Convergence Rates for Regularized Optimal Transport via Quantization

Stephan Eckstein† Marcel Nutz‡

August 31, 2022

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 all 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

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

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 $\text{OT}$ without regularization. We are interested in the speed of convergence $\text{OT}_{f,\varepsilon} \to \text{OT}$ as the regularization parameter $\varepsilon$ tends to zero, especially its dependence on the marginals $\mu_i$ and the divergence $D_f$. 

†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, 39, 54] 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 $\text{OT}_{f,\varepsilon}$ is often called the entropic optimal transport problem (e.g., [48, 54]). Other divergences, especially $L^p$ regularization, are being used in applications where sparse optimizers are desired or weak penalization (small $\varepsilon$) causes numerical instabilities with entropic regularization [11, 26, 31, 41, 57]. For multi-marginal transport and the related Wasserstein barycenters, see for instance [2, 6, 14, 16, 17]. Literature more specific to the convergence $\text{OT}_{f,\varepsilon} \to \text{OT}$ is discussed below.

In this paper, we propose a novel methodology to estimate $\text{OT}_{f,\varepsilon} - \text{OT}$ 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, 52].

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

$$\text{OT}_{f,\varepsilon} - \text{OT} \leq \beta \varepsilon \log \left( \frac{1}{\varepsilon} \right) + K \varepsilon \quad \text{for entropic regularization},$$

$$\text{OT}_{f,\varepsilon} - \text{OT} \leq K \varepsilon^{\frac{1}{(p-1)\beta+1}} \quad \text{for } L^p \text{ regularization},$$

where $\beta$ reflects a certain quantization dimension. In our first result (Theorem 3.3), $\beta$ encodes the optimal quantization rate for one of the marginals—if $\mu_i$ are measures on $\mathbb{R}^{d_i}$, this leads to $\beta \leq d_1 \wedge d_2$. 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 $p \geq 1$ 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 $\text{OT}_{f, \varepsilon}$. Specifically, a rate is found by choosing the number of points for the quantization of the marginals relative to the regularization parameter $\varepsilon$ 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 $\beta$ 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 $\text{OT}$) 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 3.12). In its proof, Minty’s trick [47] 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 [59] 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 $\text{OT}_{f, \varepsilon} - \text{OT} = O(\varepsilon)$ 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 [43] where $\chi^2$ 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 $\text{OT}_{f, \varepsilon}$ and $\text{OT}$ goes back to [45, 46] in the Schrödinger bridge problem (which is closely related to entropic optimal transport with quadratic cost; cf. [38]). Gamma-convergence was shown in [37]; 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 [52] 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 $\text{OT}_{f,\varepsilon} - \text{OT} \leq \frac{d \varepsilon \log(1/\varepsilon)}{\varepsilon} + O(\varepsilon)$ 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. 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 $\text{OT}_{f,\varepsilon} - \text{OT} \leq \frac{d \varepsilon}{2} \log\left(\frac{1}{\varepsilon}\right) + O(\varepsilon)$. The proof is deep and based on the fine regularity of the Kantorovich potential; namely, a quadratic bound on the integrated difference between a $\lambda$-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 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 $\pi^*$, so that it needs to be applied together with a regularity result for $\pi^*$. 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 $\text{OT}_{f, \varepsilon}$, two related question are the convergence of the optimal couplings and optimal dual potentials. See [9, 15, 37, 38] and [8, 21, 35, 49, 55], respectively, and the references therein. As seen in [9, 21], the convergence is also related to the stability of $\text{OT}_{f, \varepsilon}$ wrt. the marginals [17, 28, 34, 50].

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 $\mu$ 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 $(X_i, d_{X_i})$, $i = 1, \ldots, N$ 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.,

$$|c(x)| \leq C(1 + d_{X,p}(x, \hat{x})^p)$$
for some $C \geq 0$ and $\hat{x} \in X$. The optimal transport problem is

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

where $\Pi(\mu_1, \ldots, \mu_N) \subset \mathcal{P}_p(X)$. The growth of $c$ ensures that $\text{OT}$ 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 $\mu, \nu$ 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

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

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

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

for constants $C \geq 0$ and $\alpha > 0$. The takeaway of the following is that if the support of $\mu$ 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 $\text{quant}_p(C, \alpha)$ holds with $\alpha = 1/d$ for some $C \geq 0$. More precisely, [36, Theorem 6.2] shows that the exact asymptotic constant

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

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 $\mu$. In particular, $C_\alpha > 0$ as soon as $\mu$ 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 $\text{quant}_p(C, \alpha)$ is given in [36, Corollary 6.7]; its proof yields an explicit constant valid for all $n \geq 1$ depending only on $p, \delta, d$ and $\int |x|^{p+\delta} \mu(dx)$.\footnote{The result in [36, 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.}

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

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

for constants $C \geq 0$ and $\alpha > 0$. This condition clearly implies $\text{quant}_p(C, \alpha)$; 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 $d \geq 2p$, then $\text{quant}^{em}_p(C, \alpha)$ holds with $\alpha = 1/d$ and a constant $C$ depending only on $d, p, \delta$ and the $(2p + \delta)$-moment of $\mu$. Rates for other regimes are also obtained and moreover, these even hold when $\mu^n$ are i.i.d. samples of $\mu$. 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 [60].

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 case $d = 1$ has been studied in detail [7, 61]. Here the optimal rate is $\alpha = 1$ if $\mu$ 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 \( \mu_i \) supported on \( n \) points has relative entropy \( D_f(\pi, \mu_1 \otimes \mu_2) \leq \log n \). 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, \( \mathcal{P}^{n, \text{em}}(X_i) \) the empirical measures on \( n \) points, and \( P = \mu_1 \otimes \cdots \otimes \mu_N \).

**Lemma 2.3** (Divergence bounds). Let \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \) and define \( \varphi \) by \( f(x) = x \varphi(x) \).

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

(ii) If \( \mu_i \in \mathcal{P}^{n_i, \text{em}}(X_i) \) for \( i = 2, \ldots, N \), then \( D_f(\pi, P) \leq \varphi(\prod_{i=2}^N n_i) \).

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

**Proof.** Denote by \( \pi_{2:N} \) the marginal of \( \pi \) on \( X_2 \times \cdots \times X_N \). In particular, \( P_{2:N} = \mu_2 \otimes \cdots \otimes \mu_N \). Let \( \sigma \) be the counting measure on the (finite) support of \( P_{2:N} \). Disintegrating \( \pi = \mu_1 \otimes K \), we then have \( \frac{d\pi}{dP} = \frac{dK}{dP_{2:N}} \leq \frac{d\sigma}{dP_{2:N}} \), hence

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

In the case (i) where \( N = 2 \), Jensen’s inequality yields

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

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

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 \( \mu_N \) as second marginal,

\[
D_f(\pi, P) \leq \log(n_N) + D_f(\pi_{1:N-1}, P_{1:N-1}).
\]

Iterating this argument yields \( D_f(\pi, P) \leq \sum_{i=2}^N \log(n_i) \), which was the claim. \( \square \)
2.4 Shadows

Given \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \), the shadow \( \tilde{\pi} \) of \( \pi \) on another vector \((\tilde{\mu}_1, \ldots, \tilde{\mu}_N)\) of marginals is a particular \( W_p \)-projection of \( \pi \) onto \( \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N) \) 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 \( \mu_1 \) using a \( W_p \)-optimal transport, then follow the transport \( \pi \) moving \( \mu_1 \) into \( \mu_2 \), and finally move \( \mu_2 \) to \( \tilde{\mu}_2 \) using a \( W_p \)-optimal transport. The general definition follows.

**Definition 2.4** ([28]). Let \( p \in [1, \infty) \) and \( \mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i), i = 1, \ldots, N \). 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 \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \), its shadow \( \tilde{\pi} \) on \((\tilde{\mu}_1, \ldots, \tilde{\mu}_N)\) 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 \( \mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i), i = 1, \ldots, N \). Given \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \), its shadow \( \tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N) \) satisfies

\[
W_p(\pi, \tilde{\pi})^p = \sum_{i=1}^{N} W_p(\mu_i, \tilde{\mu}_i)^p,
\]

\[
D_f(\tilde{\pi}, \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N) \leq D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N).
\]

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 re-proving the (known) convergence \( \text{OT}_{f, \varepsilon} \to \text{OT} \) in our general setting.

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

**Proof.** Using tightness of \( \{\mu_i\} \), we can construct measures \( \mu_i^n \) supported on \( n \) points with \( W_p(\mu_i^n, \mu_i) \to 0 \) for \( i = 1, \ldots, N \). Let \( \pi^* \in \Pi(\mu_1, \ldots, \mu_N) \) be an optimizer of \( \text{OT} \). We introduce another coupling \( \pi^n \in \Pi(\mu_1, \ldots, \mu_N) \) as follows: first, let \( \tilde{\pi} \) be the shadow of \( \pi^* \) onto \((\mu_1^n, \mu_2^n, \ldots, \mu_N^n)\); then, define \( \pi^n \) as the shadow of \( \tilde{\pi} \) onto \((\mu_1, \ldots, \mu_N)\). Using the triangle inequality and
Lemma 2.5, this implies
\[ W_p(\pi^n, \pi^*) \leq W_p(\tilde{\pi}, \pi^*) \leq 2 \left( \sum_{i=1}^N W_p(\mu_i^n, \mu_i) \right)^{1/p} \to 0. \]

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
\[ D_f(\pi^n, P) \leq D_f(\tilde{\pi}, \mu_1^n \otimes \mu_2^n \otimes \ldots \otimes \mu_N^n) < \infty, \]
where the finiteness is trivial by discreteness of \( \mu_i^n \). Given \( \delta > 0 \), choose \( n \) such that \( \int c d\pi^n - \int c d\pi^* \leq \delta \), and then \( \varepsilon_0 > 0 \) such that \( \varepsilon_0 D_f(\pi^n, P) \leq \delta \).

As \( \pi^n \) is an admissible coupling for \( \text{OT}_{f,\varepsilon} \), we have shown \( \text{OT}_{f,\varepsilon} - \text{OT} \leq 2\delta \) for all \( \varepsilon \leq \varepsilon_0 \).

### 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 \( \mu_i \in \mathcal{P}_p(X_i), i = 1, \ldots, N \). Given constants \( L, C \geq 0 \), we say that \( c \) satisfies \((A_{L,C})\) if for all \( \tilde{\pi} \) with \( W_p(\tilde{\mu}_i, \mu_i) \leq C, i = 2, \ldots, N \), we have
\[ \left| \int c d(\pi - \tilde{\pi}) \right| \leq L W_p(\pi, \tilde{\pi}) \quad (A_{L,C}) \]
for all \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \) and \( \tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N) \).

Clearly \((A_{L,C})\) 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 \( \mu_i \) 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 \( i = 1, \ldots, N \). Assume that \( \mu_i \) satisfies \( \text{quant}_p(C, \alpha_i) \) for \( i = 2, \ldots, N \) and that \( c \) satisfies \((A_{L,C})\), for some \( \alpha_2, \ldots, \alpha_N \in (0, 1] \) and \( L, C \geq 0 \).
(i) Let $f(x) = x \log(x)$. Then for all $\varepsilon \in (0, 1]$,

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

(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 $\tilde{f}_{\text{inv}}$ and $\varphi$ is nondecreasing on $[y_0, \infty)$. Suppose also that either $N = 2$ and $\varphi$ is concave and nondecreasing, or the $\mu_i$ satisfy $\text{quant}_{p}^m(C, \alpha_i)$ instead of $\text{quant}_{p}(C, \alpha_i)$. Set $S_\varepsilon = \tilde{f}_{\text{inv}}(1/\varepsilon)$, 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 $S_\varepsilon \geq y_0^{1/\beta} + 1$,

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

While the quantity $S_\varepsilon$ 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 $\tilde{g}_{\text{inv}}$. Then $\tilde{g}_{\text{inv}} \leq \tilde{f}_{\text{inv}}$ and hence $1/S_\varepsilon \leq 1/\tilde{g}_{\text{inv}}(1/\varepsilon)$, so that Theorem 3.3 (ii) implies

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

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

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

**Remark 3.5** (On $\mu_1$). In Theorem 3.3, nothing is assumed about the quantization of $\mu_1$. In an application, one would thus label $\mu_1$ 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 \( \pi^* \in \Pi(\mu_1, \ldots, \mu_N) \) be an optimizer of \( \text{OT} \). By our assumption, there exist empirical quantizations \( \mu_i^{n_i} \) for the marginals \( i = 2, \ldots, N \) such that \( W_p(\mu_i^{n_i}, \mu_i) \leq Cn_i^{-\alpha_i} \). We introduce a coupling \( \pi \in \Pi(\mu_1, \ldots, \mu_N) \) (depending on \( n_2, \ldots, n_N \)) as a double shadow: first, let \( \tilde{\pi} \) be the shadow of \( \pi^* \) onto \( (\mu_1, \mu_2^{n_2}, \ldots, \mu_N^{n_N}) \); then, define \( \pi \) as the shadow of \( \tilde{\pi} \) onto \( (\mu_1, \ldots, \mu_N) \). Using the triangle inequality and Lemma 2.5,

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

Combining this with our assumption \((A_{L,C})\), we deduce

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

On the other hand, Lemma 2.5 again yields

\[
D_f(\pi, P) \leq D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \ldots \otimes \mu_N^{n_N}).
\]

As \( \pi \) is an admissible coupling for \( \text{OT}_{f,\varepsilon} \), we have proved

\[
\text{OT}_{f,\varepsilon} - \text{OT} \leq 2LC \left( \sum_{i=2}^{N} n_i^{-\alpha_ip} \right)^{1/p} + \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \ldots \otimes \mu_N^{n_N}) \quad (3.1)
\]

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 \( \varepsilon \) 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

\[
\varrho(S) := \left( \frac{1}{N - 1} \sum_{i=2}^{N} \left[ \frac{Sp}{|S^{1/\alpha_i}|^{\alpha_i p}} \right]^{1/p} \right)
\]

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

\[
\left( \sum_{i=2}^{N} |S^{1/\alpha_i}|^{-\alpha_i p} \right)^{1/p} = \varrho(S)(N - 1)^{1/p}/S \leq \frac{2(N - 1)^{1/p}}{S}.
\]

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

\[
\left( \sum_{i=2}^{N} n_i^{-\alpha_ip} \right)^{1/p} = \varrho(S_\varepsilon)(N - 1)^{1/p}/S_\varepsilon \leq 2(N - 1)^{1/p}\varepsilon,
\]
and Lemma 2.3(iii) bounds the divergence term by
\[ \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \ldots \mu_N^{n_N}) \leq \varepsilon \sum_{i=2}^{N} \log(n_i) \leq \varepsilon \sum_{i=2}^{N} \frac{1}{\alpha_i} \log \left( \frac{1}{\varepsilon} \right). \]

In view of (3.1), the claim follows.

(ii) Set \( n_i = [S_\varepsilon^{1/\alpha_i}] \) for \( i = 2, \ldots, N \), where \( S_\varepsilon \) was defined in the theorem. Similarly as in (i),
\[ \left( \sum_{i=2}^{N} n_i^{-\alpha_i p} \right)^{1/p} \leq \frac{g(S_\varepsilon)(N-1)^{1/p}}{S_\varepsilon} \leq 2(N-1)^{1/p} \frac{1}{S_\varepsilon}. \]

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 \text{quant}_p^e(C, \alpha_i), Lemma 2.3(ii) and monotonicity of \( \varphi \) on \([y_0, \infty)\) yield
\[ \varepsilon D_f(\tilde{\pi}, \mu_1 \otimes \mu_2^{n_2} \otimes \ldots \mu_N^{n_N}) \leq \varepsilon \varphi \left( \prod_{i=2}^{N} n_i \right) \leq \varepsilon \varphi(S_\varepsilon^\beta) = \varepsilon \frac{\tilde{f}(S_\varepsilon)}{S_\varepsilon} = \varepsilon \frac{\tilde{f}_{\text{inv}}(\frac{1}{\varepsilon})}{S_\varepsilon} = \frac{1}{S_\varepsilon} \]
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).

\[ \square \]

**Remark 3.6 (On the constant).** The constant 4 in Theorem 3.3(i),(ii) can be replaced by \( 2g(1/\varepsilon) \) and \( 2g(S_\varepsilon) \), respectively, where \( g(\cdot) \) is defined in (3.2) and satisfies \( 1 \leq g(\cdot) \leq 2 \). As \( g(S) = 1 + o(1/S) \), 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 \( \varepsilon \) and \( \varepsilon \log(1/\varepsilon) \). However, \( S_\varepsilon = \tilde{f}_{\text{inv}}(\frac{1}{\varepsilon}) \) 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, \((A_L, C)\) 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 \( \alpha \) 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 \( i = 1, \ldots, N \). Assume that \( c \in C^2(X) \) has bounded second derivative:

\[
 w^\top c''(x)w \leq B|w|^2 \quad \text{for all } x, w \in X, \text{ for some } B \geq 0,
\]

and that \( \text{OT} \) admits an optimal transport \( \pi^* \) satisfying \( \text{quant}_p(C, \alpha) \) for some \( \alpha \in (0, 1] \) and \( C > 0 \).

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

\[
 \text{OT}_{f, \varepsilon} - \text{OT} \leq \frac{N-1}{2\alpha} \varepsilon \log \left( \frac{1}{\varepsilon} \right) + 8BC\varepsilon.
\]

(ii) Let \( N = 2 \), \( f(x) = x\varphi(x) \) with \( \varphi \) nondecreasing and concave, let \( \beta = \frac{1}{2\alpha} \) and \( \bar{f}(x) = x\varphi(x^\beta) \). Assume that for some \( x_0 \geq 0 \), \( \bar{f} \) is strictly increasing on \( [x_0, \infty) \) with inverse \( \bar{f}^{-1} \). Set \( S_\varepsilon = \bar{f}^{-1}(\frac{1}{\varepsilon}) \), 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 \( S_\varepsilon \geq 1 \),

\[
 \text{OT}_{f, \varepsilon} - \text{OT} \leq \frac{8BC + 1}{S_\varepsilon}.
\]

**Proof.** For \( n \geq 1 \), let \( \tilde{\pi} \in \mathcal{P}(X) \) be the optimal \( W_2 \)-quantizer of \( \pi^* \) on \( n \) points and let \( \theta \in \Pi(\tilde{\pi}, \pi^*) \) be a coupling attaining \( W_2(\tilde{\pi}, \pi^*) \); that is,

\[
 \int |x - y|^2 \theta(dx, dy) = W_2(\tilde{\pi}, \pi^*)^2.
\]
It is known (e.g., [51, Proposition 5.1, p.139]) that \( \theta \) is a martingale; i.e., the kernel \( \kappa \) in its disintegration \( \theta = \tilde{\pi} \otimes \kappa \) satisfies \( \int y \kappa(x, dy) = x \) for \( \tilde{\pi} \)-almost all \( x \). This implies that \( \int h(x) \cdot (y - x) \theta(dx, dy) = 0 \) for any measurable function \( h : X \to \mathbb{R}^{d_1 + \ldots + d_N} \) 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

\[
\left| \int c \, d\pi^* - \int c \, d\tilde{\pi} \right| = \left| \int (c(y) - c(x)) \theta(dx, dy) \right| \\
\leq B \int |x - y|^2 \theta(dx, dy) = BW_2^2(\tilde{\pi}, \pi^*)^2. \tag{3.5}
\]

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

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

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 \( \theta_i \) instead of \( W_2 \)-optimal transports between \( \mu^n_i \) and \( \mu_i \). Namely, decomposing \( \theta_i = \mu^n_i \otimes K_i \) 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 \( \gamma \). Probabilistically speaking, this means that we take the (possibly dependent) components of the vector martingale \( \theta \) and combine their laws into a new vector martingale \( \gamma \) 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 \( \theta_i \). Repeating the argument for (3.5) with \( \gamma \) instead of \( \theta \), inserting the definition of \( \gamma \) and using (3.6), we conclude that

\[
\left| \int c \, d\pi - \int c \, d\tilde{\pi} \right| \leq B \int |x - y|^2 \gamma(dx, dy) \\
= B \sum_{i=1}^N \int |x_i - y_i|^2 \theta_i(dx_i, dy_i) = BW_2^2(\tilde{\pi}, \pi^*)^2.
\]
In view of (3.5), the triangle inequality and the assumption on \( \pi^* \) then yield
\[
\int c \, d\pi - \int c \, d\pi^* \leq 2BW_2(\tilde{\pi}, \pi^*)^2 \leq 2BCn^{-2\alpha}. \tag{3.7}
\]
On the other hand, by the data processing inequality (e.g., [48, Lemma 1.6]), the construction of \( \pi \) implies
\[
D_f(\pi, P) \leq D_f(\tilde{\pi}, \mu^n_1 \otimes \cdots \otimes \mu^n_N).
\]
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 \( g(S) \) for \( S \in [1, \infty) \) as
\[
g(S) := \left( \frac{S^{\frac{1}{2\alpha}}}{\lfloor S^{\frac{1}{2\alpha}} \rfloor} \right)^{2\alpha} \tag{3.8}
\]
so that \( 1 \leq g(S) \leq 2^{2\alpha} \leq 4 \) and \( \lim_{S \to \infty} g(S) = 1 \). In particular,
\[
[S^{\frac{1}{2\alpha}}]^{-2\alpha} = g(S)S^{-1} \leq 4S^{-1}. \tag{3.9}
\]
(i) Let \( n = \lfloor \varepsilon^{-\frac{1}{2\alpha}} \rfloor \). Then (3.7) and (3.9) for \( S = S_\varepsilon = 1/\varepsilon \) imply
\[
\int c \, d\pi - \int c \, d\pi^* \leq 2BCg(S_\varepsilon)S_\varepsilon^{-1} \leq 8BCg(S_\varepsilon)\varepsilon
\]
while Lemma 2.3(iii) yields \( D_f(\tilde{\pi}, \mu^n_1 \otimes \cdots \otimes \mu^n_N) \leq (N-1)\log(n) \), completing the proof of (i).

(ii) Here we define \( n = \lfloor S^{\frac{1}{2\alpha}}_\varepsilon \rfloor \), then (3.7) and (3.9) yield
\[
\int c \, d\pi - \int c \, d\pi^* \leq \frac{2BCg(S_\varepsilon)}{S_\varepsilon} \leq \frac{8BC}{S_\varepsilon}
\]
while (recall \( N = 2 \)) Lemma 2.3(i) yields \( D_f(\tilde{\pi}, \mu^n_1 \otimes \mu^n_2) \leq \varphi(n) \) 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. \( \square \)

Similarly as in Remark 3.6, the asymptotic constant in Theorem 3.8 can be improved from 8 to 2. We also have the following analogue of Example 3.4.
Example 3.9 ($L^\rho$ regularization). For the $L^\rho$ regularization $f(x) = \frac{1}{\rho} x^\rho - 1$ with $\rho > 1$, Theorem 3.8 (ii) implies that for all $\varepsilon \in (0,1]$, 

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

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

Remark 3.10 (Comparison with Theorem 3.3). Let $N = 2$ for simplicity. As any quantization of the coupling $\pi^*$ 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 $\pi^*$ 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 $\text{OT}$ admits an optimal transport $\pi^*$ satisfying $\text{quant}_2(C,\alpha)$ for some good $\alpha$. Indeed, $\text{quant}_2(C,\alpha)$ 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 $\pi^*$ is given by a Lipschitz transport map over $X_1$, then $\pi^*$ inherits the quantization rate from $\mu_1$, 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 \[40, 42\], and clearly a Lipschitz map can almost never be expected for unbounded marginals. On the other hand, as emphasized in \[44\], 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 3.11. 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 $\text{quant}_2(C,1/d)$ for some $C > 0$. 

17
Proof. Let \( \Delta = \{(x, x) : x \in \mathbb{R}^d\} \) be the diagonal and \( \text{proj}^\Delta : \mathbb{R}^{2d} \to \Delta \) 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 
\[ \eta := \text{proj}^\Delta_\# \pi \]
which is concentrated on \( \Delta \); we claim that \( \eta \) satisfies \( \text{quant}_2(C, 1/d) \). 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 \( \text{proj}^\Delta \) can be written as \( (u, v) \mapsto (u, 0) \). Thus \( \eta \) 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 \( \eta \) satisfies \( \text{quant}_2(C, 1/d) \).
To show the same rate for \( \pi \), we use Minty’s trick [47] along the lines of [3]. Recall that the support \( \Gamma := \text{spt} \pi \) is \( c \)-cyclically monotone (e.g., [58]), 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 \( \text{proj}^\Delta \) admits an inverse map \( \ell : \text{proj}^\Delta(\Gamma) \to \Gamma, (u, 0) \mapsto (u, v) \) and moreover \( \ell \) is \( \sqrt{2} \)-Lipschitz. By Kirszbraun’s theorem, we can extend \( \ell \) to a \( \sqrt{2} \)-Lipschitz map \( \Delta \to \mathbb{R}^d \times \mathbb{R}^d \), still denoted \( \ell \). Note that \( \pi = \ell_\# \eta \) and any quantization of \( \eta \) on \( \Delta \) pushes forward to a quantization of \( \pi \). In view of the \( \sqrt{2} \)-Lipschitz property, we conclude that \( \pi \) satisfies \( \text{quant}_2(\sqrt{2}C, 1/d) \). \( \square \)

The following combines Lemma 3.11 with Theorem 3.8 and Example 3.9.

**Corollary 3.12** (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
\[ \text{OT}_{f, \varepsilon} - \text{OT} \leq \frac{d}{2} \varepsilon \log \left( \frac{1}{\varepsilon} \right) + K\varepsilon, \quad \varepsilon \in (0, 1]. \]
(ii) Let \( f(x) = \frac{1}{\rho} x^\rho - 1 \) with \( \rho > 1 \). There exists \( K > 0 \) such that
\[
\text{OT}_{f,\varepsilon} - \text{OT} \leq K\varepsilon^{\frac{1}{(\rho-1)d^2+1}}, \quad \varepsilon \in (0, 1].
\]

Next, we aim to generalize Lemma 3.11 from quadratic to more general costs. Following [44], 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 \( c \in C^2(X) \). We say that \( c \) is nondegenerate if \( D_{xy}^2 c(x, y) \) is invertible for all \( (x, y) \in X \). Here \( D_{xy}^2 c(x, y) \) denotes the \( d \times d \) matrix \( [\partial_{x,y}^2 c(x, y)]_{1\leq i,j\leq d} \). We follow the terminology of [44]; the condition is called (A2) in [42] 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 3.13.** Let \( X_1, X_2 \subset \mathbb{R}^d \) be convex and let \( c \in C^2(X) \) be nondegenerate. If \( \mu_1, \mu_2 \) are compactly supported, then any optimal transport satisfies \( \text{quant}_2(C, 1/d) \) for some \( C > 0 \).

For a proof, see Steps 1 and 2 in the proof of Lemma 3.14 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 3.11 and [44, Theorem 1.1] with a cut-off argument. We denote by \( |M| \) the operator norm of the matrix \( M \).

**Lemma 3.14.** Let \( X_1, X_2 \subset \mathbb{R}^d \) be convex and let \( c \in C^2(X) \) be nondegenerate. Suppose that \( D_{xy}^2 c(x, y) \) is uniformly continuous and \( \|D_{xy}^2 c\| \), \( \|(D_{xy}^2 c)^{-1}\| \) are bounded on \( X \). Let \( d' > d \). If \( \mu_1, \mu_2 \in \mathcal{P}_q(\mathbb{R}^d) \) for \( q := \frac{2d^2+d}{d^2+d} \), then any optimal transport satisfies \( \text{quant}_2(C, 1/d') \) for some \( C > 0 \).

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

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

\[
= D_{xy}^2 c(x, -M^{-1}y)M^{-1} - D_{xy}^2 c(x_0, y_0)M^{-1}
\]

19
and hence
\[ \|D^2_{xy}G(x, y)\| \leq \|M^{-1}\| \|D^2_{xy}c(x, -M^{-1}y) - D^2_{xy}c(x_0, y_0)\|. \]
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 $\|D^2_{xy}G(x, y)\| \leq \frac{1}{2}$ 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 $(x, -M^{-1}y), (x', -M^{-1}y') \in Q \cap \text{spt}\, \pi$. Then the $c$-cyclical monotonicity of spt $\pi$ yields
\[ c(x, -M^{-1}y) + (x', -M^{-1}y') \leq c(x, -M^{-1}y) + c(x', -M^{-1}y') \]
or equivalently
\[ x \cdot y + G(x, y) + x' y' + G(x', y') \geq x \cdot y' + G(x, y') + x' \cdot y + G(x', y). \quad (3.10) \]
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 [44, 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
\[ |\Delta u|^2 - |\Delta v|^2 = 2 \Delta x \cdot \Delta y \geq -|\Delta x| |\Delta y| \]
\[ \geq -\frac{1}{2} (|\Delta x|^2 + |\Delta y|^2) = -\frac{1}{2} (|\Delta u|^2 + |\Delta v|^2) \]
and thus
\[ |\Delta v| \leq \sqrt{3} |\Delta u|. \quad (3.11) \]
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 $\eta := a \# \pi|_Q$, we see that 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 \(\|M\| = \|D_{x'y'}c(x_0, y_0)\|\) is uniformly bounded. Recall from Remark 2.1 that a measure on \(\mathbb{R}^d\) with bounded support satisfies \(\text{quant}_2(C_0, 1/d)\) 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, \(\eta\) satisfies \(\text{quant}_2(C_0, 1/d)\) with a constant \(C_0\) independent of \((x_0, y_0)\).

The map \(a\) admits a Lipschitz inverse \(\ell : a(Q \cap \text{spt} \pi) \to Q \cap \text{spt} \pi\), with a Lipschitz constant \(L\) independent of \((x_0, y_0)\) due to the boundedness of \(\|(D_{x'y'}c)^{-1}\|\) and (3.11). Again, by Kirszbraun’s theorem, \(\ell\) extends to a Lipschitz map \(\ell : I \to \mathbb{R}^d \times \mathbb{R}^d\) with the same Lipschitz constant. As \(\pi = \ell_\# \eta\), we deduce that \(\pi\) satisfies \(\text{quant}_2(C, 1/d)\) 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 \(\tilde{\nu}_i\) satisfies \(\text{quant}_2(C, \alpha)\) 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 \(\sum_{i=1}^m \nu_i(X) \tilde{\nu}_i\) of probability measures (and the zero measure, if necessary). Noting that given \(n = km\) points, we can allocate \(k\) points to each of the \(\tilde{\nu}_i\), it is easy to see that \(\tilde{\nu}\) satisfies \(\text{quant}_2(m^\alpha C, \alpha)\) 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 \(R = N\tau\) 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 \(\pi|Q_R\) satisfies \(\text{quant}_2((R/r)^2 C, 1/d)\) for \(n \geq (R/r)^{2d}\).

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

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

\[
R(n) := r \left[\frac{1}{n^d} - \frac{1}{d} \right]^{1/2}.
\]

Note \(\lim_{n \to \infty} R(n) = \infty\) and \(n \geq (R(n)/r)^{2d}\) and \((R(n)/r)^{2n^{-1/d}} \leq n^{-1/d'}\). Writing \(\pi_n := \pi|Q_R(n)\), the above shows that there exist \(\nu_n \in \mathbb{P}^n(\mathbb{R}^{2d})\) such that \(W_2(\nu_n, \pi_n) \leq Cn^{-1/d'}\). 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 \(\gamma := \frac{4d}{d-\alpha} = \frac{4}{d} \left(\frac{1}{2} - \frac{1}{d'}\right)^{-1} \); then \(R(n)^{-\gamma} \leq C'n^{-2/d'}\) for a constant \(C' > 0\) and as \(2 + \gamma = 2\frac{d+4}{d-\alpha}\), 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 $\pi_n$ that $\pi$ satisfies $\text{quant}_2(C, 1/d')$ for a (different) constant $C$.

Lemma 3.13 and Lemma 3.14 have immediate corollaries similar to Corollary 3.12; we omit the statements for brevity.

The nondegeneracy condition can be extended to the multi-marginal transport problem, and is used in [53, 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 [53, Remark 2.13]. 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.

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 $\text{OT}_{f, \epsilon} - \text{OT}$ are naturally obtained from the dual problem of $\text{OT}_{f, \epsilon}$.

**Lemma 4.1.** Let $\hat{h}_i \in L^1(\mu_i)$, $i = 1, 2$ be Kantorovich potentials for $\text{OT}$ 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 $f^*_\epsilon(y) := \epsilon f^*(\frac{1}{\epsilon}y)$. Then

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

$$\geq \sup_{a \in \mathbb{R}} \left( a - f^*_\epsilon(a) \int 1_{a \geq \hat{c}} \, d(\mu_1 \otimes \mu_2) \right) - \epsilon f^*(0).$$

**Proof.** Recall (e.g., [29, 57]) the duality

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

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

$$\text{OT}_{f, \epsilon} - \text{OT} \geq \sup_{a \in \mathbb{R}} \left( a - \int f^*_\epsilon(a - \hat{c}) \, d(\mu_1 \otimes \mu_2) \right).$$
As \( f_\varepsilon^* \) is nondecreasing, \( \hat{c} \geq 0 \) and \( f_\varepsilon^*(0) = \varepsilon f^*(0) \geq -\varepsilon f(1) = 0 \), we also have \( \int f_\varepsilon^*(a - \hat{c}) \, d(\mu_1 \otimes \mu_2) \leq f_\varepsilon^*(a) \int 1_{a \geq \varepsilon} \, d(\mu_1 \otimes \mu_2) + f_\varepsilon^*(0) \), 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 \( L^\rho \) 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 \( \varepsilon > 0 \),

\[
\text{OT}_{f,\varepsilon} - \text{OT} \geq d \varepsilon \log(1/\varepsilon) - (2^d - 1) \varepsilon.
\]

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

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

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 \( \int_{\mathbb{R}} e^{x/\varepsilon} \, du = 2\varepsilon \) of the Laplace distribution,

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

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

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

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

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

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

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

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

$$\text{OT}_{f, \varepsilon} - \text{OT} \geq \sup_{a \in \mathbb{R}_+} \left( a - 2^d f^*_\varepsilon(a)a^d - \varepsilon f^*(0) \right)$$

(4.1)

$$= \sup_{a \in \mathbb{R}_+} \left( a - 2^d \frac{a^{d+q}}{q^{\varepsilon^{-1}}} - 2^d \varepsilon a^d - \varepsilon \right).$$

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

**Remark 4.3.** We can similarly show the sharpness of Corollary 3.12(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 $1_{a \geq c} \leq \prod_{i=1}^d 1_{\{|x_i - y_i| \leq \sqrt{a}\}}$, and thus $\text{OT}_{f, \varepsilon} - \text{OT} \geq K\varepsilon^{1/(q-1)} + O(\varepsilon)$. 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 $\frac{1}{2\varepsilon} \log \left( \frac{1}{\varepsilon} \right)$ 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, 52]. Very recently, [18] showed that this term is sharp for the broad class of nondegenerate (as defined before Lemma 3.13) 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 $c \in C^2(X)$ be nondegenerate.
(i) Let \( f(x) = x \log(x) \). Then
\[
\text{OT}_{f, \varepsilon} - \text{OT} \geq \frac{d}{2} \varepsilon \log(1/\varepsilon) + O(\varepsilon).
\]
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
\[
\text{OT}_{f, \varepsilon} - \text{OT} \geq K \varepsilon \frac{1}{(\rho - 1)d/2 + 1} + O(\varepsilon)
\]
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 3.9.

Proof. See [18, Proposition 4.4] for (i). To show (ii), we argue that there exist constants \( C_0, C > 0 \) such that
\[
\text{OT}_{f, \varepsilon} \geq \text{OT} + \sup_{a \leq C_0} \left( a - C f^*_\varepsilon(a) d^{d/2} - \max\{0, f^*_\varepsilon(0)\} \right).
\] (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 \( A = \bigcup_{i=1}^n A_i \) of the compact set \( \{\hat{c} = 0\} \cap (X_1 \times X_2) \) satisfying the following:

(a) On the compact \( B := (X_1 \times X_2) \setminus A \) 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 \( v \in \mathbb{R}^d \),
\[
\int_{A_i} \mathbf{1}_{a \geq \hat{c}} d(\mu_1 \otimes \mu_2) \leq C_1 \int_{B_r} \int_{B_r} \mathbf{1}_{a \geq |u - r_v|^2/4} dudv,
\]
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
\[
\int_{B_r} \mathbf{1}_{a \geq |u - r_v|^2/4} du \leq \int_{\mathbb{R}^d} \mathbf{1}_{a \geq |u|^2/4} du \leq \prod_{i=1}^d \int_{\mathbb{R}} \mathbf{1}_{|u_i| \leq 2 \sqrt{n}} du \leq 4^d a^{d/2},
\]
we obtain
\[
\int_{A \cap (X_1 \times X_2)} \mathbf{1}_{a \geq \hat{c}} d(\mu_1 \otimes \mu_2) \leq Ca^{d/2}.
\]
for a constant $C > 0$. In view of (a), this shows

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

and now (4.2) follows by Lemma 4.1.

\section{Appendix}

The following is well known in the entropic case [48, Section 5]. For completeness, we provide an extension to the $f$-divergences under consideration.

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

\textbf{Proof.} If there exists an optimal transport plan $\pi^*$ with finite divergence, clearly $\text{OT}_{f, \varepsilon} - \text{OT} \leq \varepsilon D_f(\pi^*, P) = O(\varepsilon)$. Conversely, let $\pi_\varepsilon$ be an optimizer of $\text{OT}_{f, \varepsilon}$. If $\text{OT}_{f, \varepsilon} - \text{OT} = O(\varepsilon)$, 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 $d\pi_\varepsilon/ dP$, meaning that $(\pi_\varepsilon_n)$ converge set-wise. The limit $\pi_0$ is again a coupling. We have $\int c d\pi_0 \leq \lim \inf \text{OT}_{f, \varepsilon_n} = \text{OT}$ by a generalized Fatou’s lemma [56, p.231] and the growth condition on $c$, showing that $\pi_0$ is an optimal transport. The same Fatou’s lemma shows $D_f(\pi_0, P) \leq \lim \inf D_f(\pi_\varepsilon_n, P) < \infty$, completing the proof. \hfill \Box

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

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

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

(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 $\omega$, then choosing $\theta \in \Pi(\pi, \tilde{\pi})$ attaining $W_1(\pi, \tilde{\pi})$ yields

$$
\left| \int c d(\pi - \tilde{\pi}) \right| \leq \int |c(x) - c(y)| \theta(dx, dy)
\leq \int \omega(d_{X,1}(x, y)) \theta(dx, dy) = \omega(W_1(\pi, \tilde{\pi}))
$$
by Jensen’s inequality. Going through the proof of Theorem 3.3 with (A.1), we obtain instead of (3.1) that

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

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 \( \varepsilon \) 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

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

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

References


