi(y) - \frac{1}{2}\|y\|_2^2 + \langle g(u), y\rangle)\, \xi(\mathrm{d}y)}{\int_{\mathcal{Y}} \exp(\psi(y) - \frac{1}{2}\|y\|_2^2 + \langle g(u), y\rangle)\, \xi(\mathrm{d}y)} ,$$

it follows by the multivariate Faá di Bruno formula that

$$\mathrm{D}^k \bar\phi(u) = -\mathrm{D}^k \log(\exp[-\bar\phi(u)]) = \sum_{j_1, \ldots j_\kappa} \lambda_{\kappa, j_1, \ldots, j_\kappa} \prod_{i=1}^\kappa \Psi_{j_i}(u) , \tag{23}$$

where the sum runs over all sub-indices $j_1, \ldots, j_\kappa$ that partition k, i.e., $(j_1, \ldots, j_\kappa)$ is equal to a permuted version of k, and the $\lambda_{\kappa, j_1, \ldots, j_\kappa}$ are some constants related to the derivatives of the logarithm. We show that

$$|\Psi_k(u)| \lesssim G^{(\kappa)} \sum_{r=1}^\kappa \sum_{q_1, \ldots, q_r = 1}^d \frac{\int_{\mathcal{Y}} \prod_{t=1}^r |y_{q_t}| \exp(\psi(y) - \frac{1}{2}\|y\|_2^2 + \langle g(u), y\rangle)\, \xi(\mathrm{d}y)}{\int_{\mathcal{Y}} \exp(\psi(y) - \frac{1}{2}\|y\|_2^2 + \langle g(u), y\rangle)\, \xi(\mathrm{d}y)} ,$$

where the implicit constant only depends on κ. Together with (23) and using the bound by Mena and Niles-Weed (2019, Lemma 3 in Appendix B) for each summand, we obtain the desired result.

It suffices to show that for $k \in [\![s]\!]^\kappa$ it holds with $m = 0, \dots, \kappa - 1$ and indices tuples $|j| = 0, \dots, m$ that

$$|\mathrm{D}^j \Phi^k_{\kappa-m}(u,y)| \lesssim G^{(\kappa-m+|j|)} \sum_{r=1}^{\kappa-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \,,$$

where the implicit constant only depends on κ. Then, with $m = 0 = |j|$ we can conclude.

We prove this similar to Lemma 16 by double induction. In the following, we write $\Phi^k_m \equiv \Phi^k_m(u,y)$. First, we do induction over $\kappa = |k|$.

For the base case $\kappa = 1$ and thus $m = 0 = |j|$, it holds due to (22) that

$$|\mathrm{D}^j \Phi^k_{\kappa-m}| = |\Phi^k_1| = \Big| \sum_{q=1}^{d} \frac{\partial}{\partial u_{k_1}} g(u)_q y_q \Big| \lesssim G^{(1)} \sum_{q=1}^{d} |y_q| = G^{(\kappa-m)} \sum_{r=1}^{\kappa-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \,.$$

Let the bound on $|\mathrm{D}^j \Phi^k_{\kappa-m}|$ hold for all $|k| = \kappa$, $m = 0, \dots, \kappa - 1$ and $|j| = 0, \dots, m$. We extend this to $\kappa + 1$. Again, we only have to bound the new diagonal $m = 0, \dots, \kappa$ and $|j| = m$. We do this by reverse induction on m.

The base case $m = \kappa = |j|$ holds as

$$\begin{aligned} |\mathrm{D}^j \Phi^k_{\kappa+1-m}| = |\mathrm{D}^j \Phi^k_1| &= \Big| \sum_{q=1}^{d} \mathrm{D}^{(j,k_1)} g(u)_q y_q \Big| \\ &\lesssim G^{(|j|+1)} \sum_{q=1}^{d} |y_q| = G^{(\kappa+1-m)} \sum_{r=1}^{\kappa+1-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \,. \end{aligned}$$

Suppose that $|\mathrm{D}^i \Phi^k_{\kappa+1-m}|$ is bounded as required for $m \le |i| \le \kappa$. Then, we need to extend this to $m - 1$. Using (22) and the multivariate Leibniz rule yields for $|j| = m - 1$ that

$$\begin{aligned} &|\mathrm{D}^j \Phi^k_{\kappa+1-(m-1)}| \\ &\quad = \Big| \mathrm{D}^{(j,k_{\kappa+2-m})} \Phi^k_{\kappa+1-m} + \sum_{q=1}^{d} \mathrm{D}^j \left[\Phi^k_{\kappa+1-m} \frac{\partial}{\partial u_{k_{\kappa+2-m}}} g(u)_q y_q \right] \Big| \\ &\quad = \Big| \mathrm{D}^{(j,k_{\kappa+2-m})} \Phi^k_{\kappa+1-m} + \sum_{q=1}^{d} y_q \sum_{i \subseteq j} \binom{j}{i} \mathrm{D}^{(i,k_{\kappa+2-m})} [g(u)_q] \, \mathrm{D}^{j-i} \Phi^k_{\kappa+1-m} \Big| \\ &\quad \lesssim G^{(\kappa+1-m+|j|+1)} \sum_{r=1}^{\kappa+1-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \\ &\qquad\qquad + \sum_{q=1}^{d} |y_q| \sum_{i \subseteq j} \binom{j}{i} G^{(|i|+1)} \left[G^{(\kappa+1-m+|j|-|i|)} \sum_{r=1}^{\kappa+1-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \right] \end{aligned}$$

$$\lesssim G^{(\kappa+1-m+|j|+1)} \sum_{r=1}^{\kappa+1-m} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}|$$
$$+ G^{(\kappa+1-m+|j|+1)} \sum_{r=1}^{\kappa+1-m+1} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \sum_{i \subseteq j} \binom{j}{i}$$
$$\lesssim G^{(\kappa+1-(m-1)+|j|)} \sum_{r=1}^{\kappa+1-(m-1)} \sum_{q_1,\dots,q_r=1}^{d} \prod_{t=1}^{r} |y_{q_t}| \,. \qquad \blacksquare$$

Lemma 37. *Let $\sigma > 0$ and $\mu \in \mathrm{SG}_d(\sigma^2)$, $\nu \in \mathrm{SG}_d(\sigma^2)$. Let σ_n be the infimum over all $\tau > 0$ such that μ, $\hat{\mu}_n$, ν, $\hat{\nu}_n \in \mathrm{SG}_d(\tau^2)$. Then, it holds that*

$$\mathbb{P}(\sigma_n^2 > 6\sigma^2) \leq 4n^{-1} \,,$$
$$\mathbb{E}[\sigma_n^{2k}] \leq 2k^k \sigma^{2k} \quad (k \in \mathbb{N}) \,.$$

Proof The second assertion follows from Mena and Niles-Weed (2019, Lemma B.4 in supplement). For the first assertion, denote

$$\tau_{1,n} \coloneqq \mathbb{E}_{X \sim \hat{\mu}_n}[\exp_{4d\sigma^2}(\|X\|_2^2)] \,, \qquad \tau_1 \coloneqq \mathbb{E}_{X \sim \mu}[\exp_{4d\sigma^2}(\|X\|_2^2)] \,,$$
$$\tau_{2,n} \coloneqq \mathbb{E}_{Y \sim \hat{\nu}_n}[\exp_{4d\sigma^2}(\|Y\|_2^2)] \,, \qquad \tau_2 \coloneqq \mathbb{E}_{Y \sim \nu}[\exp_{4d\sigma^2}(\|Y\|_2^2)] \,.$$

As in the proof of Lemma B.4 in the supplement of Mena and Niles-Weed (2019), it follows that $\sigma_n^2 \leq 2\sigma^2 \max(\tau_{1,n}, \tau_{2,n})$. Hence, it holds that

$$\mathbb{P}(\sigma_n^2 > 6\sigma^2) \leq \mathbb{P}(\max(\tau_{1,n}, \tau_{2,n}) > 3) \leq \mathbb{P}(\tau_{1,n} > 3) + \mathbb{P}(\tau_{2,n} > 3) \,.$$

We bound the first term and conclude, the second term can be dealt with analogously. Using the Chebyshev inequality, we get

$$\mathbb{P}(\tau_{1,n} > 3) \leq \mathbb{P}(|\tau_{1,n} - \tau_1| > 3 - \tau_1) \leq \frac{\mathrm{Var}_{X \sim \mu}[\exp_{4d\sigma^2}(\|X\|_2^2)]}{(3 - \tau_1)^2} n^{-1} \leq 2n^{-1} \,,$$

where we used that by sub-Gaussianity $\tau_1 \leq 2$ as well as

$$\mathrm{Var}_{X \sim \mu}[\exp_{4d\sigma^2}(\|X\|_2^2)] \leq \mathbb{E}_{X \sim \mu}[\exp_{2d\sigma^2}(\|X\|_2^2)] \leq 2 \,. \qquad \blacksquare$$

Proof of Theorem 23. First, recall that according to Lemma 39 we can consider the case that $I = 1$. Given empirical probability measures $\hat{\mu}_n, \hat{\nu}_n$, we have due to Lemma 37 a random σ_n^2 such that $\mu, \hat{\mu}_n, \nu, \hat{\nu}_n$ are all in $\mathrm{SG}_d(\sigma_n^2)$. Similar to Corollary 2 in Mena and Niles-Weed (2019) (and Lemma 4), we get

$$\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{T}_{c,\varepsilon}(\mu, \nu)|] \leq 2\,\mathbb{E}\left[\sup_{\phi \in \mathcal{F}_{\sigma_n}} \left| \int_{\mathcal{X}} \phi \,\mathrm{d}[\mu - \hat{\mu}_n] \right| \right] \tag{24}$$
$$+ 2\,\mathbb{E}\left[\sup_{\phi \in \mathcal{F}_{\sigma_n}} \left| \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\hat{\mu}_n)} \,\mathrm{d}[\nu - \hat{\nu}_n] \right| \right] .$$

We now bound the two terms separately and conclude. We decompose the first term of (24) into

$$\begin{aligned}\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\phi\,\mathrm{d}[\mu-\hat{\mu}_n]\right|\right] &\le \mathbb{E}\left[\sigma_n^{4\alpha}\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\sigma_n^{-4\alpha}[\phi-\frac{1}{2}\|\bullet\|_2^2]\,\mathrm{d}[\mu-\hat{\mu}_n]\right|\right]\\ &\quad+\mathbb{E}\left[\left|\int_{\mathcal{X}}\frac{1}{2}\|\bullet\|_2^2\,\mathrm{d}[\mu-\hat{\mu}_n]\right|\right]\\ &\lesssim (\mathbb{E}[\sigma_n^{8\alpha}])^{1/2}\left(\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\sigma_n^{-4\alpha}[\phi-\frac{1}{2}\|\bullet\|_2^2]\,\mathrm{d}[\mu-\hat{\mu}_n]\right|^2\right]\right)^{1/2}\\ &\quad+r^2n^{-1/2}\,,\end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality. The expectation of $\sigma_n^{8\alpha}$ can be bounded via Lemma 37 by an explicit constant that depends on α times $\sigma^{8\alpha}$. Further, by definition the function class $\mathcal{F}_{\sigma_n}$ is bounded by $6d^2r^2\sigma_n^4$, and hence $[\mathcal{F}_{\sigma_n}-\frac{1}{2}\|\bullet\|_2^2]$ is bounded by $6d^2r^2\sigma_n^4+\frac{1}{2}r^2\le 8d^2r^2\sigma_n^4$. In conjunction with Lemma 21 we thus infer that

$$\sigma_n^{-4\alpha}[\mathcal{F}_{\sigma_n}-\frac{1}{2}\|\bullet\|_2^2]\subseteq\mathcal{C}_M^\alpha(\mathcal{U})\text{ with }M\coloneqq[G^{(\alpha)}]^{s/\alpha}+8d^2r^2\,,$$

where the first term in M controls the derivatives of functions in $\sigma_n^{-4\alpha}[\mathcal{F}_{\sigma_n}-\frac{1}{2}\|\bullet\|_2^2]$ and the second term arises from our aforementioned upper bound. Note that the function class on the right-hand side is deterministic and independent of σ_n^2. Hence, we can apply Theorem 2.14.5 from van der Vaart and Wellner (1996), Theorem 3.5.1 from Giné and Nickl (2015) and Lemma 43 to obtain

$$\begin{aligned}&n\,\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\sigma_n^{-4\alpha}[\phi-\frac{1}{2}\|\bullet\|_2^2]\,\mathrm{d}[\mu-\hat{\mu}_n]\right|^2\right]\\ &\quad\lesssim\left(\sqrt{n}\,\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\sigma_n^{-4\alpha}[\phi-\frac{1}{2}\|\bullet\|_2^2]\,\mathrm{d}[\mu-\hat{\mu}_n]\right|\right]+M\right)^2\\ &\quad\lesssim\left(\mathbb{E}\left[\int_0^M\sqrt{\log 2N(\delta,\,\mathcal{C}_M^\alpha(\mathcal{U}),\,\|\bullet\|_\infty)}\,\mathrm{d}\delta\right]+M\right)^2\\ &\quad\lesssim\left(\int_0^M\sqrt{1+M^{s/\alpha}\delta^{-s/\alpha}}\,\mathrm{d}\delta+M\right)^2\\ &\quad\lesssim(M+M^{s/(2\alpha)}M^{1-s/(2\alpha)}+M)^2\lesssim M^2\,,\end{aligned}$$

where we used that $s/\alpha<2$ and the implicit constant only depends on α, d, s and $\mathcal{U}$. Combining these inequalities we obtain

$$\sqrt{n}\,\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{X}}\phi\,\mathrm{d}[\mu-\hat{\mu}_n]\right|\right]\lesssim([G^{(\alpha)}]^{s/\alpha}+d^2r^2)\sigma^{4\alpha}+r^2\,.$$

For the second term of (24), we decompose into the parts $\sigma_n^2 > 6\sigma^2$ and $\sigma_n^2 \leq 6\sigma^2$ and use the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{Y}}\phi^{(c,\varepsilon,\hat{\mu}_n)}\,\mathrm{d}[\nu-\hat{\nu}_n]\right|\right] \\
&\quad= \mathbb{E}\left[[\mathbb{I}(\sigma_n^2>6\sigma^2)+\mathbb{I}(\sigma_n^2\leq 6\sigma^2)]\sup_{\phi\in\mathcal{F}_{\sigma_n}}\left|\int_{\mathcal{Y}}\phi^{(c,\varepsilon,\hat{\mu}_n)}\,\mathrm{d}[\nu-\hat{\nu}_n]\right|\right] \\
&\quad\leq (\mathbb{P}[\sigma_n^2>6\sigma^2])^{1/2}\left(\mathbb{E}\left[\left(\sup_{\phi\in\mathcal{F}_{\sigma_n}}\int_{\mathcal{Y}}|\phi^{(c,\varepsilon,\hat{\mu}_n)}|\,\mathrm{d}\nu+\sup_{\phi\in\mathcal{F}_{\sigma_n}}\int_{\mathcal{Y}}|\phi^{(c,\varepsilon,\hat{\mu}_n)}|\,\mathrm{d}\hat{\nu}_n\right)^2\right]\right)^{1/2} \\
&\qquad+\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sqrt{6}\sigma}}\left|\int_{\mathcal{Y}}\phi^{(c,\varepsilon,\hat{\mu}_n)}\,\mathrm{d}[\nu-\hat{\nu}_n]\right|\right].
\end{aligned}$$

The probability can be bounded via Lemma 37. Furthermore, employing that $\mathcal{F}_\sigma$ is bounded in uniform norm by $6d^2r^2\sigma^4$, the definition of the entropic transform and that for $x\in\mathcal{X}$, $y\in\mathcal{Y}$: $\|x-y\|^2\leq r^2+2r\|y\|_2+\|y\|_2^2\leq 4r^2+4r\|y\|_2^2$, it follows for $\phi\in\mathcal{F}_\sigma$ that

$$|\phi^{(c,\varepsilon,\hat{\mu}_n)}(y)|\leq 8d^2r^2\sigma^4+2r\|y\|_2^2\,. \tag{25}$$

Hence, it holds that

$$\sup_{\phi\in\mathcal{F}_{\sigma_n}}\int_{\mathcal{Y}}|\phi^{(c,\varepsilon,\hat{\mu}_n)}|\,\mathrm{d}\nu\leq 8d^2r^2\sigma_n^4+2r\,\mathbb{E}_{Y\sim\nu}[\|Y\|_2^2]$$

and similar for $\hat{\nu}_n$. Thus,

$$\begin{aligned}
&\left(\mathbb{E}\left[\left(\sup_{\phi\in\mathcal{F}_{\sigma_n}}\int_{\mathcal{Y}}|\phi^{(c,\varepsilon,\hat{\mu}_n)}|\,\mathrm{d}\nu+\sup_{\phi\in\mathcal{F}_{\sigma_n}}\int_{\mathcal{Y}}|\phi^{(c,\varepsilon,\hat{\mu}_n)}|\,\mathrm{d}\hat{\nu}_n\right)^2\right]\right)^{1/2} \\
&\quad\leq (\mathbb{E}[(16d^2r^2\sigma_n^4+2r\,\mathbb{E}_{Y\sim\nu}[\|Y\|_2^2]+2r\,\mathbb{E}_{Y\sim\hat{\nu}_n}[\|Y\|_2^2])^2])^{1/2} \\
&\quad\lesssim d^2r^2\sigma^4\,,
\end{aligned} \tag{26}$$

where we used the Cauchy-Schwarz inequality and the moment bounds provided in Lemma 37 and Lemma B.1 in the supplement of Mena and Niles-Weed (2019). Recalling (25), we have

$$\begin{aligned}
b^2 \coloneqq \sup_{\phi\in\mathcal{F}_{\sqrt{6}\sigma}}\|\phi^{(c,\varepsilon,\hat{\mu}_n)}\|_{L^2(\hat{\nu}_n)}^2 &\leq \mathbb{E}_{Y\sim\hat{\nu}_n}[(8d^2r^2(\sqrt{6}\sigma)^4+2r\|Y\|_2^2)^2] \\
&= \mathbb{E}_{Y\sim\hat{\nu}_n}[(288d^2r^2\sigma^4+2r\|Y\|_2^2)^2].
\end{aligned}$$

Employing Theorem 3.5.1 from Giné and Nickl (2015) in combination with Lemma 5 and Proposition 22, we obtain

$$\begin{aligned}\sqrt{n}\,\mathbb{E}\left[\sup_{\phi\in\mathcal{F}_{\sqrt{6}\sigma}}\left|\int_{\mathcal{Y}}\phi^{(c,\varepsilon,\hat{\mu}_n)}\,\mathrm{d}[\nu-\hat{\nu}_n]\right|\right] &\lesssim \mathbb{E}\left[\int_0^b\sqrt{\log 2N(\delta/2,\,\mathcal{F}_{\sqrt{6}\sigma},\,\|\bullet\|_\infty)}\,\mathrm{d}\delta\right]\\ &\lesssim \mathbb{E}[b+[G^{(\alpha)}]^{s/(2\alpha)}\sigma^{3s/2}b^{1-s/(2\alpha)}]\\ &\leq (1+[G^{(\alpha)}]^{s/(2\alpha)}\sigma^{3s/2})\,\mathbb{E}[1+b]\\ &\leq (1+[G^{(\alpha)}]^{s/(2\alpha)}\sigma^{3s/2})\left(1+\mathbb{E}[b^2]^{1/2}\right)\\ &\lesssim (1+[G^{(\alpha)}]^{s/(2\alpha)}\sigma^{3s/2})d^2r^2\sigma^4,\end{aligned}$$

where for the last inequality we used Lemma B.1 in the supplement of Mena and Niles-Weed (2019), Lemma 37 and $d, r, \sigma \geq 1$ to upper bound $\mathbb{E}[b^2]^{1/2}$, and the implicit constant only depends on α, d, s, and $\mathcal{U}$. Putting everything together, we obtain that

$$\begin{aligned}\sqrt{n}\,\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{T}_{c,\varepsilon}(\mu,\nu)|] &\lesssim ([G^{(\alpha)}]^{s/\alpha}+r^2d^2)\sigma^{4\alpha}+r^2+d^2r^2\sigma^4\\ &\quad+(1+[G^{(\alpha)}]^{s/(2\alpha)}\sigma^{3s/2})d^2r^2\sigma^4\\ &\lesssim (1+[G^{(\alpha)}]^{s/\alpha})r^2\sigma^{4\alpha\vee(4+3s/2)}\,,\end{aligned}$$

where the implicit constant only depends on α, d, s, and $\mathcal{U}$. ∎

Proof of Corollary 24. Consider the case $\widehat{\mathrm{T}}_{c,\varepsilon,n}=\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)$, the one-sample plug-in estimators can be dealt with analogously. Furthermore, w.l.o.g. we can assume that $0<\varepsilon<1$. By Remark 45, it holds that

$$\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{T}_{c,\varepsilon}(\mu,\nu)|]=\varepsilon\,\mathbb{E}[|\mathrm{T}_{c,1}(\hat{\mu}_n^\varepsilon,\hat{\nu}_n^\varepsilon)-\mathrm{T}_{c,1}(\mu^\varepsilon,\nu^\varepsilon)|]\,,$$

where μ^ε is supported on $\varepsilon^{-1/2}\mathcal{X}=\bigcup_{i=1}^I\varepsilon^{-1/2}g_i(\mathcal{U}_i)$ with $\sup_{x\in\varepsilon^{-1/2}\mathcal{X}}\|x\|_2\leq\varepsilon^{-1/2}r$ and μ^ε, ν^ε are σ^2/ε-sub-Gaussian. Hence, it follows from Theorem 23 that

$$\begin{aligned}&\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{T}_{c,\varepsilon}(\mu,\nu)|]\\ &\qquad\lesssim\varepsilon\left(\sum_{i=1}^I 1+[G_\varepsilon^{(i,\alpha)}]^{s/\alpha}\right)[\varepsilon^{-1/2}r]^2[\varepsilon^{-1/2}\sigma]^{4\alpha\vee(4+3s/2)}n^{-1/2}\,,\end{aligned}$$

where $G_\varepsilon^{(i,\alpha)}$ are the constants from Lemma 21 where g_i is substituted with $\varepsilon^{-1/2}g_i$. As $\varepsilon<1$, it follows from the definition that $G_\varepsilon^{(i,\alpha)}\leq\varepsilon^{-\alpha/2}G^{(i,\alpha)}$. As a consequence,

$$\begin{aligned}&\mathbb{E}[|\mathrm{T}_{c,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{T}_{c,\varepsilon}(\mu,\nu)|]\\ &\qquad\lesssim\left(\sum_{i=1}^I 1+[G^{(i,\alpha)}]^{s/\alpha}\right)r^2\sigma^{4\alpha\vee(4+3s/2)}\varepsilon^{-[2\alpha\vee(2+3s/4)]-s/2}n^{-1/2}\,.\end{aligned}$$

∎

A.3 Computational Complexity

For our computational analysis of a computable estimator for the empirical EOT cost we make use of the following characterization of dual potentials which arise from Sinkhorn iterations. Its proof is based on an insight which was previously formulated for cost chosen as the squared Euclidean norm by Pooladian and Niles-Weed (2021, Proof of Theorem 6).

Lemma 38 (Potentials from Sinkhorn algorithm)**.** *Let Assumption* **(C)** *hold and consider probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. For $\psi_0 \in L_\varepsilon^{\exp}(\nu)$ define its single and double (c,ε)-transform as $\phi \coloneqq \psi_0^{(c,\varepsilon,\nu)}$ and $\psi \coloneqq \phi^{(c,\varepsilon,\mu)}$ Further, define the measure $\tilde{\nu}$ on $\mathcal{Y}$ by*

$$\frac{\mathrm{d}\tilde{\nu}}{\mathrm{d}\nu}(y) \coloneqq \int_{\mathcal{X}} \exp_\varepsilon\left(\phi(x) + \psi_0(y) - c(x,y)\right) \mu(\mathrm{d}x)\,.$$

Then, the measure $\tilde{\nu}$ is a probability measure on $\mathcal{Y}$. Further, the potentials (ϕ,ψ) are the dual optimizers of (10) for μ and $\tilde{\nu}$, and it holds that

$$\mathrm{T}_{c,\varepsilon}(\mu,\tilde{\nu}) = \int_{\mathcal{X}} \phi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \psi \,\mathrm{d}\tilde{\nu} = \max_{\phi \in \mathcal{F}_{c,\varepsilon}} \int_{\mathcal{X}} \phi \,\mathrm{d}\mu + \int_{\mathcal{Y}} \phi^{(c,\varepsilon,\mu)} \,\mathrm{d}\tilde{\nu}\,.$$

Proof For the first assertion note that $\tilde{\nu}$ has a non-negative density with respect to ν, hence it suffices to show that $\tilde{\nu}$ integrates to one. Invoking the Tonelli-Fubini theorem it follows by definition of the (c,ε)-transform of ψ_0 that

$$\begin{aligned}
\tilde{\nu}(\mathcal{Y}) &= \int_{\mathcal{Y}} \int_{\mathcal{X}} \exp_\varepsilon\left(\phi(x) + \psi_0(y) - c(x,y)\right) \mu(\mathrm{d}x)\, \nu(\mathrm{d}y) \\
&= \int_{\mathcal{X}} \exp_\varepsilon(\phi(x)) \left[\int_{\mathcal{Y}} \exp_\varepsilon\left(\psi_0(y) - c(x,y)\right) \nu(\mathrm{d}y)\right] \mu(\mathrm{d}x) \\
&= \int_{\mathcal{X}} \exp_\varepsilon(\phi(x) - \phi(x))\, \mu(\mathrm{d}x) = 1.
\end{aligned}$$

For the second assertion we employ the characterization for optimality of EOT plans and potentials from Theorem 1. To this end, we define the measure π on $\mathcal{X} \times \mathcal{Y}$,

$$\mathrm{d}\pi \coloneqq \exp_\varepsilon(\phi \oplus \psi - c)\, \mathrm{d}[\mu \otimes \tilde{\nu}]$$

and verify that $\pi \in \Pi(\mu,\tilde{\nu})$. Once this is confirmed, the remaining assertions follow at once from Theorem 1 and Proposition 3. To show the marginal constraints we first observe for any $y \in \mathcal{Y}$ that the marginal density of π on $\mathcal{Y}$ fulfills

$$\begin{aligned}
&\int_{\mathcal{X}} \exp_\varepsilon\left(\phi(x) + \psi(y) - c(x,y)\right) \mu(\mathrm{d}x) \\
&= \exp_\varepsilon\left(\phi^{(c,\varepsilon,\mu)}(y)\right) \int_{\mathcal{X}} \exp_\varepsilon\left(\phi(x) - c(x,y)\right) \mu(\mathrm{d}x) \\
&= \exp_\varepsilon\left(\phi^{(c,\varepsilon,\mu)}(y) - \phi^{(c,\varepsilon,\mu)}(y)\right) = 1.
\end{aligned}$$

For the marginal density of π on $\mathcal{X}$ we note by definition of ϕ and $\tilde{\nu}$ for any $x \in \mathcal{X}$ that

$$\begin{aligned}
&\int_{\mathcal{Y}} \exp_\varepsilon\left(\phi(x) + \psi(y) - c(x,y)\right) \tilde{\nu}(\mathrm{d}y) \\
&= \int_{\mathcal{Y}} \frac{\exp_\varepsilon\left(\phi(x) - c(x,y)\right)}{\int_{\mathcal{X}} \exp_\varepsilon\left(\phi(\tilde{x}) - c(\tilde{x},y)\right) \mu(\mathrm{d}\tilde{x})} \tilde{\nu}(\mathrm{d}y) \\
&= \int_{\mathcal{Y}} \frac{\exp_\varepsilon\left(\phi(x) + \psi_0(y) - c(x,y)\right)}{\int_{\mathcal{X}} \exp_\varepsilon\left(\phi(\tilde{x}) + \psi_0(y) - c(\tilde{x},y)\right) \mu(\mathrm{d}\tilde{x})} \tilde{\nu}(\mathrm{d}y) \\
&= \int_{\mathcal{Y}} \exp_\varepsilon\left(\phi(x) + \psi_0(y) - c(x,y)\right) \nu(\mathrm{d}y) \\
&= \exp_\varepsilon\left(\psi_0^{(c,\varepsilon,\nu)}(x)\right) \int_{\mathcal{Y}} \exp_\varepsilon\left(\psi_0(y) - c(x,y)\right) \nu(\mathrm{d}y) \\
&= \exp_\varepsilon\left(\psi_0^{(c,\varepsilon,\nu)}(x) - \psi_0^{(c,\varepsilon,\nu)}(x)\right) = 1.
\end{aligned}$$

■

A.4 Implications to the Entropic Gromov-Wasserstein Distance

Proof of Lemma 30. For $A \in \mathcal{D}$, denote

$$U_\varepsilon^{\mu,\nu}(A) \coloneqq 32\|A\|_2^2 + \mathrm{T}_{c_A,\varepsilon}(\mu,\nu)\,,$$

such that by Theorem 29 it holds

$$\mathrm{GW}_{2,\varepsilon}(\mu,\nu) = \min_{A \in \mathcal{D}} U_\varepsilon^{\mu,\nu}(A)\,.$$

In particular, there exist $A, \tilde{A} \in \mathcal{D}$ such that

$$\mathrm{GW}_{2,\varepsilon}(\mu,\nu) = U_\varepsilon^{\mu,\nu}(A)\,, \quad \mathrm{GW}_2(\tilde{\mu},\tilde{\nu}) = U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(\tilde{A})\,.$$

By optimality, it follows that

$$U_\varepsilon^{\mu,\nu}(A) - U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(A) \le U_\varepsilon^{\mu,\nu}(A) - U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(\tilde{A}) \le U_\varepsilon^{\mu,\nu}(\tilde{A}) - U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(\tilde{A})\,.$$

Hence,

$$\begin{aligned}
|\mathrm{GW}_{2,\varepsilon}(\mu,\nu) - \mathrm{GW}_{2,\varepsilon}(\tilde{\mu},\tilde{\nu})| &\le |U_\varepsilon^{\mu,\nu}(A) - U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(A)| + |U_\varepsilon^{\mu,\nu}(\tilde{A}) - U_\varepsilon^{\tilde{\mu},\tilde{\nu}}(\tilde{A})| \\
&= |\mathrm{T}_{c_A,\varepsilon}(\mu,\nu) - \mathrm{T}_{c_A,\varepsilon}(\tilde{\mu},\tilde{\nu})| + |\mathrm{T}_{c_{\tilde{A}},\varepsilon}(\mu,\nu) - \mathrm{T}_{c_{\tilde{A}},\varepsilon}(\tilde{\mu},\tilde{\nu})| \\
&\le 2 \sup_{A \in \mathcal{D}} |\mathrm{T}_{c_A,\varepsilon}(\mu,\nu) - \mathrm{T}_{c_A,\varepsilon}(\tilde{\mu},\tilde{\nu})|\,.
\end{aligned}$$

An application of Lemma 4 yields the second inequality. ■

Proof of Lemma 31. First note by the assumption on $\mathcal{F}$ that elements of $\mathcal{F}^{(\mathcal{D},\varepsilon,\tilde{\mu})}$ are well-defined and real-valued. W.l.o.g. we can assume that covering numbers on the right-hand side of (20) are finite since otherwise the bound is vacuous. As in the proof of Lemma 5, we construct a $\delta/2$-covering $\{\phi_1, \dots, \phi_N\} \subseteq \mathcal{F}$ of $\mathcal{F}$ with $N \coloneqq N(\delta/4, \mathcal{F}, \|\bullet\|_\infty)$. Furthermore,

let $\{A_1, \dots, A_M\} \subseteq \mathcal{D}$ with $M \coloneqq N(\delta/[64r^2], \mathcal{D}, \|\bullet\|_\infty)$ be a $\delta/[64r^2]$-covering of $\mathcal{D}$. We show that

$$\left\{ \phi_i^{(c_{A_j},\varepsilon,\tilde{\mu})} \;\middle|\; i \in \{1,\dots,N\},\, j \in \{1,\dots,M\} \right\}$$

is a δ-covering of $\mathcal{F}^{(\mathcal{D},\varepsilon,\tilde{\mu})}$ and conclude. For $\psi \in \mathcal{F}^{(\mathcal{D},\varepsilon,\tilde{\mu})}$, by definition there is a $\phi \in \mathcal{F}$ and $A \in \mathcal{D}$ such that $\psi = \phi^{(c_A,\varepsilon,\tilde{\mu})}$. In particular, there exists ϕ_i and A_j with $\|\phi - \phi_i\|_\infty \leq \delta/2$ and $\|A - A_j\|_\infty \leq \delta/[64r^2]$. The triangle inequality yields that

$$\begin{aligned}
\|\phi^{(c_A,\varepsilon,\tilde{\mu})} - \phi_i^{(c_{A_j},\varepsilon,\tilde{\mu})}\|_\infty &\leq \|\phi^{(c_A,\varepsilon,\tilde{\mu})} - \phi_i^{(c_A,\varepsilon,\tilde{\mu})}\|_\infty + \|\phi_i^{(c_A,\varepsilon,\tilde{\mu})} - \phi_i^{(c_{A_j},\varepsilon,\tilde{\mu})}\|_\infty \\
&\leq \|\phi - \phi_i\|_\infty + \|c_A - c_{A_j}\|_\infty \\
&\leq \|\phi - \phi_i\|_\infty + 32r^2\|A - A_j\|_\infty \leq \delta/2 + \delta/2 = \delta\,. \qquad \blacksquare
\end{aligned}$$

Proof of Theorem 32. Consider the two-sample estimator $\widehat{\mathrm{GW}}_{\varepsilon,n} = \mathrm{GW}_\varepsilon(\hat{\mu}_n, \hat{\nu}_n)$ and note that the one-sample estimators from (18) can be handled similarly. First, note that by the proof of Theorem 2 from Zhang et al. (2022) and Assumption **(GW)**, it holds that

$$\mathbb{E}[|\widehat{\mathrm{GW}}_{\varepsilon,n} - \mathrm{GW}_\varepsilon(\mu,\nu)|] \lesssim r^4 n^{-1/2} + \mathbb{E}[|\mathrm{GW}_{2,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GW}_{2,\varepsilon}(\mu,\nu)|]\,.$$

Hence, it remains to bound the second term involving $\mathrm{GW}_{2,\varepsilon}$. Following the proof of Theorem 6 with Lemma 30 and Lemma 31, we see that if there exist constants K_ε, $k > 0$ such that for $\delta > 0$ suffices small it holds

$$\log N(\delta/4,\, \mathcal{F}_{\mathcal{D},\varepsilon},\, \|\bullet\|_\infty) + \log N(\delta/[64r^2],\, \mathcal{D},\, \|\bullet\|_\infty) \leq K_\varepsilon \delta^{-k}\,,$$

then

$$\mathbb{E}[|\mathrm{GW}_{2,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GW}_{2,\varepsilon}(\mu,\nu)|] \lesssim \sqrt{1 + K_\varepsilon}\, n^{-1/2}\,.$$

Since for any $\delta > 0$ it holds,

$$\log N(\delta,\, \mathcal{D},\, \|\bullet\|_\infty) \lesssim sdr^2\delta^{-1}\,,$$

it remains to show that the uniform covering numbers of $\mathcal{F}_{\mathcal{D},\varepsilon}$ are suitably bounded. Furthermore, we need to rescale $\{c_A\}_{A\in\mathcal{D}}$ by a constant such that Assumption **(C)** is met for each element. To this end, note by Theorem 29 for $a > 0$ that

$$\frac{1}{a}\,\mathrm{GW}_{2,\varepsilon}(\mu,\nu) = \min_{A\in\mathcal{D}} \frac{32}{a}\|A\|_2^2 + \mathrm{T}_{c_A/a,\varepsilon/a}(\mu,\nu)\,,$$

which only depends on the scaled cost functions c_A/a. Furthermore, we have uniformly over all $A \in \mathcal{D}$ that

$$\|\mathrm{D}^k c_A\|_\infty \leq \begin{cases} 20r^4 & \text{if } |k| = 0\,, \\ (8 + 16\sqrt{d})r^3 & \text{if } |k| = 1\,, \\ 8r^2 & \text{if } |k| = 2\,, \\ 0 & \text{if } |k| > 2\,, \end{cases} \tag{27}$$

and can therefore set $a \coloneqq 20r^4$. In particular, we see that the functions classes $\mathcal{F}_{c_A,\varepsilon}$ are α-Hölder smooth for any $\alpha \in \mathbb{N}$ with uniform Hölder constant over all $A \in \mathcal{D}$. Hence, as

in the proof of Proposition 17 we obtain the desired upper bound on the uniform covering numbers of $\mathcal{F}_{\mathcal{D},\varepsilon}$. Putting everything together, via the rescaling

$$\mathbb{E}[|\mathrm{GW}_{2,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{GW}_{2,\varepsilon}(\mu, \nu)|] = a\,\mathbb{E}[|a^{-1}\,\mathrm{GW}_{2,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - a^{-1}\,\mathrm{GW}_{2,\varepsilon}(\mu, \nu)|]\,,$$

and Remark 46, the assertion follows. ∎

Proof of Remark 34. We follow the proof of Theorem 32 and consider $\widehat{\mathrm{GW}}_{0,n} = \mathrm{GW}_0(\hat{\mu}_n, \hat{\nu}_n)$. First, note that since the decomposition $\mathrm{GW}_\varepsilon(\mu, \nu) = \mathrm{GW}_{1,1}(\mu, \nu) + \mathrm{GW}_{2,\varepsilon}(\mu, \nu)$ for centered μ, ν also holds in the unregularized case $\varepsilon = 0$ (Zhang et al., 2022, Section 4), it remains to bound the statistical error of $\mathrm{GW}_{2,0}(\hat{\mu}_n, \hat{\nu}_n)$. Considering Corollary 1 from Zhang et al. (2022) and the proofs of Lemma 2.1 and Theorem 2.2 from Hundrieser et al. (2024b), we see that Lemma 30 and Lemma 31 remain valid for $\varepsilon = 0$, where the entropic (c, ε)-transform is to be understood as the (measure-independent) c-transform. Hence, we can apply an adjusted version of Theorem 2.2 from Hundrieser et al. (2024b) and are left with providing suitable bounds on the uniform metric entropy of the function class $\mathcal{F}_{\mathcal{D},0}$. To this end, note that the cost functions $\{c_A\}_{A \in \mathcal{D}}$ are semi-concave and Lipschitz continuous in the first component with uniform moduli only depending on r and d, see (27). Thus, $\mathcal{F}_{\mathcal{D},0}$ is contained in the class of uniformly bounded, Lipschitz continuous and semi-concave functions on $\mathcal{X}$ and Lemma 42 provides the required bound on the uniform metric entropy. ∎

Appendix B. Uniform Metric Entropy

In this section, we give bounds on the uniform metric entropy of certain function classes. Recall that the uniform metric entropy is given by the logarithm of the covering numbers with respect to the uniform norm $\|\cdot\|_\infty$. For $\delta > 0$, the covering numbers of a function class $\mathcal{F}$ on $\mathcal{X}$ w.r.t. $\|\cdot\|_\infty$ are in turn defined as

$$N(\delta,\, \mathcal{F},\, \|\cdot\|_\infty) \coloneqq \inf\{n \in \mathbb{N} \mid \exists f_1, \dots, f_n : \mathcal{X} \to \mathbb{R} \text{ s.t. } \sup_{f \in \mathcal{F}} \min_{1 \le i \le n} \|f - f_i\|_\infty \le \delta\}\,.$$

To apply Theorem 6, we are interested in the uniform metric entropy of the class $\mathcal{F}_{c,\varepsilon}$ introduced in Proposition 3. Motivated by the following two lemmata, we consider in Section 3 the setting of $\mathcal{X} = \bigcup_{i=1}^I g_i(\mathcal{U}_i)$.

Lemma 39 (Union bound, Hundrieser et al. 2024b, Lemma 3.1)**.** *Let $\mathcal{F}$ be a class of functions on $\mathcal{X} = \bigcup_{i=1}^I \mathcal{X}_i$ for $I \in \mathbb{N}$ subsets $\mathcal{X}_i \subseteq \mathcal{X}$. Furthermore, denote with $\mathcal{F}\big|_{\mathcal{X}_i} \coloneqq \{\phi\big|_{\mathcal{X}_i} : \mathcal{X}_i \to \mathbb{R} \mid \phi \in \mathcal{F}\}$ the collection of functions in $\mathcal{F}$ restricted to $\mathcal{X}_i$. Then, it follows for any $\delta > 0$ that*

$$\log N(\delta,\, \mathcal{F},\, \|\cdot\|_\infty) \le \sum_{i=1}^I \log N(\delta,\, \mathcal{F}\big|_{\mathcal{X}_i},\, \|\cdot\|_\infty)\,.$$

Lemma 40 (Composition bound, Hundrieser et al. 2024b, Lemma A.1)**.** *Let $g : \mathcal{U} \to \mathcal{X}$ be a surjective map between the sets $\mathcal{U}$ and $\mathcal{X}$, and denote with $\mathcal{F}$ a function class on $\mathcal{U}$. Then, it follows for the composed function class $\mathcal{F} \circ g \coloneqq \{\phi \circ g \mid \phi \in \mathcal{F}\}$ on $\mathcal{X}$ for any $\delta > 0$ that*

$$N(\delta,\, \mathcal{F},\, \|\cdot\|_\infty) \le N(\delta,\, \mathcal{F} \circ g,\, \|\cdot\|_\infty)\,.$$

Hence, bounding the uniform metric entropy for the setting $\mathcal{X} = \bigcup_{i=1}^{I} g_i(\mathcal{U}_i)$ reduces to controlling $\mathcal{F}_{c,\varepsilon} \circ g_i$ for one i (provided that they all are of a similar structure). Furthermore, denoting $\tilde{c}_i \coloneqq c \circ (g_i, \mathrm{id}_{\mathcal{Y}})$, it holds for $\psi^{(c,\varepsilon,\xi)} \in \mathcal{F}_{c,\varepsilon}$ that $\psi^{(c,\varepsilon,\xi)} \circ g_i = \psi^{(\tilde{c}_i,\varepsilon,\xi)}$. In particular, this implies that $\mathcal{F}_{c,\varepsilon} \circ g_i = \mathcal{F}_{\tilde{c}_i,\varepsilon}$. By the definition of the entropic (c,ε)-transform, it can thus be seen that certain properties of $\tilde{c}_i$ are inherited to the function class $\mathcal{F}_{c,\varepsilon} \circ g_i$. In the following, we give uniform metric entropy bounds for function classes that contain $\mathcal{F}_{c,\varepsilon} \circ g_i$ for suitably chosen g_i. These results can then be combined with Theorem 6.

Lemma 41 (Kolmogorov and Tikhomirov 1961, Inequality (238))**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a connected metric space and $k > 0$ such that*

$$N(\delta,\, \mathcal{X},\, d_{\mathcal{X}}) \lesssim \delta^{-k} \qquad \text{for } \delta > 0 \text{ sufficiently small.}$$

Then, it follows for the same values of δ and the class $\mathcal{F}$ of 1-Lipschitz continuous functions on $\mathcal{X}$ which are bounded by one that

$$\log N(\delta,\, \mathcal{F},\, \|\bullet\|_\infty) \lesssim \delta^{-k}\,,$$

where the implicit constant only depends on k and $\mathcal{X}$.

Lemma 42. *Let $\mathcal{X} \subseteq \mathbb{R}^s$ be bounded and convex with $s \in \mathbb{N}$. Then, it follows for $\delta > 0$ sufficiently small and the class $\mathcal{F}$ of 1-Lipschitz continuous and 1-semi-concave functions on $\mathcal{X}$ which are bounded by one that*

$$\log N(\delta,\, \mathcal{F},\, \|\bullet\|_\infty) \lesssim \delta^{-s/2}\,,$$

where the implicit constant only depends on $\mathcal{X}$.

Proof The assertion follows from the proof of Lemma A.3 in Hundrieser et al. (2024b); for the sake of completeness we spell it out. By rotation and translation we may assume that $\mathcal{X}$ is contained in the linear subspace $V = \mathbb{R}^{\tilde{s}} \times \{0\}^{s-\tilde{s}}$ for $\tilde{s} \leq s$ and admits non-empty relative interior. By definition of $\mathcal{F}$, every function $\tilde{f}\colon \mathcal{X} \to \mathbb{R}, x \mapsto f(x) - \|x\|^2$ for $f \in \mathcal{F}$ is concave, L-Lipschitz with $L \coloneqq 1 + 2\,\mathrm{diam}(\mathcal{X})$ on $\mathcal{X}$ and bounded in absolute value by $1 + \mathrm{diam}(\mathcal{X})^2$. Hence, by Dragomirescu and Ivan (1992, Theorem 1 and Remark 2(ii)) there exists a concave L-Lipschitz extension $\overline{f}$ of $\tilde{f}$ onto a compact cube $\mathcal{D} \subseteq V$ with non-empty relative interior such that $\mathcal{X} \subseteq \mathcal{D}$. In particular, $\overline{f}$ is absolutely bounded by $B \coloneqq 1 + \mathrm{diam}(\mathcal{X})^2 + L\,\mathrm{diam}(\mathcal{D})$, and thus $\overline{f}$ is contained in the class $C_{B,L}(\mathcal{D})$ of concave functions that are bounded by B and L-Lipschitz. We thus conclude for $\delta > 0$ sufficiently small that

$$\begin{aligned}\log N(\delta, \mathcal{F}, \|\cdot\|_{\infty,\mathcal{X}}) &= \log N(\delta, \mathcal{F} - \|\cdot\|^2, \|\cdot\|_{\infty,\mathcal{X}}) \leq \log N(\delta, C_{B,L}(\mathcal{D}), \|\cdot\|_{\infty,\mathcal{X}}) \\ &\leq \log N(\delta, C_{B,L}(\mathcal{D}), \|\cdot\|_{\infty,\mathcal{D}}) \lesssim \delta^{-\tilde{s}/2} \lesssim \delta^{-s/2},\end{aligned}$$

where the second to last inequality follows by uniform metric entropy bounds on the class $C_{B,L}(\mathcal{D})$ detailed in Bronshtein (1976) or Guntuboyina and Sen (2013). In particular, the suppressed constant depends on $\mathcal{D}$, B and L, which again depends on $\mathcal{X}$. ■

Lemma 43 (van der Vaart and Wellner 1996, Theorem 2.7.1)**.** *Let $\mathcal{X} \subset \mathbb{R}^s$ be bounded and convex with nonempty interior. Then, it follows for $\delta > 0$ and the class $\mathcal{C}_M^\alpha(\mathcal{X})$ of α-Hölder smooth functions with $\alpha > 0$ and $M > 0$ that*

$$\log N(\delta,\, \mathcal{C}_M^\alpha(\mathcal{X}),\, \|\bullet\|_\infty) \lesssim M^{s/\alpha}\delta^{-s/\alpha}\,,$$

where the implicit constant only depends on s, α and $\mathcal{X}$.

Appendix C. Rescaling Properties

This appendix summarizes some useful insights on how the entropic optimal transport cost and corresponding convergence statements change under rescaling.

Remark 44 (Rescaling)**.** *Suppose that c is a bounded cost function that does not satisfy Assumption* **(C)***, i.e., it holds that $\|c\|_\infty \in (1,\infty)$. As for any $a > 0$, $b \in \mathbb{R}$ and $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ the EOT cost fulfills the rescaling property*

$$\mathrm{T}_{ac+b,\varepsilon}(\mu,\nu) = a \cdot \mathrm{T}_{c,\varepsilon/a}(\mu,\nu) + b,$$

we obtain for the empirical estimators $\widehat{\mathrm{T}}_{c,\varepsilon,n}$ in (5) with $a \coloneqq \|c\|_\infty$ that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] = a\,\mathbb{E}[|\widehat{\mathrm{T}}_{c/a,\varepsilon/a,n} - \mathrm{T}_{c/a,\varepsilon/a}(\mu,\nu)|]\,,$$

where the underlying cost function on the right-hand side satisfies Assumption **(C)***.*

Remark 45 (Rescaling of squared Euclidean distance)**.** *Denote by c the squared Euclidean distance, consider measurable sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ and let $r > 0$. For probability measures $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ denote by μ^{r^2} the pushforward of μ w.r.t. to the map $x \mapsto r^{-1}x$ and likewise define ν^{r^2}. Then, since $c(r^{-1}x, r^{-1}y) = r^{-2}c(x,y)$, we observe for any $\varepsilon > 0$ that*

$$\mathrm{T}_{c/r^2,\varepsilon}(\mu,\nu) = \mathrm{T}_{c,\varepsilon}(\mu^{r^2},\nu^{r^2})\,, \tag{28}$$

where μ^{r^2} and ν^{r^2} are supported on $r^{-1}\mathcal{X}$ and $r^{-1}\mathcal{Y}$, respectively. Moreover, by combining Remark 44 with the rescaling property (28) for $r = \varepsilon^{1/2} > 0$, it also follows that

$$\mathrm{T}_{c,\varepsilon}(\mu,\nu) = \varepsilon\,\mathrm{T}_{c/\varepsilon,1}(\mu,\nu) = \varepsilon\,\mathrm{T}_{c,1}(\mu^\varepsilon,\nu^\varepsilon).$$

In particular, μ^ε and ν^ε are supported in $\varepsilon^{-1/2}\mathcal{X}$ and $\varepsilon^{-1/2}\mathcal{Y}$, respectively. In addition, if μ and ν are σ^2-sub-Gaussian for some $\sigma^2 > 0$, then μ^ε and ν^ε are σ^2/ε-sub-Gaussian.

Remark 46 (Constants in convergence statements under rescaling)**.** *Under Assumption* **(C)** *and Assumption* **(SC)** *we have the constraint that the cost function as well as the Lipschitz and semi-concavity moduli must all be bounded by 1. As mentioned before, if one of these constraints is violated, then we can rescale appropriately via Remark 44 and still obtain bounds for the statistical error. We now discuss how this rescaling affects the constants.*

1. *First, we treat Assumption* **(Lip)** *and Assumption* **(SC)** *together (for the former, ignore the semi-concavity). Denote by L the Lipschitz modulus of the cost and by Λ its semi-concavity modulus. If $a \coloneqq [\|c\|_\infty \vee L \vee \Lambda] \geq 1$, then rescaling yields the additional factor a for the statistical error bounds in Theorem 14 and Theorem 15. This is due to the fact they do not depend on ε.*

2. *For Theorem 18, the effect of rescaling is more complicated. Indeed, it affects the Hölder constants $C^{(i,\alpha)}$ as well as the entropic regularization parameter ε. W.l.o.g. consider the case that $I = 1$, $g_1 = \mathrm{id}_\mathcal{X}$ and drop the index i. Let $a \coloneqq \|c\|_\infty > 1$ and denote the normalized cost $\tilde{c} \coloneqq c/a$ with its version $\tilde{C}^{(\alpha)}$ of $C^{(\alpha)}$. Due to the recursive*

and increasing structure, the relationship of $\tilde{C}^{(\alpha)}$ *and* $C^{(\alpha)}$ *is not straightforward. To simplify this, let*

$$C \coloneqq \max_{|j|=1,\dots,\alpha} \|\mathrm{D}^j c\|_\infty \,, \qquad \tilde{C} \coloneqq \max_{|j|=1,\dots,\alpha} \|\mathrm{D}^j \tilde{c}\|_\infty \,.$$

Then, it holds that $C = a\tilde{C}$ *and*

$$\tilde{C}^{(\alpha)} \leq \begin{cases} \tilde{C}^\alpha & \tilde{C} \geq 1 \,, \\ \tilde{C} & \tilde{C} < 1 \,, \end{cases} = a^{-\eta} C^\eta \qquad \text{with } \eta \coloneqq \begin{cases} \alpha & C \geq a \,, \\ 1 & C < a \,. \end{cases}$$

Using this in combination with Theorem 18 yields that

$$\mathbb{E}[|\widehat{\mathrm{T}}_{c,\varepsilon,n} - \mathrm{T}_{c,\varepsilon}(\mu,\nu)|] \lesssim a^{1+\frac{s}{2}[\frac{\alpha-1}{\alpha}-\frac{\eta}{\alpha}]} C^{\frac{\eta s}{2\alpha}} (\varepsilon \wedge a)^{-\frac{s}{2}\frac{\alpha-1}{\alpha}} \begin{cases} n^{-1/2} & s/\alpha < 2 \,, \\ n^{-1/2}\log(n+1) & s/\alpha = 2 \,, \\ n^{-\alpha/s} & s/\alpha > 2 \,, \end{cases}$$

Hence, we observe that the rescaling affects the statistical error bound polynomially.

References

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In I. Guyon, U. Von Luxburg, et al., editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. M. Altschuler, J. Niles-Weed, and A. J. Stromme. Asymptotics for semidiscrete entropic optimal transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741, 2022.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

E. Bayraktar, S. Eckstein, and X. Zhang. Stability and sample complexity of divergence regularized optimal transport. *Bernoulli [to appear, preprint arXiv:2212.00367]*, 2022.

E. Bernton, P. Ghosal, and M. Nutz. Entropic optimal transport: Geometry and large deviations. *Duke Mathematical Journal*, 171(16):3363–3400, 2022.

D. P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

D. P. Bertsekas and D. A. Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96, 1989.

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

E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539–563, 2014.

N. Bonneel and J. Digne. A survey of optimal transport for computer graphics and computer vision. *Computer Graphics Forum*, 42(2):439–460, 2023.

E. M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17 (3):393–398, 1976.

L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. In H. Larochelle, M. Ranzato, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269. Curran Associates, Inc., 2020.

C. W. Commander. A survey of the quadratic assignment problem, with applications. *Morehead Electronic Journal of Applicable Mathematics*, 4:MATH–2005–01, 2005.

G. Constantine and T. Savits. A multivariate Faa di Bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2):503–520, 1996.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In T. Calders and F. Esposito, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In I. Guyon, U. Von Luxburg, et al., editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

N. Deb, P. Ghosal, and B. Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In M. Ranzato, A. Beygelzimer, et al., editors, *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L^2-Wasserstein distance. *The Annals of Statistics*, 27(4): 1230–1239, 1999.

E. del Barrio, A. González-Sanz, J.-M. Loubes, and J. Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM Journal on Mathematics of Data Science*, 5(3):639–669, 2023.

A. Delalande. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In G. Camps-Valls, F. J. R. Ruiz, et al., editors, *International Conference on Artificial Intelligence and Statistics*, pages 1619–1642. PMLR, 2022.

F. Dragomirescu and C. Ivan. The smallest convex extensions of a convex function. *Optimization*, 24(3-4):193–206, 1992.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In J. Dy and A. Krause, editors, *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018.

S. Eckstein and M. Nutz. Convergence rates for regularized optimal transport via quantization. *Mathematics of Operations Research*, 49(2):1223–1240, 2024.

S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592, 2012.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 2016.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

A. Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, Apr. 2019.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu. Statistical inference with regularized optimal transport. *Information and Inference: A Journal of the IMA*, 13(1):iaad056, 2024a.

Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu. Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electronic Journal of Statistics*, 18(1):980–1041, 2024b.

A. González-Sanz and S. Hundrieser. Weak limits for empirical entropic optimal transport: Beyond smooth costs. *Preprint arXiv:2305.09745*, 2023.

A. González-Sanz, J.-M. Loubes, and J. Niles-Weed. Weak limits of entropy regularized optimal transport; potentials, plans and divergences. *Preprint arXiv:2207.07427*, 2022.

E. Grave, A. Joulin, and Q. Berthet. Unsupervised alignment of embeddings with Wasserstein procrustes. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, 2019.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. Von Luxburg, et al., editors, *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Guntuboyina and B. Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

M. Hallin and G. Mordant. Center-outward multiple-output Lorenz curves and Gini indices a measure transportation approach. *Preprint arXiv:2211.10822*, 2022.

M. Hallin, D. Hlubinka, and Š. Hudecová. Efficient fully distribution-free center-outward rank tests for multiple-output regression and MANOVA. *Journal of the American Statistical Association*, pages 1–17, 2022.

N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. In D. Precup and Y. W. Teh, editors, *International Conference on Machine Learning*, pages 1501–1509. PMLR, 2017.

S. Hundrieser, M. Klatt, and A. Munk. Limit distributions and sensitivity analysis for empirical entropic optimal transport on countable spaces. *The Annals of Applied Probability*, 34(1B):1403–1468, 2024a.

S. Hundrieser, T. Staudt, and A. Munk. Empirical optimal transport between different measures adapts to lower complexity. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):824–846, 2024b.

L. Kantorovitch. On the translocation of masses. *Doklady Akademii Nauk URSS*, 37:7–8, 1942.

L. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.

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

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. In S. Cernikov, N. Cernikova, et al., editors, *Twelve Papers on Algebra and Real Functions*, American Mathematical Society Translations–series 2, pages 277–364. American Mathematical Society, 1961.

J. M. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2013.

T. Lin and H. Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

J. Luo, D. Yang, and K. Wei. Improved complexity analysis of the Sinkhorn and Greenkhorn algorithms for optimal transport. *Preprint arXiv:2305.14939*, 2023.

T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135, 2024.

S. D. Marino and A. Gerolin. An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *Journal of Scientific Computing*, 85(2): 1–28, 2020.

F. Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In H. Wallach, H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

G. Monge. Mémoire sur la théeorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

G. Mordant and J. Segers. Measuring dependence between random vectors via optimal transport. *Journal of Multivariate Analysis*, 189:104912, 2022.

A. Munk and C. Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241, 1998.

T. G. Nies, T. Staudt, and A. Munk. Transport dependency: Optimal transport based dependency measures. *Preprint arXiv:2105.02073*, 2021.

J. Niles-Weed and P. Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

M. Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

M. Nutz and J. Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

J. Orlin. A faster strongly polynomial minimum cost flow algorithm. In *Proceedings of the Twentieth annual ACM symposium on Theory of Computing*, pages 377–387, 1988.

S. Pal. On the difference between entropic cost and the optimal transport cost. *The Annals of Applied Probability*, 34(1B):1003–1028, 2024.

V. M. Panaretos and Y. Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019.

V. M. Panaretos and Y. Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In M. F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *Preprint arXiv:2109.12004*, 2021.

S. Rachev and L. Rüschendorf. *Mass transportation problems - Volume I: Theory*. Springer, 1998a.

S. Rachev and L. Rüschendorf. *Mass transportation problems - Volume II: Applications*. Springer, 1998b.

P. Rigollet and A. J. Stromme. On the sample complexity of entropic optimal transport. *Preprint arXiv:2206.13472*, 2022.

G. Rioux, Z. Goldfeld, and K. Kato. Entropic Gromov-Wasserstein distances: Stability, algorithms, and distributional limits. *Preprint arXiv:2306.00182*, 2023.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):1–11, 2015.

J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics*, 35(4):1–13, 2016.

M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(1):219–238, 2018.

T. Staudt and S. Hundrieser. Convergence of empirical optimal transport in unbounded settings. *Bernoulli [to appear, preprint arXiv:2306.11499]*, 2024.

A. J. Stromme. Minimum intrinsic dimension scaling for entropic optimal transport. *Preprint arXiv:2306.03398*, 2023.

K.-T. Sturm. The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. *Preprint arXiv:1208.0434v2*, 2012.

A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

C. Tameling, S. Stoldt, T. Stephan, J. Naas, S. Jakobs, and A. Munk. Colocalization for super-resolution microscopy via optimal transport. *Nature computational science*, 1(3): 199–211, 2021.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

C. Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

S. Wang, T. T. Cai, and H. Li. Optimal estimation of Wasserstein distance on a tree with an application to microbiome studies. *Journal of the American Statistical Association*, 116 (535):1237–1253, 2021.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

C. A. Weitkamp, K. Proksch, C. Tameling, and A. Munk. Distribution of distances based object matching: Asymptotic inference. *Journal of the American Statistical Association*, 2022.

Z. Zhang, Z. Goldfeld, Y. Mroueh, and B. K. Sriperumbudur. Gromov-Wasserstein distances: Entropic regularization, duality, and sample complexity. *The Annals of Statistics [to appear, preprint arXiv:2212.12848]*, 2022.

B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.