Convergence Rates for Regularized Optimal Transport via Quantization*

Stephan Eckstein† Marcel Nutz‡

June 21, 2023

Abstract

We study the convergence of divergence-regularized optimal transport as the regularization parameter vanishes. Sharp rates for general divergences including relative entropy or L^p regularization, general transport costs and multi-marginal problems are obtained. A novel methodology using quantization and martingale couplings is suitable for non-compact marginals and achieves, in particular, the sharp leading-order term of entropically regularized 2-Wasserstein distance for marginals with finite $(2+\delta)$-moment.

Keywords Entropic Optimal Transport; f-Divergence; Regularization; Quantization
AMS 2010 Subject Classification 90C25; 49N05

1 Introduction

We study regularized optimal transport problems of the form

$$\mathsf{OT}_{f,\varepsilon} := \inf_{\pi\in\Pi(\mu_1,\dots,\mu_N)} \int c\, d\pi + \varepsilon D_f(\pi, \mu_1\otimes\cdots\otimes\mu_N)$$

where D_f is an f-divergence, for example relative entropy (Kullback–Leibler divergence) or L^p regularization. (Notation is detailed in Section 2.) Note that $\varepsilon = 0$ yields the classical optimal transport problem OT without regularization. We are interested in the speed of convergence $\mathsf{OT}_{f,\varepsilon} \to \mathsf{OT}$ as the regularization parameter ε tends to zero, especially its dependence on the marginals μ_i and the divergence D_f.

*The authors thank Guillaume Carlier, Lénaïc Chizat, Nicolas Juillet, Harald Luschgy, Jon Niles-Weed, Gilles Pagès and Luca Tamanini for discussions and encouragement.

†Department of Mathematics, ETH Zurich, seckstein@ethz.ch.

‡Departments of Statistics and Mathematics, Columbia University, mnutz@columbia.edu. Research supported by an Alfred P. Sloan Fellowship and NSF Grants DMS-1812661, DMS-2106056.

Regularized optimal transport has attracted a great deal of research in recent years, chiefly because regularization enables the use of efficient numerical algorithms (e.g, [10, 25, 40, 55] and the references therein) to approximate OT in high-dimensional applications—whence the interest in the speed of convergence. The most important divergence is relative entropy which gives rise to Sinkhorn's algorithm (or IPFP); here $\mathsf{OT}_{f,\varepsilon}$ is often called the *entropic optimal transport* problem (e.g., [49, 55]). Other divergences, especially L^p regularization, are being used in applications where sparse optimizers are desired or weak penalization (small ε) causes numerical instabilities with entropic regularization [11, 26, 31, 42, 58]. For multi-marginal transport and the related Wasserstein barycenters, see for instance [2, 6, 14, 16, 17]. Literature more specific to the convergence $\mathsf{OT}_{f,\varepsilon} \to \mathsf{OT}$ is discussed below.

In this paper, we propose a novel methodology to estimate $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT}$ based on quantization. It is simultaneously more general and, arguably, easier than previous arguments, allowing us to obtain convergence rates for a wide class of f-divergences, unbounded cost functions and multi-marginal problems in a unified manner—the methodology may be as important as the results themselves. Even for entropic optimal transport with two marginals and quadratic cost, we substantially improve on the existing results, by allowing for arbitrary marginals with finite $(2+\delta)$-moments where previous techniques required compact supports and uniformly bounded densities [18, 22, 24, 53].

To give an informal preview, let us focus on $N = 2$ marginals with entropic or L^p regularization ($p > 1$) for simplicity. In those examples, we obtain non-asymptotic bounds of the form

$$\begin{aligned} &\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le \beta\varepsilon \log\left(\frac{1}{\varepsilon}\right) + K\varepsilon && \text{for entropic regularization,} \\ &\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le K\varepsilon^{\frac{1}{(p-1)\beta+1}} && \text{for } L^p \text{ regularization,} \end{aligned}$$

where β reflects a certain quantization dimension. In our first result (Theorem 3.3), β encodes the optimal quantization rate for one of the marginals—if μ_i are measures on $\mathbb{R}^{d_i}$, this leads to $\beta \le d_1 \wedge d_2$. In this result, we assume that the integrated cost $\pi \mapsto \int c\,d\pi$ is Lipschitz when restricted to a certain set of couplings; this is satisfied for Lipschitz functions c but also, e.g., for $|x-y|^p$ with $p \ge 1$ on $\mathbb{R}^d \times \mathbb{R}^d$. The stated estimates are sharp in certain examples (see Section 4), up to the constant K.

The key idea is to use so-called *shadows* to transfer explicit divergence bounds for discrete measures into continuous couplings with controlled divergence, while also bounding the Wasserstein distance. As quantization

theory has long studied how fast general measures can be approximated with discrete ones, this enables us to control both the transport and divergence terms in $\mathsf{OT}_{f,\varepsilon}$. Specifically, a rate is found by choosing the number of points for the quantization of the marginals relative to the regularization parameter ε such as to balance the transport and divergence terms. At a high level, the shadow construction is a substitute for the widely used *block approximation* method first introduced in [15]. Employing quantization and Wasserstein geodesics instead of building blocks explicitly, our construction fully exploits the flexibility of the p-Wasserstein distance, making it very suitable for unbounded domains and costs.

Our main result (Theorem 3.8) pertains to cost functions on $\mathbb{R}^d \times \cdots \times \mathbb{R}^d$ admitting a bounded second derivative, in particular the quadratic cost, and improves the value of β to $d/2$ under sufficient regularity. Here, smoothness leads to the factor $1/2$ while d reflects the quantization rate for an optimal transport plan (of the unregularized problem OT) rather than the marginals. The key idea is a martingale argument that seems to be novel: the martingale property of 2-Wasserstein quantization can be used to eliminate the first-order term in the integrated Taylor expansion of the cost function. The remaining leading term is then of second order, whence the factor $1/2$. Once again, the martingale methodology lends itself to the unbounded setting; moreover, the rates are sharp in wide class of examples. In particular, we establish the leading-order term $\frac{d}{2}\varepsilon \log\left(\frac{1}{\varepsilon}\right)$ for entropically regularized 2-Wasserstein distance whenever the marginals have finite moments of order $2+\delta$ for some $\delta > 0$ (Corollary 3.14). In its proof, Minty's trick [48] is used to establish the quantization rate for an optimal transport plan.

For discrete problems, the study of entropic regularization and its convergence goes back to [23]; see also [60] for a non-asymptotic result, [5] for a semi-discrete problem, and [4] for multi-marginal transport. Here, we are mainly interested in continuous problems. As $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} = O(\varepsilon)$ if and only if there exists an optimal transport with finite divergence (Proposition A.1) and as the latter typically fails for continuous marginals, we shall be dealing with convergence slower than $O(\varepsilon)$. In the continuous case, we are not aware of works addressing the multi-marginal problem, and for two marginals, almost all results are on the entropic regularization; an exception is [44] where χ^2 divergence is studied in a compact setting and an upper bound of order $\varepsilon^{1/(d+1)}$ is found. Returning to the entropic case, the link between $\mathsf{OT}_{f,\varepsilon}$ and OT goes back to [46, 47] in the Schrödinger bridge problem (which is closely related to entropic optimal transport with quadratic cost; cf. [39]). Gamma-convergence was shown in [38]; see also [15] for a proof in a setting closer to ours. A stochastic control viewpoint is presented in [19]. Early

quantitative results for quadratic cost, from a large deviations viewpoint, are [1, 27, 30], later extended in [53] to cost functions closely modeled on the quadratic. While these are first-order results, a second-order expansion of the optimal cost was obtained in [24] for the Schrödinger bridge setting and in [22] for entropic optimal transport, all with quadratic cost. These results require strong regularity assumptions in addition to compactly supported marginals.

The most comparable results by far were obtained in the very recent (and partly concurrent) work [18] which addresses general cost functions and obtains rates similar to ours, at least for compactly supported marginals, in the case of entropic regularization with two marginals. Remarkably, the methods used are quite different. For Lipschitz cost functions and compactly supported marginals, [18, Proposition 3.1] finds that $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq d\varepsilon \log(1/\varepsilon) + O(\varepsilon)$ where d is the minimum of the two marginal dimensions. A potentially more general result is obtained with a notion of upper Rényi dimension of the marginals, however a more concrete bound is only available through the box dimension which requires compactness to be finite.[1] The proof proceeds through a block approximation, applying the Lipschitz property on each block. Our Theorem 3.3 (specialized to the entropic divergence on two marginals) obtains a bound of the same form but with the dimension defined by quantization. Using p-Wasserstein distance with finite p, quantization is well-behaved also for unbounded domains, so that the bound can be established for general marginals with finite $(p+\delta)$-moments. Moreover, Theorem 3.3 applies to costs like $|x-y|^p$, $p \geq 1$ as the Lipschitz property is only required in an integrated form. Shadows are a convenient and robust tool in this context, as is also exemplified by their application to adapted (causal) optimal transport in [29].

For cost functions of class $C^{1,1}$ (thus with a.e. bounded second derivative) and compactly supported marginals with uniformly bounded Lebesgue densities, [18, Proposition 3.4] shows that $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \frac{d}{2}\varepsilon \log\left(\frac{1}{\varepsilon}\right) + O(\varepsilon)$. The proof is deep and based on the fine regularity of the Kantorovich potential; namely, a quadratic bound on the integrated difference between a λ-convex function and its first-order Taylor expansion [18, Lemma 3.6]. This bound depends directly on the diameter of the domain and the density assumption is needed to pass from Lebesgue measure to the actual marginals. By contrast, the martingale argument used for our Theorem 3.8 applies to

[1] Note added in proof: This statement referred to the preprint version of [18]. The final published version provides an improved result; namely, the upper Rényi dimension is bounded by the Euclidean dimension as soon as the marginal has a finite logarithmic moment.

unbounded domains and is fairly robust; for instance, it easily extends to the multi-marginal case. It does, however, take as its input the quantization rate of an optimal transport plan π^*, so that it needs to be applied together with a regularity result for π^*. For quadratic cost, we prove that the rate is indeed $1/d$ in great generality, assuming only finite moments of order $2+\delta$. For compactly supported marginals, a quite generic sufficient condition for this rate is the nondegeneracy of the cost; i.e., invertibility of the mixed derivative $D^2_{xy}c(x,y)$. For unbounded but sufficiently integrable marginals, we show a rate arbitrarily close to $1/d$ if nondegeneracy holds in a uniform sense.

In [18], the authors also obtain a matching lower bound for the convergence rate (for entropic regularization), for cost functions satisfying the aforementioned nondegeneracy condition and sufficiently regular marginals. The proof is again based on a fine analysis of the Kantorovich potential. The key tool is a quadratic detachment estimate [18, Lemma 4.2] which we reuse in Section 4 to obtain matching lower bounds for L^p regularization as well.

While the present work focuses on the convergence of the optimal cost $\mathsf{OT}_{f,\varepsilon}$, two related question are the convergence of the optimal couplings and optimal dual potentials. See [9, 15, 38, 39] and [8, 21, 36, 50, 56], respectively, and the references therein. As seen in [9, 21], the convergence is also related to the stability of $\mathsf{OT}_{f,\varepsilon}$ wrt. the marginals [17, 28, 35, 51].

The remainder of this paper is organized as follows. Section 2 formally introduces the problem and notation, then gathers preliminaries on quantization, divergence bounds for discrete couplings, and shadows. Section 3 contains the main results on convergence rates. Section 4 provides instances where the rates are sharp and Appendix A gathers two additional results.

2 Preliminaries

2.1 Setting and Notation

Let (Y, d_Y) be a Polish space and $\mathcal{P}(Y)$ its set of Borel probability measures. Fix $p \in [1,\infty)$ and denote by $\mathcal{P}_p(Y)$ the subset of measures μ with finite p-th moment; i.e., $\int d_Y(x,\hat{x})^p \, \mu(dx) < \infty$ for some (and then all) $\hat{x} \in Y$. The p-Wasserstein distance $W_p(\mu,\nu)$ between $\mu,\nu \in \mathcal{P}_p(Y)$ is defined via

$$W_p(\mu,\nu)^p = \inf_{\pi \in \Pi(\mu,\nu)} \int d_Y(x,y)^p \, \pi(dx,dy).$$

Fix $N \in \mathbb{N}$ and let (X_i, d_{X_i}), $i = 1,\dots,N$ be Polish probability spaces with measures $\mu_i \in \mathcal{P}(X_i)$. We denote by $X = \prod_{i=1}^N X_i$ the product space and use

the particular product metric $d_{X,p}(x,y) := (\sum_{i=1}^N d_{X_i}(x_i,y_i)^p)^{1/p}$ to induce the p-Wasserstein distance on X.

Let $c : X \to \mathbb{R}$ be continuous with growth of order p; i.e.,

$$|c(x)| \le C\big(1 + d_{X,p}(x,\hat{x})^p\big)$$

for some $C \ge 0$ and $\hat{x} \in X$. The optimal transport problem is

$$\mathsf{OT} := \inf_{\pi \in \Pi(\mu_1,\dots,\mu_N)} \int c \, d\pi,$$

where $\Pi(\mu_1,\dots,\mu_N) \subset \mathcal{P}_p(X)$ denotes the set of couplings of the marginal measures $\mu_i \in \mathcal{P}_p(X_i)$. The growth of c ensures that OT is finite.

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex, lower bounded function with $f(1) = 0$ and $\lim_{x\to\infty} f(x)/x = \infty$. The *$f$-divergence* $D_f(\mu,\nu)$ between probabilities μ, ν on a common space is defined by

$$D_f(\mu,\nu) := \int f\left(\frac{d\mu}{d\nu}\right) d\nu \quad \text{for} \quad \mu \ll \nu$$

and $D_f(\mu,\nu) := \infty$ for $\mu \not\ll \nu$. The D_f-regularized transport problem is

$$\mathsf{OT}_{f,\varepsilon} := \inf_{\pi \in \Pi(\mu_1,\dots,\mu_N)} \int c \, d\pi + \varepsilon D_f(\pi, P), \quad P := \mu_1 \otimes \cdots \otimes \mu_N,$$

where $\varepsilon > 0$ is the regularization parameter. In particular, entropic optimal transport corresponds to $f(x) = x\log(x)$.

2.2 Quantization

On a Polish space Y, we denote by $\mathcal{P}^n(Y) \subset \mathcal{P}(Y)$ the set of probability measures supported on at most n points. Given $p \in [1,\infty)$ and $\mu \in \mathcal{P}_p(Y)$, our results depend on an approximation rate of the form

$$\exists\, \mu^n \in \mathcal{P}^n(Y) : \quad W_p(\mu^n,\mu) \le C n^{-\alpha}, \quad n \ge 1 \qquad (\mathsf{quant}_p(C,\alpha))$$

for constants $C \ge 0$ and $\alpha > 0$. The takeaway of the following is that if the support of μ is d-dimensional, this property typically holds with $\alpha = 1/d$.

Remark 2.1 (Quantization rate on $\mathbb{R}^d$)**.** Let $Y = \mathbb{R}^d$. If $\mu \in \mathcal{P}_{p+\delta}(Y)$ for some $\delta > 0$, then $\mathsf{quant}_p(C,\alpha)$ holds with $\alpha = 1/d$ for some $C \ge 0$. More precisely, [37, Theorem 6.2] shows that the exact asymptotic constant

$$C_a := \lim_{n\to\infty} n^{1/d} \inf_{\mu^n \in \mathcal{P}^n(\mathbb{R}^d)} W_p(\mu^n,\mu)$$

can be expressed through a dimensional constant related to the p-quantization of the uniform measure on the unit cube and a moment of the density of the absolutely continuous part of μ. In particular, $C_a > 0$ as soon as μ is not mutually singular wrt. Lebesgue measure, showing that the rate $\alpha = 1/d$ is then optimal. A bound for the (non-asymptotic) constant C in $\mathsf{quant}_p(C, \alpha)$ is given in [37, Corollary 6.7]; its proof yields an explicit constant valid for all $n \geq 1$ depending only on p, δ, d and $\int |x|^{p+\delta} \, \mu(dx)$.[2]

For some variations of our results (in fact, only in the multi-marginal case of Theorem 3.3 with non-entropic divergence), we use a slightly stronger notion, sometimes called (deterministic) empirical quantization, where the approximating measures are required to be uniform. Let $\mathcal{P}^{n,em}(Y) \subset \mathcal{P}(Y)$ be the set of uniform measures on n points; i.e., measures $\mu^n = n^{-1} \sum_{i=1}^n \delta_{y_i}$ for some $y_i \in Y$. Similarly as above, we introduce

$$\exists \, \mu^n \in \mathcal{P}^{n,em}(Y) : \quad W_p(\mu^n, \mu) \leq C n^{-\alpha}, \quad n \geq 1 \qquad (\mathsf{quant}_p^{em}(C, \alpha))$$

for constants $C \geq 0$ and $\alpha > 0$. This condition clearly implies $\mathsf{quant}_p(C, \alpha)$; but at least in the high-dimensional regime, the optimal rate is in fact the same, as summarized in the following remark.

Remark 2.2 (Empirical quantization rate on $\mathbb{R}^d$)**.** Let $Y = \mathbb{R}^d$. The well known [33, Theorem 1] shows, among other things, that if $\mu \in \mathcal{P}_{2p+\delta}(\mathbb{R}^d)$ with $d > 2p$, then $\mathsf{quant}_p^{em}(C, \alpha)$ holds with $\alpha = 1/d$ and a constant C depending only on d, p, δ and the $(2p + \delta)$-moment of μ. In particular, this bound for the empirical rate coincides with the bound $1/d$ given for (arbitrary) quantization in Remark 2.1. Rates for other regimes ($d \leq 2p$) are also obtained in [33, Theorem 1]. Notably, the rates derived in [33] are not based on a deterministic construction of μ^n, but hold a.s. when μ^n are i.i.d. samples of μ. More precise constants for this result, and non-asymptotic bounds, can be found in the very recent work [32]. Rates for i.i.d. samples of measures supported on compact submanifolds are studied in [61].

For measures with bounded support, a deterministic construction in [20, Theorem 3] provides the rate $\alpha = 1/d$ and an explicit constant C for $p < d$; for $p = d$, a logarithmic correction is added, whereas for $p > d$, the rate is at least $\alpha = 1/p$. For unbounded measures, [20, Corollary 1] shows a slightly looser bound for the rate under the condition $\mu \in \mathcal{P}_{p+\delta}(Y)$. The univariate

[2] The result in [37, Corollary 6.7] is stated for all $n \geq C_3$ instead of $n \geq 1$, for a certain constant C_3, in order to have a statement whose constants do not depend on the moment $\int |x|^{p+\delta} \, \mu(dx)$. For our purposes, we do not mind such a dependence, and we can easily deduce a result valid for all $n \geq 1$ by adjusting the constants.

case $d = 1$ has been studied in detail [7, 62]. Here the optimal rate is $\alpha = 1$ if μ has a positive density on its support and is sufficiently integrable, whereas $\alpha < 1$ is known in several other cases (see [7, Table 1] for an overview).

2.3 Elementary Divergence Bounds

For our purposes, discrete measures are useful because they admit straightforward divergence bounds. The best-known example is that a coupling $\pi \in \Pi(\mu_1, \mu_2)$ of marginals μ_i supported on n points has relative entropy $D_f(\pi, \mu_1 \otimes \mu_2) \le \log n$. The following lemma collects some extensions of that fact for later reference. We recall that $\mathcal{P}^n(X_i)$ denotes the probabilities supported on at most n points, $\mathcal{P}^{n,em}(X_i)$ the empirical measures on n points, and $P = \mu_1 \otimes \cdots \otimes \mu_N$.

Lemma 2.3 (Divergence bounds)**.** *Let $\pi \in \Pi(\mu_1, \dots, \mu_N)$ and define φ by $f(x) = x\varphi(x)$. Assume that φ is nondecreasing.*

(i) If $N = 2$, φ is concave, and $\mu_2 \in \mathcal{P}^{n_2}(X_2)$, then $D_f(\pi, P) \le \varphi(n_2)$.

(ii) If $\mu_i \in \mathcal{P}^{n_i,em}(X_i)$ for $i = 2, \dots, N$, then $D_f(\pi, P) \le \varphi\big(\prod_{i=2}^N n_i\big)$.

(iii) If $\varphi(x) = \log(x)$ and $\mu_i \in \mathcal{P}^{n_i}(X_i)$ for $i = 2, \dots, N$, then $D_f(\pi, P) \le \sum_{i=2}^N \log(n_i)$.

Proof. Denote by $\pi_{2:N}$ the marginal of π on $X_2 \times \cdots \times X_N$. In particular, $P_{2:N} = \mu_2 \otimes \cdots \otimes \mu_N$. We similarly define $\pi_{1:N-1}$ and $P_{1:N-1}$ as the marginals on $X_1 \times \cdots \times X_{N-1}$. Let σ be the counting measure on the (finite) support of $P_{2:n}$. Disintegrating $\pi = \mu_1 \otimes K$, we then have $\frac{d\pi}{dP} = \frac{dK}{dP_{2:n}} \le \frac{d\sigma}{dP_{2:n}}$, hence

$$D_f(\pi, P) = \int \varphi\left(\frac{d\pi}{dP}\right) d\pi \le \int \varphi\left(\frac{d\sigma}{dP_{2:N}}\right) d\pi.$$

In the case (i) where $N = 2$, Jensen's inequality yields

$$\int \varphi\left(\frac{d\sigma}{dP_{2:N}}\right) d\pi = \int \varphi\left(\frac{d\sigma}{d\mu_2}\right) d\mu_2 \le \varphi(n_2).$$

Whereas in (ii), $\frac{d\sigma}{dP_{2:N}}$ is constant and thus $\int \varphi\big(\frac{d\sigma}{dP_{2:N}}\big)\, d\pi = \varphi\big(\prod_{i=2}^N n_i\big)$. To see (iii), we write $\frac{d\pi}{dP} = \frac{d\pi}{d(\pi_{1:N-1} \otimes \mu_N)} \frac{d(\pi_{1:N-1} \otimes \mu_N)}{dP} = \frac{d\pi}{d(\pi_{1:N-1} \otimes \mu_N)} \frac{d(\pi_{1:N-1})}{dP_{1:N-1}}$. As $\varphi(x) = \log(x)$, this yields

$$D_f(\pi, P) = D_f(\pi, \pi_{1:N-1} \otimes \mu_N) + D_f(\pi_{1:N-1}, P_{1:N-1}).$$

To bound the first term, we apply (i) with μ_N as second marginal,

$$D_f(\pi, P) \le \log(n_N) + D_f(\pi_{1:N-1}, P_{1:N-1}).$$

Iterating this argument yields $D_f(\pi, P) \le \sum_{i=2}^{N} \log(n_i)$, which was the claim. □

2.4 Shadows

Given $\pi \in \Pi(\mu_1, \dots, \mu_N)$, the shadow $\tilde{\pi}$ of π on another vector $(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ of marginals is a particular W_p-projection of π onto $\Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ that enjoys a control on its divergence. Intuitively, for $N = 2$, the shadow $\tilde{\pi}$ is obtained by concatenating three transports: move $\tilde{\mu}_1$ to μ_1 using a W_p-optimal transport, then follow the transport π moving μ_1 into μ_2, and finally move μ_2 to $\tilde{\mu}_2$ using a W_p-optimal transport. The general definition follows.

Definition 2.4 ([28])**.** Let $p \in [1, \infty)$ and $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$, $i = 1, \dots, N$. Let $\kappa_i \in \Pi(\mu_i, \tilde{\mu}_i)$ be a coupling attaining $W_p(\mu_i, \tilde{\mu}_i)$ and $\kappa_i = \mu_i \otimes K_i$ a disintegration. Given $\pi \in \Pi(\mu_1, \dots, \mu_N)$, its *shadow* $\tilde{\pi}$ on $(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ is defined as the second marginal of $\pi \otimes K \in \mathcal{P}(X \times X)$, where the kernel $K : X \to \mathcal{P}(X)$ is defined as $K(x) = K_1(x_1) \otimes \cdots \otimes K_N(x_N)$.

The definition and the data processing inequality readily imply the following properties; see [28, Lemma 3.2] for a detailed proof.

Lemma 2.5 (Shadow bounds)**.** *Let $p \in [1, \infty)$ and $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$, $i = 1, \dots, N$. Given $\pi \in \Pi(\mu_1, \dots, \mu_N)$, its shadow $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ satisfies*

$$W_p(\pi, \tilde{\pi})^p = \sum_{i=1}^{N} W_p(\mu_i, \tilde{\mu}_i)^p,$$
$$D_f(\tilde{\pi}, \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N) \le D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N).$$

3 Main Results

One novel idea in this paper is to use a "double shadow" through auxiliary discrete marginals to approximate a given (typically singular) transport plan with one that has controlled divergence. To illustrate this, we start by reproving the (known) convergence $\mathsf{OT}_{f,\varepsilon} \to \mathsf{OT}$ in our general setting.

Proposition 3.1. *Let $p \in [1, \infty)$ and $\mu_i \in \mathcal{P}_p(X_i)$ for $i = 1, \dots, N$. If c is continuous with growth of order p, then $\lim_{\varepsilon \to 0} \mathsf{OT}_{f,\varepsilon} = \mathsf{OT}$.*

Proof. Using tightness of $\{\mu_i\}$, we can construct measures μ_i^n supported on n points with $W_p(\mu_i^n, \mu_i) \to 0$ for $i = 1, \dots, N$. Let $\pi^* \in \Pi(\mu_1, \dots, \mu_N)$ be an optimizer of OT. We introduce another coupling $\pi^n \in \Pi(\mu_1, \dots, \mu_N)$ as follows: first, let $\tilde{\pi}$ be the shadow of π^* onto $(\mu_1^n, \mu_2^n, \dots, \mu_N^n)$; then, define π^n as the shadow of $\tilde{\pi}$ onto $(\mu_1, \dots, \mu_N)$. Using the triangle inequality and Lemma 2.5, this implies

$$W_p(\pi^n, \pi^*) \le W_p(\pi^n, \tilde{\pi}) + W_p(\tilde{\pi}, \pi^*) \le 2 \left(\sum_{i=1}^N W_p(\mu_i^n, \mu_i)^p \right)^{1/p} \to 0.$$

As c is continuous with growth of order p, we conclude $\int c \, d\pi^n \to \int c \, d\pi^*$. On the other hand, Lemma 2.5 yields

$$D_f(\pi^n, P) \le D_f(\tilde{\pi}, \mu_1^n \otimes \mu_2^n \otimes \dots \mu_N^n) < \infty,$$

where the finiteness is trivial by discreteness of μ_i^n. Given $\delta > 0$, choose n such that $\int c \, d\pi^n - \int c \, d\pi^* \le \delta$, and then $\varepsilon_0 > 0$ such that $\varepsilon_0 D_f(\pi^n, P) \le \delta$. As π^n is an admissible coupling for $\mathsf{OT}_{f,\varepsilon}$, we have shown $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le 2\delta$ for all $\varepsilon \le \varepsilon_0$. □

3.1 Rate for Lipschitz-type Costs

To enable a quantitative version of Proposition 3.1, we need to control the speed of convergence $\int c \, d\pi^n \to \int c \, d\pi^*$ in its proof. We introduce the following adaptation of the condition (A_L) of [28], stating that the integrated transport cost is Lipschitz with respect to the coupling.

Definition 3.2. Let $p \in [1, \infty)$ and $\mu_i \in \mathcal{P}_p(X_i)$, $i = 1, \dots, N$. Given constants $L, C \ge 0$, we say that c satisfies $(\mathrm{A}_{L,C})$ if for all $\tilde{\mu}_i \in \mathcal{P}_p(X_i)$ with $W_p(\tilde{\mu}_i, \mu_i) \le C$, $i = 1, \dots, N$, we have

$$\left| \int c \, d(\pi - \tilde{\pi}) \right| \le L W_p(\pi, \tilde{\pi}) \tag{$\mathrm{A}_{L,C}$}$$

for all $\pi \in \Pi(\mu_1, \dots, \mu_N)$ and $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$.

Clearly $(\mathrm{A}_{L,C})$ is satisfied (for all C) if c is L-Lipschitz, but as discussed in [28, Example 3.6], the condition also captures various non-Lipschitz costs, like $c(x_1, x_2) = |x_1 - x_2|^p$ on $\mathbb{R}^d \times \mathbb{R}^d$ with $p \in [1, \infty)$. In that case, the constant L depends on the moments of the μ_i and on C. (The condition does not capture $|x_1 - x_2|^r$ for $0 < r < 1$. An extension with a modulus of continuity instead of a Lipschitz constant is discussed in Remark A.2.)

Theorem 3.3. *Let $p \in [1,\infty)$ and $\mu_i \in \mathcal{P}_p(X_i)$ for $i = 1,\dots,N$. Assume that μ_i satisfies $\mathsf{quant}_p(C,\alpha_i)$ for $i = 2,\dots,N$ and that c satisfies $(\mathrm{A}_{L,C})$, for some $\alpha_2,\dots,\alpha_N \in (0,1]$ and $L, C \geq 0$.*[3]

(i) Let $f(x) = x\log(x)$. Then for all $\varepsilon \in (0,1]$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \left(\sum_{i=2}^{N} \frac{1}{\alpha_i}\right) \varepsilon \log\left(\frac{1}{\varepsilon}\right) + 4(N-1)^{1/p} LC\varepsilon.$$

(ii) Let $f(x) = x\varphi(x)$, $\beta = \sum_{i=2}^{N} \frac{1}{\alpha_i}$, $\tilde{f}(x) = x\varphi(x^\beta)$. Assume that for some $x_0, y_0 \geq 0$, $\tilde{f}$ is strictly increasing on $[x_0,\infty)$ with inverse $\tilde{f}_{\mathrm{inv}}$ and φ is nondecreasing. Suppose also that either $N = 2$ and φ is concave, or the μ_i satisfy $\mathsf{quant}_p^{em}(C,\alpha_i)$ instead of $\mathsf{quant}_p(C,\alpha_i)$. Set $S_\varepsilon = \tilde{f}_{\mathrm{inv}}(\frac{1}{\varepsilon})$, which satisfies $\lim_{\varepsilon\to 0} S_\varepsilon = \infty$ and $\lim_{\varepsilon\to 0} \varepsilon S_\varepsilon = 0$. Then for all $\varepsilon \in [0, 1/x_0]$ small enough such that $S_\varepsilon \geq y_0^{1/\beta} + 1$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \frac{4(N-1)^{1/p} LC + 1}{S_\varepsilon}.$$

While the quantity S_ε in Theorem 3.3 (ii) may not admit a closed-form expression, we can deduce more explicit bounds as follows.

Example 3.4 (Explicit bounds)**.** Choose a function $\psi \geq \varphi$ such that $\tilde{g}(x) := x\psi(x^\beta)$ is strictly increasing with inverse denoted $\tilde{g}_{\mathrm{inv}}$. Then $\tilde{g}_{\mathrm{inv}} \leq \tilde{f}_{\mathrm{inv}}$ and hence $1/S_\varepsilon \leq 1/\tilde{g}_{\mathrm{inv}}(1/\varepsilon)$, so that Theorem 3.3 (ii) implies

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq (4(N-1)^{1/p} LC + 1) \frac{1}{\tilde{g}_{\mathrm{inv}}(1/\varepsilon)}.$$

We thus aim to choose ψ so that $\tilde{g}_{\mathrm{inv}}$ has an explicit expression. As an example, consider the L^ρ regularization given by $f(x) = \frac{1}{\rho}(x^\rho - 1)$ with $\rho > 1$. Here $\varphi(x) = \frac{1}{\rho}x^{\rho-1} - \frac{1}{\rho x} \leq \frac{1}{\rho}x^{\rho-1} =: \psi(x)$. With this choice of ψ, we have $\tilde{g}(x) = \frac{1}{\rho}x^{(\rho-1)\beta+1}$ and the explicit inverse $\tilde{g}_{\mathrm{inv}}(x) = \rho x^{1/[(\rho-1)\beta+1]}$. As a result, for all $\varepsilon \in (0,1]$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq K\varepsilon^{\frac{1}{(\rho-1)\beta+1}}, \quad K := (4(N-1)^{1/p} LC + 1)/\rho.$$

[3] Exponents $\alpha_i > 1$ could be accommodated with minor changes in the constants below. In view of Remark 2.1, the condition $\alpha_i \leq 1$ is not restrictive in practice.

Remark 3.5 (On μ_1)**.** In Theorem 3.3, nothing is assumed about the quantization of μ_1. In an application, one would thus label μ_1 the marginal with the slowest quantization rate. In particular, for $N = 2$ marginals on $\mathbb{R}^{d_i}$, we typically have $1/\alpha_2 = d_1 \wedge d_2$ by Remark 2.1.

Proof of Theorem 3.3. Let $\pi^* \in \Pi(\mu_1, \dots, \mu_N)$ be an optimizer of OT. By our assumption, there exist empirical quantizations $\mu_i^{n_i}$ for the marginals $i = 2, \dots, N$ such that $W_p(\mu_i^{n_i}, \mu_i) \le C n_i^{-\alpha_i}$. We introduce a coupling $\pi \in \Pi(\mu_1, \dots, \mu_N)$ (depending on $n_2, \dots, n_N$) as a double shadow: first, let $\tilde{\pi}$ be the shadow of π^* onto $(\mu_1, \mu_2^{n_2}, \dots, \mu_N^{n_N})$; then, define π as the shadow of $\tilde{\pi}$ onto $(\mu_1, \dots, \mu_N)$. Using the triangle inequality and Lemma 2.5,

$$W_p(\pi, \pi^*) \le W_p(\pi, \tilde{\pi}) + W_p(\tilde{\pi}, \pi^*) \le 2 \left(\sum_{i=2}^{N} W_p(\mu_i^{n_i}, \mu_i)^p \right)^{1/p}.$$

Combining this with our assumption $(\mathrm{A}_{L,C})$, we deduce

$$\int c \, d\pi - \int c \, d\pi^* \le 2L \left(\sum_{i=2}^{N} W_p(\mu_i^{n_i}, \mu_i)^p \right)^{1/p} \le 2LC \left(\sum_{i=2}^{N} n_i^{-\alpha_i p} \right)^{1/p}.$$

On the other hand, Lemma 2.5 again yields

$$D_f(\pi, P) \le D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \dots \mu_N^{n_N}).$$

As π is an admissible coupling for $\mathsf{OT}_{f,\varepsilon}$, we have proved

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le 2LC \left(\sum_{i=2}^{N} n_i^{-\alpha_i p} \right)^{1/p} + \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \dots \mu_N^{n_N}) \tag{3.1}$$

and the last divergence term can be bounded by Lemma 2.3. In the remainder of the proof, we choose n_i as a suitable function of ε to balance the decay of the two terms on the right-hand side of (3.1). As n_i is an integer, we need to deal with a rounding error: given $S \in [1, \infty)$, we define $\varrho(S) > 0$ as

$$\varrho(S) := \left(\frac{1}{N-1} \sum_{i=2}^{N} \frac{S^p}{\lfloor S^{1/\alpha_i} \rfloor^{\alpha_i p}} \right)^{1/p} \tag{3.2}$$

so that $1 \le \varrho(S) \le 2^{\max_{i \ge 2} \alpha_i} \le 2$ and $\lim_{S \to \infty} \varrho(S) = 1$. We then have

$$\left(\sum_{i=2}^{N} \lfloor S^{1/\alpha_i} \rfloor^{-\alpha_i p} \right)^{1/p} = \frac{\varrho(S)(N-1)^{1/p}}{S} \le \frac{2(N-1)^{1/p}}{S}. \tag{3.3}$$

(i) Set $n_i = \lfloor \varepsilon^{-1/\alpha_i} \rfloor$ for $i = 2, \dots, N$. For $S = S_\varepsilon = 1/\varepsilon$, (3.3) yields

$$\left(\sum_{i=2}^{N} n_i^{-\alpha_i p} \right)^{1/p} = \frac{\varrho(S_\varepsilon)(N-1)^{1/p}}{S_\varepsilon} \leq 2(N-1)^{1/p}\varepsilon,$$

and Lemma 2.3 (iii) bounds the divergence term by

$$\varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \dots \mu_N^{n_N}) \leq \varepsilon \sum_{i=2}^{N} \log(n_i) \leq \varepsilon \sum_{i=2}^{N} \frac{1}{\alpha_i} \log\left(\frac{1}{\varepsilon}\right).$$

In view of (3.1), the claim follows.

(ii) Set $n_i = \lfloor S_\varepsilon^{1/\alpha_i} \rfloor$ for $i = 2, \dots, N$, where S_ε was defined in the theorem. Similarly as in (i),

$$\left(\sum_{i=2}^{N} n_i^{-\alpha_i p} \right)^{1/p} \leq \frac{\varrho(S_\varepsilon)(N-1)^{1/p}}{S_\varepsilon} \leq 2(N-1)^{1/p} \frac{1}{S_\varepsilon}.$$

On the other hand, $S_\varepsilon \geq y_0^{1/\beta} + 1$ implies $y_0 \leq \prod_{i=2}^{N} n_i \leq S_\varepsilon^\beta$ by elementary arguments. Under $\mathsf{quant}_p^{em}(C, \alpha_i)$, Lemma 2.3 (ii) and monotonicity of φ on $[y_0, \infty)$ yield

$$\begin{aligned} \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \dots \mu_N^{n_N}) \leq \varepsilon\varphi\left(\prod_{i=2}^{N} n_i \right) \leq \varepsilon\varphi(S_\varepsilon^\beta) = \frac{\varepsilon \tilde{f}(S_\varepsilon)}{S_\varepsilon} \\ = \frac{\varepsilon \tilde{f}(\tilde{f}_{\mathrm{inv}}(\frac{1}{\varepsilon}))}{S_\varepsilon} = \frac{1}{S_\varepsilon} \end{aligned}$$

and now the claim again follows from (3.1). For the claim under $N = 2$, we use Lemma 2.3 (i) instead of Lemma 2.3 (ii). □

Remark 3.6 (On the constant)**.** The constant 4 in Theorem 3.3 (i),(ii) can be replaced by $2\varrho(1/\varepsilon)$ and $2\varrho(S_\varepsilon)$, respectively, where $\varrho(\cdot)$ is defined in (3.2) and satisfies $1 \leq \varrho(\cdot) \leq 2$. As $\varrho(S) = 1 + o(1/S)$, this improves the asymptotic constant for $\varepsilon \to 0$ in Theorem 3.3 from 4 to 2.

Remark 3.7 (On the proof)**.** In Theorem 3.3 and its proof, the entropic case (i) is treated separately from the general case (ii) to obtain an expression that is more explicit and more in line with the literature. In fact, the bound in Theorem 3.3 (ii) is slightly sharper even for the entropic divergence, as its proof is based on the optimal tradeoff between the transport and divergence

terms: both have the same rate $1/S_\varepsilon$, whereas in the proof of (i) they have differing rates ε and $\varepsilon \log(1/\varepsilon)$. However, $S_\varepsilon = \tilde{f}_{\mathrm{inv}}(\frac{1}{\varepsilon})$ does not admit an explicit expression in the entropic case, so we chose instead $S_\varepsilon = 1/\varepsilon$ to obtain an explicit statement. The leading-order term nevertheless turns out to be sharp; see Proposition 4.2.

3.2 Rate for Twice Differentiable Costs

For the main result, we focus on the exponent $p = 2$ for the Wasserstein metric and on closed convex sets $X_i \subset \mathbb{R}^{d_i}$ endowed with the Euclidean norm $|\cdot|$. We recall that $X = X_1 \times \cdots \times X_N$ then also carries the Euclidean metric and write $c \in C^2(X)$ to indicate that c is defined and twice continuously differentiable on a neighborhood of $X \subset \mathbb{R}^{d_1+\cdots+d_N}$.

For costs with bounded second derivative and an additional regularity condition, we shall improve upon the dimension-dependence in Theorem 3.3 by a factor $1/2$, at least for marginals of equal dimension. For that improvement, $(\mathrm{A}_{L,C})$ is too weak (as evidenced in Proposition 4.2). Instead, we shall use a martingale argument to achieve a full cancellation of the integrated first-order term in the Taylor expansion of c. For this, we directly quantize an optimal transport, not just the marginals. In the following statement, its quantization rate α is taken as given—we shall elaborate below on how to bound it in practice.

Theorem 3.8. *Let $X_i \subset \mathbb{R}^{d_i}$ be convex and $\mu_i \in \mathcal{P}_2(X_i)$ for $i = 1, \dots, N$. Assume that $c \in C^2(X)$ has bounded second derivative:*

$$w^\top c''(x) w \le B|w|^2 \quad \text{for all} \quad x, w \in X, \quad \text{for some} \quad B \ge 0, \tag{3.4}$$

and that OT admits an optimal transport π^* satisfying $\mathsf{quant}_2(C, \alpha)$ for some $\alpha \in (0, 1]$ and $C > 0$.

(i) Let $f(x) = x \log(x)$. Then for all $\varepsilon \in (0, 1]$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le \frac{N-1}{2\alpha} \varepsilon \log\left(\frac{1}{\varepsilon}\right) + 8BC\varepsilon.$$

(ii) Let $N = 2$, $f(x) = x\varphi(x)$ with φ nondecreasing and concave, let $\beta = \frac{1}{2\alpha}$ and $\tilde{f}(x) = x\varphi(x^\beta)$. Assume that for some $x_0 \ge 0$, $\tilde{f}$ is strictly increasing on $[x_0, \infty)$ with inverse $\tilde{f}_{\mathrm{inv}}$. Set $S_\varepsilon = \tilde{f}_{\mathrm{inv}}(\frac{1}{\varepsilon})$, which satisfies $\lim_{\varepsilon \to 0} S_\varepsilon = \infty$ and $\lim_{\varepsilon \to 0} \varepsilon S_\varepsilon = 0$. Then for all $\varepsilon \in (0, \frac{1}{x_0}]$ small enough such that $S_\varepsilon \ge 1$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \le \frac{8BC + 1}{S_\varepsilon}.$$

Before proving the theorem, we recall the martingale property of W_2-quantization; see, e.g., [52, Proposition 5.1, p. 139] for a proof. This property and its interplay with the Taylor expansion in (3.5) below explain why our result is limited to $p = 2$.

Lemma 3.9. *Given a probability $\eta \in \mathcal{P}_2(Y)$ on a Polish space Y and $n \geq 1$, there exists $\eta^n \in \arg\min_{\eta^n \in \mathcal{P}^n(Y)} W_2(\eta^n, \eta)$, called an optimal W_2-quantizer of η on n points. There is a coupling $\theta \in \Pi(\eta^n, \eta)$ attaining $W_2(\eta^n, \eta)$, meaning that $\int |x - y|^2 \, \theta(dx, dy) = W_2(\eta^n, \eta)^2$, and it is a martingale: the kernel κ in its disintegration $\theta = \eta^n \otimes \kappa$ satisfies $\int y \, \kappa(x, dy) = x$ for $\tilde{\pi}$-almost all x.*

Proof of Theorem 3.8. For $n \geq 1$, let $\tilde{\pi} \in \mathcal{P}(X)$ be an optimal W_2-quantizer of π^* on n points and let $\theta \in \Pi(\tilde{\pi}, \pi^*)$ be the coupling attaining $W_2(\tilde{\pi}, \pi^*)$; cf. Lemma 3.9. The martingale property of θ implies that $\int h(x) \cdot (y - x) \, \theta(dx, dy) = 0$ for any measurable function $h : X \to \mathbb{R}^{d_1 + \cdots + d_N}$ of linear growth. As c has bounded second derivative, its first derivative c' has linear growth and thus

$$\int c'(x) \cdot (y - x) \, \theta(dx, dy) = 0.$$

Considering the Taylor expansion of $c(y)$, this shows that the integral of the first-order term vanishes, and then the bound on the second derivative yields

$$\begin{aligned} \left| \int c \, d\pi^* - \int c \, d\tilde{\pi} \right| &= \left| \int (c(y) - c(x)) \, \theta(dx, dy) \right| \\ &\leq B \int |x - y|^2 \, \theta(dx, dy) = B W_2(\tilde{\pi}, \pi^*)^2. \end{aligned} \tag{3.5}$$

Denote by μ_i^n the marginal of $\tilde{\pi}$ on X_i and by θ_i the marginal of θ on $X_i \times X_i$. We observe that $\theta_i \in \Pi(\mu_i^n, \mu_i)$ is again a martingale coupling. Furthermore, as we are using the Euclidean norm,

$$\sum_{i=1}^{N} \int |x_i - y_i|^2 \, \theta_i(dx_i, dy_i) = \int |x - y|^2 \, \theta(dx, dy) = W_2(\tilde{\pi}, \pi^*)^2. \tag{3.6}$$

Next, we construct a coupling $\pi \in \Pi(\mu_1, \ldots, \mu_N)$ that is reminiscent of the shadow of $\tilde{\pi}$ but uses the kernels of θ_i instead of W_2-optimal transports between μ_i^n and μ_i. Namely, decomposing $\theta_i = \mu_i^n \otimes K_i$ and writing $K(x) := K_1(x_1) \otimes \cdots \otimes K_N(x_N)$, we set $\gamma := \tilde{\pi} \otimes K \in \mathcal{P}(X \times X)$ and define $\pi \in \Pi(\mu_1, \ldots, \mu_N)$ as the second marginal of γ. Probabilistically speaking, this means that we take the (possibly dependent) components of the vector

martingale θ and combine their laws into a new vector martingale γ with independent components. In particular, $\gamma \in \Pi(\tilde{\pi}, \pi)$ is also a martingale coupling: $\int y_i \, K(x, dy) = \int y_i \, K_i(x_i, dy_i) = x_i$ for all i by the martingale property of θ_i. Repeating the argument for (3.5) with γ instead of θ, inserting the definition of γ and using (3.6), we conclude that

$$\begin{aligned}\left| \int c \, d\pi - \int c \, d\tilde{\pi} \right| &\leq B \int |x - y|^2 \, \gamma(dx, dy) \\ &= B \sum_{i=1}^{N} \int |x_i - y_i|^2 \, \theta_i(dx_i, dy_i) = B W_2(\tilde{\pi}, \pi^*)^2.\end{aligned}$$

In view of (3.5), the triangle inequality and the assumption on π^* then yield

$$\int c \, d\pi - \int c \, d\pi^* \leq 2B W_2(\tilde{\pi}, \pi^*)^2 \leq 2BC n^{-2\alpha}. \tag{3.7}$$

On the other hand, by the data processing inequality (e.g., [49, Lemma 1.6]), the construction of π implies

$$D_f(\pi, P) \leq D_f(\tilde{\pi}, \mu_1^n \otimes \cdots \otimes \mu_N^n).$$

This bound is analogous to Lemma 2.5 (indeed the reasoning is the same).

The rest of the proof is analogous to Theorem 3.3. To deal with the rounding error, we now define $\varrho(S)$ for $S \in [1, \infty)$ as

$$\varrho(S) := \left(\frac{S^{\frac{1}{2\alpha}}}{\lfloor S^{\frac{1}{2\alpha}} \rfloor} \right)^{2\alpha} \tag{3.8}$$

so that $1 \leq \varrho(S) \leq 2^{2\alpha} \leq 4$ and $\lim_{S \to \infty} \varrho(S) = 1$. In particular,

$$\lfloor S^{\frac{1}{2\alpha}} \rfloor^{-2\alpha} = \varrho(S) S^{-1} \leq 4 S^{-1}. \tag{3.9}$$

(i) Let $n = \lfloor \varepsilon^{-\frac{1}{2\alpha}} \rfloor$. Then (3.7) and (3.9) for $S = S_\varepsilon = 1/\varepsilon$ imply

$$\int c \, d\pi - \int c \, d\pi^* \leq 2BC \varrho(S_\varepsilon) S_\varepsilon^{-1} \leq 8BC \varrho(S_\varepsilon) \varepsilon$$

while Lemma 2.3 (iii) yields $D_f(\tilde{\pi}, \mu_1^n \otimes \cdots \otimes \mu_N^n) \leq (N-1) \log(n)$, completing the proof of (i).

(ii) Here we define $n = \lfloor S_\varepsilon^{\frac{1}{2\alpha}} \rfloor$, then (3.7) and (3.9) yield

$$\int c \, d\pi - \int c \, d\pi^* \leq \frac{2BC \varrho(S_\varepsilon)}{S_\varepsilon} \leq \frac{8BC}{S_\varepsilon}$$

while (recall $N = 2$) Lemma 2.3(i) yields $D_f(\tilde{\pi}, \mu_1^n \otimes \mu_2^n) \leq \varphi(n)$ and thus

$$\varepsilon D_f(\pi, P) \leq \varepsilon\varphi(n) \leq \frac{\varepsilon\varphi(S_\varepsilon^{\frac{1}{2\alpha}})S_\varepsilon}{S_\varepsilon} = \frac{1}{S_\varepsilon},$$

completing the proof. □

Similarly as in Remark 3.6, the asymptotic constant in Theorem 3.8 can be improved from 8 to 2.

Remark 3.10 (Relaxing C^2 condition)**.** Theorem 3.8 immediately extends to slightly less regular costs: if $(c_n)_{n\in\mathbb{N}}$ is a sequence of cost functions satisfying the assumptions of Theorem 3.8 and $\lim_{n\to\infty} \|c_n - c\|_\infty = 0$ for some $c : \mathbb{R}^d \to \mathbb{R}$, then

$$\mathsf{OT}_{f,\varepsilon}(c) - \mathsf{OT}(c) \leq 2\|c_n - c\|_\infty + \mathsf{OT}_{f,\varepsilon}(c^n) - \mathsf{OT}(c^n)$$

as both $\mathsf{OT}_{f,\varepsilon}$ and OT are 1-Lipschitz with respect to $\|\cdot\|_\infty$, so that Theorem 3.8 applies to c as well.

We also have the following analogue of Example 3.4.

Example 3.11 (L^ρ regularization)**.** For the L^ρ regularization $f(x) = \frac{1}{\rho}(x^\rho - 1)$ with $\rho > 1$, Theorem 3.8 (ii) implies that for all $\varepsilon \in (0, 1]$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq K\varepsilon^{\frac{1}{(\rho-1)\beta+1}}, \quad K := (8BC + 1)/\rho,$$

by the same algebra as in Example 3.4. (Of course, β now has a different definition).

Remark 3.12 (Comparison with Theorem 3.3)**.** Let $N = 2$ for simplicity. As any quantization of the coupling π^* induces quantizations for its marginals, it is clear that $\alpha \leq \alpha_2$. In the best case, we have $\alpha = \alpha_2$, and then Theorem 3.8 yields an improvement of $1/2$ over Theorem 3.3. Note that $\alpha = \alpha_2$ will typically be the case if $d_1 = d_2 =: d$ and the support of π^* is also d-dimensional—more on this in a moment.

On the flip side, as Theorem 3.8 implicitly quantizes all the marginals, there is no immediate benefit to having a faster rate for one marginal as in Remark 3.5. Thus there are situations where Theorem 3.3 actually yields a better rate, especially if $d_1 > 2d_2$. But of course, $d_1 = d_2$ is the most important setting.

To obtain a good result from Theorem 3.8, we need to know that OT admits an optimal transport π^* satisfying $\mathsf{quant}_2(C, \alpha)$ for some good α. Indeed, $\mathsf{quant}_2(C, \alpha)$ holds trivially for $1/\alpha = d_1 + \cdots + d_N$ (under a moment condition), but that does not yield the desired improvement over Theorem 3.3. On the other hand, suppose that π^* is given by a Lipschitz transport map over X_1, then π^* inherits the quantization rate from μ_1, so that $1/\alpha = d_1$. The existence of such a map has been studied intensely in the regularity theory of optimal transport, see [12, 13] and the literature thereafter. However, the conditions are known to be very restrictive [41, 43], and clearly a Lipschitz map can almost never be expected for unbounded marginals. On the other hand, as emphasized in [45], a lower dimensional structure does not require a transport map at all.

In the following, we provide some results for $N = 2$ marginals, and remark briefly on the multi-marginal case. Generally speaking, any result on the structure of optimal transports can be combined with Theorem 3.8. The next result covers the most important example—the quadratic cost defining 2-Wasserstein distance—under a minimal condition on the marginals (which includes many situations where no coupling is given by a map).

Lemma 3.13. *Consider $c(x, y) = |x - y|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mu_1, \mu_2 \in \mathcal{P}_{2+\delta}(\mathbb{R}^d)$ for some $\delta > 0$. Then any optimal transport satisfies $\mathsf{quant}_2(C, 1/d)$ for some $C > 0$.*

Proof. Let $\Delta = \{(x, x) : x \in \mathbb{R}^d\}$ be the diagonal and $\mathrm{proj}^\Delta : \mathbb{R}^{2d} \to \Delta$ the Euclidean orthogonal projection. Let $\pi \in \Pi(\mu_1, \mu_2)$ be an optimal transport, then $\pi \in \mathcal{P}_{2+\delta}(\mathbb{R}^{2d})$ due to the assumption on the marginals. Define the pushforward measure

$$\eta := \mathrm{proj}^\Delta_\# \pi$$

which is concentrated on Δ; we claim that η satisfies $\mathsf{quant}_2(C, 1/d)$. Consider the rotated coordinates (u, v) given by

$$u = \frac{x + y}{\sqrt{2}}, \quad v = \frac{x - y}{\sqrt{2}}$$

in which $\Delta = \{(u, 0) : u \in \mathbb{R}^d\}$ and proj^Δ can be written as $(u, v) \mapsto (u, 0)$. Thus η can be seen as a measure on $\mathbb{R}^d$ and with that identification,

$$\int |u|^{2+\delta} \, d\eta = \int |(u, v)|^{2+\delta} \, d\pi = \int |(x, y)|^{2+\delta} \, d\pi < \infty.$$

By Remark 2.1, $\eta \in \mathcal{P}_{2+\delta}(\mathbb{R}^n)$ implies that η satisfies $\mathsf{quant}_2(C, 1/d)$.

To show the same rate for π, we use Minty's trick [48] along the lines of [3]. Recall that the support $\Gamma := \operatorname{spt}\pi$ is c-cyclically monotone (e.g., [59]), which for quadratic cost means

$$\langle x' - x, y' - y\rangle \geq 0, \quad (x,y),(x',y') \in \Gamma.$$

In the rotated coordinates, this implies that

$$|v' - v| \leq |u' - u|, \quad (u,v),(u',v') \in \Gamma.$$

In particular, $u = u'$ implies $v = v'$, meaning that proj^{Δ} admits an inverse map $\ell : \mathrm{proj}^{\Delta}(\Gamma) \to \Gamma$, $(u,0) \mapsto (u,v)$ and moreover ℓ is $\sqrt{2}$-Lipschitz. By Kirszbraun's theorem, we can extend ℓ to a $\sqrt{2}$-Lipschitz map $\Delta \to \mathbb{R}^d \times \mathbb{R}^d$, still denoted ℓ. Note that $\pi = \ell_{\#}\eta$ and any quantization of η on Δ pushes forward to a quantization of π. In view of the $\sqrt{2}$-Lipschitz property, we conclude that π satisfies $\mathsf{quant}_2(\sqrt{2}C, 1/d)$. □

The following combines Lemma 3.13 with Theorem 3.8 and Example 3.11.

Corollary 3.14 (Quadratic cost)**.** *Consider $c(x,y) = |x-y|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mu_1, \mu_2 \in \mathcal{P}_{2+\delta}(\mathbb{R}^d)$ for some $\delta > 0$.*

(i) Let $f(x) = x\log(x)$. There exists $K > 0$ such that

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \frac{d}{2}\varepsilon \log\left(\frac{1}{\varepsilon}\right) + K\varepsilon, \qquad \varepsilon \in (0,1].$$

(ii) Let $f(x) = \frac{1}{\rho}(x^\rho - 1)$ with $\rho > 1$. There exists $K > 0$ such that

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq K\varepsilon^{\frac{1}{(\rho-1)d/2+1}}, \qquad \varepsilon \in (0,1].$$

Next, we aim to generalize Lemma 3.13 from quadratic to more general costs. Following [45], the basic idea is that a fairly generic cost is locally equivalent to a perturbation of the quadratic cost after a change of coordinates. Let $X_1, X_2 \subset \mathbb{R}^d$ be convex and $c \in \mathcal{C}^2(X)$. We say that c is *nondegenerate* if $D^2_{xy}c(x,y)$ is invertible for all $(x,y) \in X$. Here $D^2_{xy}c(x,y)$ denotes the $d \times d$ matrix $[\partial^2_{x_i y_j} c(x,y)]_{1\leq i,j\leq d}$. We follow the terminology of [45]; the condition is called (A2) in [43] while [18] calls such c infinitesimally twisted.

If the support can be covered by finitely many such local coordinate changes, we obtain the same quantization rate as in the quadratic case. In particular, this holds for compact support.

Lemma 3.15. *Let $X_1, X_2 \subset \mathbb{R}^d$ be convex and let $c \in \mathcal{C}^2(X)$ be nondegenerate. If μ_1, μ_2 are compactly supported, then any optimal transport satisfies $\mathsf{quant}_2(C, 1/d)$ for some $C > 0$.*

For a proof, see Steps 1 and 2 in the proof of Lemma 3.16 below. Next, we address the unbounded case; here we assume that nondegeneracy holds in a uniform sense (which is automatic in the compact case) and achieve a rate arbitrarily close to $1/d$, under sufficient integrability. The proof is a combination of the proofs of Lemma 3.13 and [45, Theorem 1.1] with a cut-off argument. We denote by $\|M\|$ the operator norm of the matrix M.

Lemma 3.16. *Let $X_1, X_2 \subset \mathbb{R}^d$ be convex and let $c \in \mathcal{C}^2(X)$ be nondegenerate. Suppose that $D^2_{xy}c(x,y)$ is uniformly continuous and $\|D^2_{xy}c\|$, $\|(D^2_{xy}c)^{-1}\|$ are bounded on X. Let $d' > d$. If $\mu_1, \mu_2 \in \mathcal{P}_q(\mathbb{R}^d)$ for $q := 2\frac{d'+d}{d'-d}$, then any optimal transport satisfies $\mathsf{quant}_2(C, 1/d')$ for some $C > 0$.*

Proof. Let π be an optimal transport. Whenever a subprobability ν is given, we denote by $\widetilde{\nu} = \nu/\nu(X)$ its normalized measure.

Step 1. Consider a cube $Q = ([-r, r]^{2d} + \{(x_0, y_0)\}) \cap X$ centered at $(x_0, y_0) \in \operatorname{spt}\pi$. We show that for r sufficiently small, $\widetilde{\pi|_Q}$ satisfies $\mathsf{quant}_2(C, 1/d)$ with a constant C independent of (x_0, y_0). Let $M := D^2_{xy}c(x_0, y_0) \in \mathbb{R}^{d\times d}$ and $G(x,y) := -c(x, -M^{-1}y) - x \cdot y$. Then

$$\begin{aligned} D^2_{xy}G(x,y) &= D^2_{xy}c(x, -M^{-1}y)M^{-1} - \mathbf{1}_n \\ &= D^2_{xy}c(x, -M^{-1}y)M^{-1} - D^2_{xy}c(x_0, y_0)M^{-1} \end{aligned}$$

and hence

$$\|D^2_{xy}G(x,y)\| \le \|M^{-1}\|\|D^2_{xy}c(x, -M^{-1}y) - D^2_{xy}c(x_0, y_0)\|.$$

As $D^2_{xy}c$ is uniformly continuous and $\|(D^2_{xy}c)^{-1}\|$ is uniformly bounded, we can thus choose $r \in (0,1)$ independent of (x_0, y_0) such that $\|D^2_{xy}G(x,y)\| \le \frac{1}{2}$ for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$ with $(x, -M^{-1}y) \in Q$.

Consider $(x,y), (x',y')$ such that $(x, -M^{-1}y), (x', -M^{-1}y') \in Q \cap \operatorname{spt}\pi$. Then the c-cyclical monotonicity of $\operatorname{spt}\pi$ yields

$$c(x, -M^{-1}y) + (x', -M^{-1}y') \le c(x, -M^{-1}y) + c(x', -M^{-1}y')$$

or equivalently

$$x \cdot y + G(x,y) + x'y' + G(x',y') \ge x \cdot y' + G(x,y') + x' \cdot y + G(x',y). \quad (3.10)$$

Next, we use a second change of coordinates

$$u = \frac{x+y}{\sqrt{2}}, \quad v = \frac{x-y}{\sqrt{2}}.$$

Closely following the proof of [45, Theorem 1.2], using (3.10) with $\Delta x := x' - x$, $\Delta y := y' - y$, $\Delta u := u' - u$, $\Delta v := v' - v$ leads to

$$\Delta x \cdot \Delta y + \Delta x \cdot \int_0^1 \int_0^1 D^2_{xy} G(x + s\Delta x, y + t\Delta y)\Delta y \, ds \, dt \geq 0$$

and hence $\Delta x \cdot \Delta y \geq -\frac{1}{2}|\Delta x||\Delta y|$ as $\|D^2_{xy} G\| \leq \frac{1}{2}$ along the integration domain. Noting that $\Delta y\sqrt{2} = \Delta u + \Delta v$ and $\Delta x\sqrt{2} = \Delta u - \Delta v$, we deduce

$$\begin{aligned} |\Delta u|^2 - |\Delta v|^2 = 2\Delta x \cdot \Delta y &\geq -|\Delta x||\Delta y| \\ &\geq -\frac{1}{2}(|\Delta x|^2 + |\Delta y|^2) = -\frac{1}{2}(|\Delta u|^2 + |\Delta v|^2) \end{aligned}$$

and thus

$$|\Delta v| \leq \sqrt{3}|\Delta u|. \tag{3.11}$$

Consider the composition $a = a_3 \circ a_2 \circ a_1$ of the linear maps

$$a_1 : (x, -M^{-1}y) \mapsto (x, y), \quad a_2 : (x, y) \mapsto (u, v), \quad a_3 : (u, v) \mapsto u.$$

Clearly, the image $I = a(\mathbb{R}^{2d})$ is a d-dimensional linear subspace. Defining $\eta := a_\# \widetilde{\pi|_Q}$, we see that $\operatorname{spt} \eta$ is a bounded subset of I. Its diameter admits a bound depending only on r and the Lipschitz constant of a, and the latter is independent of (x_0, y_0) as $\|M\| = \|D_{xy} c(x_0, y_0)\|$ is uniformly bounded. Recall from Remark 2.1 that a measure on $\mathbb{R}^d$ with bounded support satisfies $\mathsf{quant}_2(C_0, 1/d)$ with a constant C_0 depending only on d and the diameter of the support (note that the diameter bounds any moment). As a result, η satisfies $\mathsf{quant}_2(C_0, 1/d)$ with a constant C_0 independent of (x_0, y_0).

The map a admits a Lipschitz inverse $\ell : a(Q \cap \operatorname{spt} \pi) \to Q \cap \operatorname{spt} \pi$, with a Lipschitz constant L independent of (x_0, y_0) due to the boundedness of $\|(D^2_{xy} c)^{-1}\|$ and (3.11). Again, by Kirszbraun's theorem, ℓ extends to a Lipschitz map $\ell : I \to \mathbb{R}^d \times \mathbb{R}^d$ with the same Lipschitz constant. As $\widetilde{\pi|_Q} = \ell_\# \eta$, we deduce that $\widetilde{\pi|_Q}$ satisfies $\mathsf{quant}_2(C, 1/d)$ for $C = LC_0$.

Step 2. We start with a general observation about sums. Let $\nu_1, \ldots, \nu_m$ be subprobabilities with a cumulative mass of at most one and suppose that each $\widetilde{\nu_i}$ satisfies $\mathsf{quant}_2(C, \alpha)$ for $n \geq 1$. Consider the quantization problem for the sum $\nu = \sum_{i=1}^m \nu_i$, which can be seen as the convex combination

$\sum_{i=1}^m \nu_i(X)\widetilde{\nu}_i$ of probability measures (and the zero measure, if necessary). Noting that given $n = km$ points, we can allocate k points to each of the $\widetilde{\nu}_i$, it is easy to see that $\widetilde{\nu}$ satisfies $\mathsf{quant}_2(m^\alpha C, \alpha)$ for all $n \in \{m, 2m, \cdots\}$, and thus for all $n \geq m$ after increasing the constant C.

For $N \in \mathbb{N}$, consider $R = Nr$ and the cube $Q_R = [-R, R]^{2d}$, which can be divided into $m := N^{2d}$ small cubes of the type in Step 1. Combining Step 1 with the observation about sums, we see that $\widetilde{\pi|_{Q_R}}$ satisfies $\mathsf{quant}_2((R/r)^2 C, 1/d)$ for $n \geq (R/r)^{2d}$.

We note that if the marginals are compactly supported, Q_R contains $\operatorname{spt}\pi$ for R sufficiently large, so that π satisfies $\mathsf{quant}_2(C, 1/d)$ after increasing C. For the noncompact case, we use the following cut-off.

Step 3. For $n \geq 1$, choose $R = R(n)$ as

$$R(n) := r \left\lfloor n^{\frac{1}{d} - \frac{1}{d'}} \right\rfloor^{1/2}.$$

Note $\lim_{n\to\infty} R(n) = \infty$ and $n \geq (R(n)/r)^{2d}$ and $(R(n)/r)^2 n^{-1/d} \leq n^{-1/d'}$. Writing $\pi_n := \pi|_{Q_{R(n)}}$, the above shows that there exist $\nu_n \in \mathcal{P}^n(\mathbb{R}^{2d})$ such that $W_2(\nu_n, \widetilde{\pi_n}) \leq C n^{-1/d'}$. On the other hand, consider $\pi - \pi_n$, which is supported outside $[-R(n), R(n)]^{2d}$. As a consequence,

$$\int |z|^2 \, d(\pi - \pi_n) \leq R(n)^{-\gamma} \int |z|^{2+\gamma} \, d\pi$$

for any $\gamma \geq 0$. Choose $\gamma := \frac{4d}{d'-d} = \frac{4}{d'}(\frac{1}{d} - \frac{1}{d'})^{-1}$; then $R(n)^{-\gamma} \leq C' n^{-2/d'}$ for a constant $C' > 0$ and as $2 + \gamma = 2\frac{d'+d}{d'-d}$, the integral is finite by our assumption on the marginals. Quantizing $\pi - \pi_n$ by a single point mass at the origin, we then see with the result for π_n that π satisfies $\mathsf{quant}_2(C, 1/d')$ for a (different) constant C. □

Lemma 3.15 and Lemma 3.16 have immediate corollaries similar to Corollary 3.14; we omit the statements for brevity.

The nondegeneracy condition can be extended to the multi-marginal transport problem, and is used in [54, Theorem 2.2] to bound the dimension of the support of an optimal transport. However, as noted by the author, the condition is no longer generic when $N > 2$, and indeed, some quite reasonable multi-marginal problems only have solutions of larger dimension [54, Remark 2.13]. On the other hand, we do expect that our results extend to $N > 2$ for particular costs like those in [34]. In any event, Theorem 3.8 separates such regularity issues from the convergence analysis, so that any available regularity result from optimal transport theory can be applied directly.

Remark 3.17. As mentioned in the Introduction, [18] previously obtained the constant $d/2$, for compactly supported marginals with uniformly bounded Lebesgue densities, and also showed its sharpness (cf. Section 4). Unlike in our result, the upper bound in [18] does not require nondegeneracy. Interestingly, Minty's trick is also used in [18], but it is employed in the proof of the sharpness rather than in the upper bound as in the present work. Above, we worked on the primal problem and used Minty's trick to estimate the dimension of optimal couplings. Whereas in [18], the authors work on the Kantorovich potentials of the dual problem to derive the upper bound, giving a quadratic control on the integrated difference between a λ-convex function and its first-order Taylor expansion.

4 Sharpness

In this section we show that the upper bounds obtained in the preceding section are sharp in certain cases. Throughout, we focus on $N = 2$ marginals and divergences given by $f(x) = x\log(x)$ and $f(x) = \frac{1}{\rho}(x^\rho - 1)$. Lower bounds for $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT}$ are naturally obtained from the dual problem of $\mathsf{OT}_{f,\varepsilon}$.

Lemma 4.1. *Let $\hat{h}_i \in L^1(\mu_i)$, $i = 1, 2$ be Kantorovich potentials for OT and $\hat{c}(x, y) := c(x, y) - \hat{h}_1(x) - \hat{h}_2(y)$ for $(x, y) \in X_1 \times X_2$. Let $f^*(y) := \sup_{x\geq 0}[xy - f(x)]$ for $y \in \mathbb{R}$ and $f^*_\varepsilon(y) := \varepsilon f^*(\frac{1}{\varepsilon}y)$. Then*

$$\begin{aligned}\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} &\geq \sup_{a\in\mathbb{R}} \left(a - \int f^*_\varepsilon\,(a - \hat{c})\; d(\mu_1 \otimes \mu_2)\right)\\ &\geq \sup_{a\in\mathbb{R}} \left(a - f^*_\varepsilon(a) \int \mathbf{1}_{a\geq\hat{c}}\, d(\mu_1 \otimes \mu_2)\right) - \varepsilon f^*(0).\end{aligned}$$

Proof. Recall (e.g., [29, 58]) the duality

$$\mathsf{OT}_{f,\varepsilon} = \sup_{h_1,h_2} \int h_1(x) + h_2(y) - f^*_\varepsilon\big(h_1(x) + h_2(y) - c(x, y)\big)\, \mu_1(dx)\mu_2(dy)$$

where the supremum ranges over $h_i \in L^1(\mu_i)$. As $\mathsf{OT} = \sum_{i=1}^2 \int \hat{h}_i\, d\mu_i$, choosing $h_1 = \hat{h}_1 + a$ and $h_2 = \hat{h}_2$ yields

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq \sup_{a\in\mathbb{R}} \left(a - \int f^*_\varepsilon(a - \hat{c})\, d(\mu_1 \otimes \mu_2)\right).$$

As f^*_ε is nondecreasing, $\hat{c} \geq 0$ and $f^*_\varepsilon(0) = \varepsilon f^*(0) \geq -\varepsilon f(1) = 0$, we also have $\int f^*_\varepsilon(a - \hat{c})\, d(\mu_1 \otimes \mu_2) \leq f^*_\varepsilon(a) \int \mathbf{1}_{a\geq\hat{c}}\, d(\mu_1 \otimes \mu_2) + f^*_\varepsilon(0)$, leading to the second inequality. □

Turning to the sharpness of the Lipschitz result (Theorem 3.3), it was observed in [18, Example 3.3] that the leading-order term $\varepsilon \log(1/\varepsilon)$ is sharp in the entropic case for the distance cost on $\mathbb{R}$. Part (i) below is a simple extension of that result to d dimensions equipped with the L^1-metric as cost, showing that the dependence on the dimension (or equivalently the quantization rate) is also sharp. For L^ρ regularization, we show in (ii) that the leading term has the sharp order and in particular the correct dimension dependence. Regarding the relation between dimension and quantization rate, recall from Remark 2.1 that $\alpha_2 = 1/d$ for absolutely continuous marginal $\mu_2 \in \mathcal{P}(\mathbb{R}^d)$.

Proposition 4.2 (Sharpness of Theorem 3.3)**.** *Let $X_1 = X_2 = \mathbb{R}^d$ with $\mu_1 = \mu_2$ the uniform distribution on $[0,1]^d$ and $c(x,y) = \sum_{i=1}^d |x_i - y_i|$.*

(i) Let $f(x) = x\log(x)$. Then for all $\varepsilon > 0$,

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq d\varepsilon \log(1/\varepsilon) - (2^d - 1)\varepsilon.$$

In particular, the leading term matches the bound in Theorem 3.3 (i).

(ii) Let $f(x) = \frac{1}{\rho}(x^\rho - 1)$ for some $\rho > 1$. Then

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq K\varepsilon^{\frac{1}{(\rho-1)d+1}} + O(\varepsilon)$$

for a constant $K > 0$. In particular, the leading term has the same exponent as the bound deduced from Theorem 3.3 (ii) in Example 3.4.

Proof. (i) Here $f^*(x) = e^x - 1$. Recalling the normalizing constant $\int_{\mathbb{R}} e^{\frac{|u-v|}{\varepsilon}}\, du = 2\varepsilon$ of the Laplace distribution,

$$\int e^{\frac{a-c}{\varepsilon}}\, d(\mu_1 \otimes \mu_2) = e^{\frac{a}{\varepsilon}} \prod_{i=1}^d \int_{[0,1]^2} e^{\frac{|x_i - y_i|}{\varepsilon}}\, dx_i dy_i \leq e^{a/\varepsilon}(2\varepsilon)^d,$$

and thus Lemma 4.1 (with $\hat{h}_1 = \hat{h}_2 = 0$) shows

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq \sup_a \left(a - 2^d \varepsilon^{d+1} e^{a/\varepsilon} + \varepsilon \right).$$

Choosing $a = d\varepsilon\log(1/\varepsilon)$, the right-hand side equals $d\varepsilon\log(1/\varepsilon) - (2^d - 1)\varepsilon$.
(ii) Here $f^*(y) = \frac{1}{q} y_+^q + \frac{1}{\rho}$ for $q := \frac{\rho}{\rho - 1}$, so that

$$f_\varepsilon^*(a) = \varepsilon f^*(a/\varepsilon) = \frac{a^q}{q\varepsilon^{q-1}} + \frac{\varepsilon}{\rho}, \quad a \geq 0.$$

The definition of c shows that $\mathbf{1}_{a\geq c} \leq \prod_{i=1}^d \mathbf{1}_{\{|x_i-y_i|\leq a\}}$ and thus

$$\int \mathbf{1}_{a\geq c}\, d(\mu_1\otimes\mu_2) \leq \prod_{i=1}^d \int_0^1\int_0^1 \mathbf{1}_{a\geq|x_i-y_i|}\, dx_i dy_i = (2a-a^2)^d \leq (2a)^d$$

for $a\in[0,1]$, with the last bound valid for $a\geq 0$. Lemma 4.1 thus yields

$$\begin{aligned}\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} &\geq \sup_{a\in\mathbb{R}_+}\Big(a - 2^d f_\varepsilon^*(a)a^d - \varepsilon f^*(0)\Big) \\ &= \sup_{a\in\mathbb{R}_+}\left(a - 2^d\frac{a^{d+q}}{q\varepsilon^{q-1}} - \frac{2^d\varepsilon a^d}{\rho} - \varepsilon\right).\end{aligned} \tag{4.1}$$

Setting $a := k\varepsilon^{\frac{1}{(\rho-1)d+1}}$, where $k>0$ is such that $K := (k - 2^d k^{q+d}/q) > 0$, we deduce $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq K\varepsilon^{\frac{1}{(p-1)d+1}} + O(\varepsilon)$ as claimed. □

Remark 4.3. We can similarly show the sharpness of Corollary 3.14 (ii) for quadratic cost. Namely, let $c(x,y) = |x-y|^2 = \sum_{i=1}^d |x_i-y_i|^2$. Going through the proof of Proposition 4.2, we now have $\mathbf{1}_{a\geq c} \leq \prod_{i=1}^d \mathbf{1}_{\{|x_i-y_i|\leq\sqrt{a}\}}$, and thus $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq K\varepsilon^{\frac{1}{(p-1)d/2+1}} + O(\varepsilon)$. A more general (if much more involved) argument for a general class of marginals is given below.

Indeed, we can establish the sharpness of Theorem 3.8 for a general class of marginals and costs. For the entropic case, it is well known that the leading term $\frac{d}{2}\varepsilon\log\left(\frac{1}{\varepsilon}\right)$ is sharp for quadratic cost $c(x,y) = |x-y|^2$ on $\mathbb{R}^d\times\mathbb{R}^d$ when the marginals are sufficiently regular [22, 24, 53]. Very recently, [18] showed that this term is sharp for the broad class of nondegenerate (as defined before Lemma 3.15) costs and regular marginals; their result is stated in (i) below for completeness. The core of the proof in [18] is a quadratic detachment estimate for the Kantorovich potentials. In (ii), we apply their technique to divergences $f(x) = \frac{1}{\rho}(x^\rho - 1)$ to show sharpness of the leading order in Theorem 3.8 (ii).

Proposition 4.4 (Sharpness of Theorem 3.8)**.** *For $i=1,2$, let $X_i\subset\mathbb{R}^d$ be convex and compact and let $\mu_i\in\mathcal{P}(X_i)$ have bounded Lebesgue density. Let $c\in\mathcal{C}^2(X)$ be nondegenerate.*

(i) Let $f(x) = x\log(x)$. Then

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq \frac{d}{2}\varepsilon\log(1/\varepsilon) + O(\varepsilon).$$

In particular, the leading term matches the bound in Theorem 3.8 (i).

(ii) Let $f(x) = \frac{1}{\rho}(x^\rho - 1)$ for some $\rho > 1$. Then

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \geq K\varepsilon^{\frac{1}{(\rho-1)d/2+1}} + O(\varepsilon)$$

for a constant $K > 0$. In particular, the leading term has the same exponent as the bound deduced from Theorem 3.8 (ii) in Example 3.11.

Proof. See [18, Proposition 4.4] for (i). To show (ii), we argue that there exist constants $C_0, C > 0$ such that

$$\mathsf{OT}_{f,\varepsilon} \geq \mathsf{OT} + \sup_{a \leq C_0} \left(a - C f_\varepsilon^*(a) a^{d/2} - \max\{0, f_\varepsilon^*(0)\} \right). \tag{4.2}$$

This bound is similar to (4.1) but with different constants, and implies the claim along the same lines. To show (4.2), we will apply Lemma 4.1 with optimal potentials $(\hat{h}_1, \hat{h}_2)$. The latter can be chosen to be continuous, so that $\hat{c}$ is also continuous. The main difficulty is to bound $\int \mathbf{1}_{a \geq \hat{c}} \, d(\mu_1 \otimes \mu_2)$. Following the proof of [18, Proposition 4.4], we find a finite open cover $A = \cup_{i=1}^n A_i$ of the compact set $\{\hat{c} = 0\} \cap (X_1 \times X_2)$ satisfying the following:

(a) On the compact $B := (X_1 \times X_2) \backslash A$ we have $\hat{c} > C_0$ for some $C_0 > 0$.

(b) There exist $r, C_1 > 0$ such that for all $i \in \{1, \ldots, n\}$, for some $r_v \in \mathbb{R}^d$ depending only on $v \in \mathbb{R}^d$,

$$\int_{A_i} \mathbf{1}_{a \geq \hat{c}} \, d(\mu_1 \otimes \mu_2) \leq C_1 \int_{B_r} \int_{B_r} \mathbf{1}_{a \geq |u - r_v|^2/4} \, du dv,$$

where $B_r \subset \mathbb{R}^d$ is the ball of radius $r > 0$ around the origin.

Bounding the inner integral in (b) according to

$$\int_{B_r} \mathbf{1}_{a \geq |u - r_v|^2/4} \, du \leq \int_{\mathbb{R}^d} \mathbf{1}_{a \geq |u|^2/4} \, du \leq \prod_{i=1}^d \int_{\mathbb{R}} \mathbf{1}_{|u_i| \leq 2\sqrt{a}} \, du \leq 4^d a^{d/2},$$

we obtain

$$\int_{A \cap (X_1 \times X_2)} \mathbf{1}_{a \geq \hat{c}} \, d(\mu_1 \otimes \mu_2) \leq C a^{d/2}$$

for a constant $C > 0$. In view of (a), this shows

$$\int_{X_1 \times X_2} \mathbf{1}_{a \geq \hat{c}} \, d(\mu_1 \otimes \mu_2) \leq C a^{d/2} \quad \text{for} \quad a \leq C_0 \tag{4.3}$$

and now (4.2) follows by Lemma 4.1. □

A Appendix

The following is well known in the entropic case [49, Section 5]. For completeness, we provide an extension to the f-divergences under consideration.

Proposition A.1. *We have $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} = O(\varepsilon)$ if and only if there exists an optimal transport plan π^* for OT with $D_f(\pi^*, P) < \infty$.*

Proof. If there exists an optimal transport plan π^* with finite divergence, clearly $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \varepsilon D_f(\pi^*, P) = O(\varepsilon)$. Conversely, let π_ε be an optimizer of $\mathsf{OT}_{f,\varepsilon}$. If $\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} = O(\varepsilon)$, it follows that $\sup_{\varepsilon\in(0,1]} D_f(\pi_\varepsilon, P) < \infty$. As f has superlinear growth, the densities $d\pi_\varepsilon/dP$ are then uniformly integrable; in particular, there exists a weak*-convergent sequence $d\pi_{\varepsilon_n}/dP$, meaning that (π_{ε_n}) converge set-wise. The limit π_0 is again a coupling. We have $\int c\, d\pi_0 \leq \liminf \mathsf{OT}_{f,\varepsilon_n} = \mathsf{OT}$ by a generalized Fatou's lemma [57, p. 231] and the growth condition on c, showing that π_0 is an optimal transport. The same Fatou's lemma shows $D_f(\pi_0, P) \leq \liminf D_f(\pi_{\varepsilon_n}, P) < \infty$, completing the proof. □

The following extension of Theorem 3.3 was prompted by a question of by G. Carlier; see also the similar [18, Remark 3.2].

Remark A.2 (Extension of Theorem 3.3 beyond Lipschitz)**.** Fix $p = 1$ and replace $(\mathrm{A}_{L,C})$ by

$$\left| \int c\, d(\pi - \tilde{\pi}) \right| \leq \omega(W_1(\pi, \tilde{\pi})), \tag{A.1}$$

where $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing and concave modulus of continuity. To motivate this, note that if the function c itself has modulus of continuity ω, then choosing $\theta \in \Pi(\pi, \tilde{\pi})$ attaining $W_1(\pi, \tilde{\pi})$ yields

$$\begin{aligned} \left| \int c\, d(\pi - \tilde{\pi}) \right| &\leq \int |c(x) - c(y)|\, \theta(dx, dy) \\ &\leq \int \omega(d_{X,1}(x, y))\, \theta(dx, dy) \leq \omega(W_1(\pi, \tilde{\pi})) \end{aligned}$$

by Jensen's inequality. Going through the proof of Theorem 3.3 with (A.1), we obtain instead of (3.1) that

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq 2\omega\left(C \sum_{i=2}^{N} n_i^{-\alpha_i} \right) + \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \dots \mu_N^{n_N})$$

and can then optimize the choice of n_i. For instance, in the entropic case, we would take $S_\varepsilon = \omega^{-1}(1/\varepsilon)$; then the first term is again of order ε while the divergence term is of order $\varepsilon \log(\omega^{-1}(1/\varepsilon))$. For $N = 2$ and $c(x, y) = d_{X,1}(x, y)^r$ with $0 < r < 1$, we end up with

$$\mathsf{OT}_{f,\varepsilon} - \mathsf{OT} \leq \frac{1}{r\alpha_2} \varepsilon \log\left(\frac{1}{\varepsilon}\right) + K\varepsilon.$$

It is worth noting the formal similarity with Theorem 3.8 (i) which would correspond to $r = 2$.

References

[1] S. Adams, N. Dirr, M. A. Peletier, and J. Zimmer. From a large-deviations principle to the Wasserstein gradient flow: a new micro-macro passage. *Comm. Math. Phys.*, 307(3):791–815, 2011.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[3] G. Alberti and L. Ambrosio. A geometrical approach to monotone functions in $\mathbf{R}^n$. *Math. Z.*, 230(2):259–316, 1999.

[4] J. M. Altschuler and E. Boix-Adserà. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Math. Program.*, 199(1-2, Ser. A):1107–1178, 2023.

[5] J. M. Altschuler, J. Niles-Weed, and A. J. Stromme. Asymptotics for semidiscrete entropic optimal transport. *SIAM J. Math. Anal.*, 54(2):1718–1741, 2022.

[6] J.-D. Benamou, G. Carlier, and L. Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numer. Math.*, 142(1):33–54, 2019.

[7] O. Bencheikh and B. Jourdain. Approximation rate in Wasserstein distance of probability measures on the real line by deterministic empirical measures. *J. Approx. Theory*, 274:Paper No. 105684, 27, 2022.

[8] R. J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge-Ampère equations. *Numer. Math.*, 145(4):771–836, 2020.

[9] E. Bernton, P. Ghosal, and M. Nutz. Entropic optimal transport: Geometry and large deviations. *Duke Math. J.*, 171(16):3363–3400, 2022.

[10] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *Preprint arXiv:1810.07717v1*, 2018.

[11] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, 2018.

[12] L. A. Caffarelli. The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, 5(1):99–104, 1992.

[13] L. A. Caffarelli. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2)*, 144(3):453–496, 1996.

[14] G. Carlier. On the linear convergence of the multi-marginal Sinkhorn algorithm. *SIAM J. Optim.*, 32(2):786–794, 2022.

[15] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

[16] G. Carlier, K. Eichinger, and A. Kroshnin. Entropic-Wasserstein barycenters: PDE characterization, regularity, and CLT. *SIAM J. Math. Anal.*, 53(5):5880–5914, 2021.

[17] G. Carlier and M. Laborde. A differential approach to the multi-marginal Schrödinger system. *SIAM J. Math. Anal.*, 52(1):709–717, 2020.

[18] G. Carlier, P. Pegon, and L. Tamanini. Convergence rate of general entropic optimal transport costs. *Calc. Var. Partial Differential Equations*, 62(4):Paper No. 116, 28, 2023.

[19] Y. Chen, T. T. Georgiou, and M. Pavon. On the relation between optimal transport and Schrödinger bridges: a stochastic control viewpoint. *J. Optim. Theory Appl.*, 169(2):671–691, 2016.

[20] J. Chevallier. Uniform decomposition of probability measures: quantization, clustering and rate of convergence. *J. Appl. Probab.*, 55(4):1037–1045, 2018.

[21] A. Chiarini, G. Conforti, G. Greco, and L. Tamanini. Gradient estimates for the Schrödinger potentials: convergence to the Brenier map and quantitative stability. *Preprint arXiv:2207.14262v1*, 2022.

[22] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269, 2020.

[23] R. Cominetti and J. San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Math. Programming*, 67(2, Ser. A):169–187, 1994.

[24] G. Conforti and L. Tamanini. A formula for the time derivative of the entropic cost and applications. *J. Funct. Anal.*, 280(11):108964, 2021.

[25] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. 2013.

[26] S. Di Marino and A. Gerolin. Optimal transport losses and Sinkhorn algorithm with general convex regularization. *Preprint arXiv:2007.00976v1*, 2020.

[27] M. H. Duong, V. Laschos, and M. Renger. Wasserstein gradient flows from large deviations of many-particle limits. *ESAIM Control Optim. Calc. Var.*, 19(4):1166–1188, 2013.

[28] S. Eckstein and M. Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM J. Math. Anal.*, 54(6):5922–5948, 2022.

[29] S. Eckstein and G. Pammer. Computational methods for adapted optimal transport. *Preprint arXiv:2203.05005v1*, 2022.

[30] M. Erbar, J. Maas, and D. R. M. Renger. From large deviations to Wasserstein gradient flows in multiple dimensions. *Electron. Commun. Probab.*, 20(89), 2015.

[31] M. Essid and J. Solomon. Quadratically regularized optimal transport on

graphs. *SIAM J. Sci. Comput.*, 40(4):A1961–A1986, 2018.

[32] N. Fournier. Convergence in expected Wasserstein distance of the empirical measure: non-asymptotic explicit bounds in $\mathbf{R}^n$. *Preprint arXiv:2209.00923v1*, 2022.

[33] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

[34] W. Gangbo and A. Święch. Optimal maps for the multidimensional Monge-Kantorovich problem. *Comm. Pure Appl. Math.*, 51(1):23–45, 1998.

[35] P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and Schrödinger bridges. *J. Funct. Anal.*, 283(9):Paper No. 109622, 2022.

[36] N. Gigli and L. Tamanini. Second order differentiation formula on $\mathrm{RCD}^*(K, N)$ spaces. *J. Eur. Math. Soc. (JEMS)*, 23(5):1727–1795, 2021.

[37] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, Berlin, 2000.

[38] C. Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *J. Funct. Anal.*, 262(4):1879–1920, 2012.

[39] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34(4):1533–1574, 2014.

[40] T. Lin, N. Ho, and M. Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 3982–3991, 2019.

[41] G. Loeper. On the regularity of solutions of optimal transportation problems. *Acta Math.*, 202(2):241–283, 2009.

[42] D. A. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. *Appl. Math. Optim.*, 83(3):1919–1949, 2021.

[43] X.-N. Ma, N. S. Trudinger, and X.-J. Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Ration. Mech. Anal.*, 177(2):151–183, 2005.

[44] L. Martins Bianco. Stochastic approximation in optimal transport. *M1 Internship Report, Université Paris-Saclay*, 2022.

[45] R. J. McCann, B. Pass, and M. Warren. Rectifiability of optimal transportation plans. *Canad. J. Math.*, 64(4):924–934, 2012.

[46] T. Mikami. Optimal control for absolutely continuous stochastic processes and the mass transportation problem. *Electron. Comm. Probab.*, 7:199–213, 2002.

[47] T. Mikami. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probab. Theory Related Fields*, 129(2):245–260, 2004.

[48] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29:341–346, 1962.

[49] M. Nutz. *Introduction to Entropic Optimal Transport.* Lecture notes, Columbia University, 2021. `https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf`.

[50] M. Nutz and J. Wiesel. Entropic optimal transport: convergence of potentials. *Probab. Theory Related Fields*, 184(1-2):401–424, 2022.

[51] M. Nutz and J. Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *Ann. Probab.*, 51(2):699–722, 2023.

[52] G. Pagès. *Numerical probability*. Universitext. Springer, Cham, 2018.

[53] S. Pal. On the difference between entropic cost and the optimal transport cost. *Preprint arXiv:1905.12206v1*, 2019.

[54] B. Pass. Multi-marginal optimal transport: theory and applications. *ESAIM Math. Model. Numer. Anal.*, 49(6):1771–1790, 2015.

[55] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[56] A.-A. Pooladian and J. Niles-Weed. Entropic estimation of optimal transport maps. *Preprint arXiv:2109.12004v1*, 2021.

[57] H. L. Royden. *Real analysis*. Macmillan, New York, second edition, 1968.

[58] D. Terjék and D. González-Sánchez. Optimal transport with f-divergence regularization and generalized Sinkhorn algorithm. volume 151 of *Proceedings of Machine Learning Research*, pages 5135–5165, 2022.

[59] C. Villani. *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

[60] J. Weed. An explicit analysis of the entropic penalty in linear programming. volume 75 of *Proceedings of Machine Learning Research*, pages 1841–1855, 2018.

[61] 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.

[62] C. Xu and A. Berger. Best finite constrained approximations of one-dimensional probabilities. *J. Approx. Theory*, 244:1–36, 2019.