Quantitative Stability of Regularized Optimal Transport and Convergence of Sinkhorn's Algorithm

Stephan Eckstein* Marcel Nutz†

June 8, 2022

Abstract

We study the stability of entropically regularized optimal transport with respect to the marginals. Lipschitz continuity of the value and Hölder continuity of the optimal coupling in p-Wasserstein distance are obtained under general conditions including quadratic costs and unbounded marginals. The results for the value extend to regularization by an arbitrary divergence. As an application, we show convergence of Sinkhorn's algorithm in Wasserstein sense, including for quadratic cost. Two techniques are presented: The first compares an optimal coupling with its so-called shadow, a coupling induced on other marginals by an explicit construction. The second transforms one set of marginals by a change of coordinates and thus reduces the comparison of differing marginals to the comparison of differing cost functions under the same marginals.

Keywords Entropic Optimal Transport; Stability; Sinkhorn's Algorithm; IPFP
AMS 2010 Subject Classification 90C25; 49N05

1 Introduction

Following advances allowing for computation in high dimensions, applications of optimal transport are thriving in areas such as machine learning, statistics, image and language processing (e.g., [4, 15, 50, 3]). Regularization plays a key role in enabling efficient algorithms with provable convergence;

*Department of Mathematics, ETH Zurich, seckstein@ethz.ch. Research supported by Landesforschungsförderung Hamburg under project LD-SODA. SE thanks Daniel Bartl, Mathias Beiglböck and Gudmund Pammer for fruitful discussions and helpful comments.

†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. MN is grateful to Guillaume Carlier, Giovanni Conforti, Flavien Léger and Luca Tamanini for their kind hospitality and advice.

see [48] for a recent monograph with numerous references. Popularized in this context by [20], entropic regularization is the most popular choice as it allows for Sinkhorn's algorithm (iterative proportional fitting procedure) that can be implemented at large scale using parallel computing and is analytically tractable. The entropically regularized transport problem can be formulated as

$$S^{\varepsilon}_{\mathrm{ent}}(\mu_1, \mu_2, c) = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int c(x, y)\, \pi(dx, dy) + \varepsilon D_{\mathrm{KL}}(\pi, \mu_1 \otimes \mu_2). \quad (1.1)$$

Here $\Pi(\mu_1, \mu_2)$ is the set of couplings of the given marginals μ_1, μ_2 and $D_{\mathrm{KL}}(\cdot, \mu_1 \otimes \mu_2)$ is the Kullback–Leibler divergence relative to the product measure $\mu_1 \otimes \mu_2$. Moreover, $\varepsilon > 0$ is a regularization parameter and c is a cost function; the most important example is quadratic cost $\|x - y\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$. The basic idea is to solve (1.1) for small $\varepsilon > 0$ to obtain an approximation of the (unregularized) optimal transport problem that corresponds to $\varepsilon = 0$. Starting with [16, 42, 43] and followed by [14, 37], the convergence as $\varepsilon \to 0$ has been studied in detail and remains a very active area of investigation; see for instance [2, 5, 6, 7, 17, 34, 45, 47, 53].

The entropic optimal transport problem (1.1) is also of its own interest. On the one hand, it is equivalent to a static formulation of the Schrödinger bridge problem that has a long history in physics (see [27, 38] for surveys); the dynamic Schrödinger bridge can be constructed by solving the static problem and combining it with a Brownian bridge. On the other hand, applied researchers have started to exploit numerous benefits resulting from entropic regularization, such as smoothness, existence of a gradient for gradient descent, improved sampling complexity (e.g., [18, 21, 30, 31]), among many others. Thus, the regularization is increasingly seen as an advantage rather than an approximation error; notions such as Sinkhorn divergence [32, 49] have become tools of their own right. We note that as long as $\varepsilon > 0$ is fixed, we can assume without loss of generality that $\varepsilon = 1$, simply dividing (1.1) by ε and using the cost function c/ε. Hence, we shall drop ε from the formulation in our results.

The main objective of the present study is to establish and quantify the stability of the value S_{ent} and its optimal coupling π^* with respect to the input marginals μ_1 and μ_2, or more generally $\mu_1, \ldots, \mu_N$ in the multimarginal setting. Distances will be quantified by Wasserstein distance W_p, thus allowing for comparison of measures with different supports, discrete and continuous measures, etc. We aim for results including unbounded marginals, replacing compactness by suitable integrability conditions such as the subgaussian tails in [41]. Schrödinger bridges are one application where

unbounded supports are very natural, as the Brownian dynamics produce unbounded intermediate marginals even if the boundary data are bounded. In this context, costs are usually quadratic, so that unbounded and non-Lipschitz cost functions are necessary. Even in applications with bounded costs, one may be interested in estimates with constants that do not depend on $\|c\|_\infty$, especially not exponentially.

To the best of our knowledge, the first stability result for entropic optimal transport is due to [12]. Here, costs are uniformly bounded, and all marginals are equivalent to a common reference measure (e.g., Lebesgue), with densities uniformly bounded above and below. Within these families, distances of measures can be quantified by the L^p norm of the difference of their densities. The authors show that the Schrödinger potentials (i.e., the dual entropic optimizers) are Lipschitz continuous relative to the marginals in L^p, for $p = 2$ and $p = \infty$. This result is obtained by a differential approach establishing invertibility of the Schrödinger system. More recently, [33] obtain the first result on stability in a general setting. Using a geometric approach called cyclical invariance, continuity of optimizers is established in the sense of weak convergence. The geometric method avoids integrability conditions almost entirely and indeed remains valid even if the value of (1.1) is infinite. On the other hand, the method relies on differentiation of measures which essentially forces the marginal spaces to be finite-dimensional. More importantly, the continuity result is purely qualitative, and that is the main difference with the present results. Most recently, and partly concurrently with the present study, a beautiful result of [22] establishes the uniform stability of Sinkhorn's algorithm with respect to the marginals, in a bounded setting. As a consequence, the authors deduce Lipschitzianity in W_1 of the optimal couplings with respect to the marginals; the assumptions include bounded Lipschitz costs and bounded spaces. The argument is based on the Hilbert–Birkhoff projective metric which has also been used successfully to show linear convergence of Sinkhorn's algorithm [13, 29]. A crucial additional step accomplished in [22] is to pass from this metric to a more standard norm on the potentials. The techniques involving the projective metric are less probabilistic in nature, which may be one reason why it is wide open how to relax the boundedness conditions. We remark that the initial result of [12] also covered the multimarginal problem which has recently become popular due to its role in the Wasserstein barycenter problem [1, 11]. At least in the context of [10], it was observed that Hilbert–Birkhoff arguments may not be equally successful beyond two marginals. Finally, we mention the follow-up [46] on the continuity of the potentials in unbounded settings.

We apply our stability result to Sinkhorn's algorithm for $N = 2$ marginals.

It is well known that each iterate π^n of the algorithm solves an entropic optimal transport problem between its own marginals, and moreover these marginals converge to the given marginals μ_i. Thus, the convergence can be seen as a particular instance of stability with respect to marginals and our results apply. Sinkhorn's algorithm has been studied over almost a century (see [48] for numerous references); the most general convergence results in this literature are due to [51]. While they treat costs that are merely measurable and show $\pi^n \to \pi^*$ in total variation, they do not cover unbounded functions like the quadratic cost in most examples, especially when both marginals have unbounded support. Applying stability results under regularity of c turns out to be fruitful in this regard: we not only obtain the convergence to the optimal value and $\pi^n \to \pi^*$ in Wasserstein distance, but even a rate of convergence. The conditions are sufficiently general to cover quadratic cost with subgaussian marginals.

1.1 Synopsis

Our first result, detailed in Theorem 3.7, is the continuity of the value S_{ent} with respect to the marginals in p-Wasserstein distance under generic conditions. If the cost c is a product of suitably integrable Lipschitz functions, then S_{ent} is also Lipschitz. This includes quadratic costs on $\mathbb{R}^d$ with possibly unbounded marginal supports. The proof is based on comparing the optimizer π^* with the "shadow" coupling it induces on other marginals. The shadow is a particular projection that we construct explicitly by gluing, controlling both the distance to π^* and its divergence. The construction is simple and flexible, thus potentially useful for other purposes. For instance, Theorem 3.7 holds for a general class of optimal transport problems regularized by a divergence D_f as previously considered in [24], Kullback–Leibler divergence is a particular case. Other divergences, especially quadratic, are being used in some applications where entropic regularization performs poorly, usually because non-equivalent optimizers are desired or weak penalization (small ε) causes numerical instabilities, see [8, 25, 39]. Theoretical results are scarce so far as these regularizations are less tractable.

By way of strong convexity, the continuity of the value S_{ent} in Theorem 3.7 leads to the continuity of the optimizer π^* with respect to the marginals. Theorem 3.11 states a nonasymptotic inequality bounding the distance of two entropic optimizers for different marginals in terms of the W_p distance of the marginals. It shows in particular that the map $(\mu_1, \dots, \mu_N) \mapsto \pi^*$ is $1/(2p)$-Hölder in W_p. Exploiting a Pythagorean-type property of relative entropy to implement the strong convexity, we achieve an unbounded

setting requiring only a transport inequality; i.e., a control of Wasserstein distance through entropy. This condition holds as soon as the marginals have a finite exponential moment; in particular, the result covers quadratic costs when marginals are σ^2-subgaussian for some (arbitrarily small) σ. We remark that Theorem 3.7 is the first quantitative stability result for unbounded costs, and in settings without differentiation of measures as assumed in [33], even the qualitative result alone would be novel.

One noteworthy feature of Theorem 3.11 is that the constants grow only linearly in c, which is particularly important for the regularized transport problem (1.1): here the effective cost function is $\tilde{c} := c/\varepsilon$ and ε is usually small. Many results on entropic optimal transport feature constants depending exponentially on the cost, typically $\exp(\|\tilde{c}\|_\infty)$ or $\exp(\|\tilde{c}\|_\infty + \operatorname{Lip}\tilde{c})$, including all previous results on stability that we are aware of. Even for well-behaved c on a fairly small domain, a choice like $\varepsilon = .01$ then leads to constants far exceeding e^{100}, potentially a concern in practical considerations.

Our second continuity result, Theorem 3.13, aims at improving the Hölder exponent in Theorem 3.11 under the more restrictive condition that the cost c is bounded (spaces may still be unbounded). For instance, we show $1/(p+1)$-Hölder continuity in W_p. More generally, Theorem 3.13 yields the Hölder exponent $p/(p+1)q$ from W_p to W_q; to wit, we can improve the exponent by measuring the distance of the marginals in a stronger norm. In particular, $p = \infty$ leads to a Lipschitz result into W_1. This choice also eliminates exponential dependence of the constant on the cost. In fact, we prove that the Lipschitz constant is *sharp* in a nontrivial discrete example. This may be surprising given that the idea of proof is somewhat circuitous and that many estimates in this area are thought to be overly conservative.

Indeed, Theorem 3.13 is based on a novel approach that may be of independent interest; the basic idea is to reduce the problem of differing marginals to one of differing cost functions (under the same marginals). In the latter problem, optimizers are measure-theoretically equivalent and comparable in the sense of Kullback–Leibler divergence. Our starting point is the observation that the regularization in our problem depends only on the relative density, but not on the geometry of the distributions. In the simplest case, a W_p-optimal coupling of the differing marginals induces an invertible transport map T that can be used as change of coordinates to achieve identical marginals. The cost is transformed at the same time and we end up comparing c with $c \circ T$. For this comparison, we can apply a separate result (Proposition 3.12) based on an entropy calculation.

The application to Sinkhorn's algorithm is summarized in Theorem 3.15

which states convergence of the entropic cost and of the Sinkhorn iterates π^n themselves. The qualitative and quantitative results follow from Theorem 3.7 and Theorem 3.11. In essence, the stability results turn a convergence rate for the Sinkhorn marginals into a convergence rate for $\pi^n \to \pi^*$. We use the sublinear rate for the marginals as obtained in [36]. As noted there, these rates are likely suboptimal—for bounded cost functions, linear convergence of Sinkhorn's algorithm is well known [10, 13, 29]—our focus at this stage is on having *some* quantitative control.

The organization of this paper is simple: Section 2 details the setting, Section 3 presents the main results, and Section 4 contains the proofs.

2 Setting and Notation

Let (Y, d_Y) be a Polish space and $\mathcal{P}(Y)$ its set of Borel probability measures. Given $p \in [1, \infty)$, we 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) $\hat{x} \in Y$. For $p = \infty$, we define $\mathcal{P}_\infty(Y)$ as the measures with bounded support. The p-Wasserstein distance $W_p(\mu, \nu)$ between $\mu, \nu \in \mathcal{P}_p(Y)$ is defined via

$$\begin{aligned} W_p(\mu,\nu)^p &= \inf_{\pi \in \Pi(\mu,\nu)} \int d_Y(x,y)^p \, \pi(dx,dy), \quad p \in [1,\infty), \\ W_\infty(\mu,\nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \operatorname*{ess\,sup}_{(x,y)\sim\pi} d_Y(x,y), \end{aligned}$$

while $\|\mu - \nu\|_{TV} = \sup_{A \subseteq Y \text{ Borel}} |\mu(A) - \nu(A)|$ is the total variation distance of $\mu, \nu \in \mathcal{P}(Y)$.

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 write $x \in X$ as $x = (x_1, \dots, x_N)$. When $p \in [1, \infty]$ is given, it will be convenient to use on X the particular product metric

$$d_{X,p}(x,y) := \begin{cases} \left(\sum_{i=1}^N d_{X_i}(x_i, y_i)^p \right)^{1/p}, & p \in [1,\infty), \\ \max_{i=1,\dots,N} d_{X_i}(x_i, y_i), & p = \infty. \end{cases}$$

Unless otherwise noted, p-Wasserstein distances on X are understood with respect to $d_{X,p}$. Similarly, the distance between two tuples of marginals will often be quantified by

$$W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N) := \begin{cases} \left(\sum_{i=1}^N W_p(\mu_i, \tilde{\mu}_i)^p \right)^{1/p}, & p \in [1,\infty), \\ \max_{i=1,\dots,N} W_\infty(\mu_i, \tilde{\mu}_i), & p = \infty. \end{cases}$$

Given a Lipschitz function $c : X \to \mathbb{R}$, we denote by $\mathrm{Lip}_p(c)$ its Lipschitz constant with respect to $d_{X,p}$.

For a strictly convex, lower bounded function $f : \mathbb{R}_+ \to \mathbb{R}$ 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 the same space is

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

and $D_f(\mu, \nu) := \infty$ for $\mu \not\ll \nu$. The main example of interest to us is the Kullback–Leibler divergence (relative entropy) $D_{\mathrm{KL}}(\mu, \nu)$ which corresponds to the choice $f(x) := x \log x$. We always assume that $(\mu, \nu) \mapsto D_f(\mu, \nu)$ is lower semicontinuous for weak convergence. This holds for D_{KL}, and more generally whenever D_f has a suitable variational representation.

Given $\mu_i \in \mathcal{P}(X_i)$ and a continuous, nonnegative[1] cost function $c \in L^1(\mu_1 \otimes \cdots \otimes \mu_N)$, we can now introduce the regularized transport problem

$$S(\mu_1, \ldots, \mu_N, c) = \inf_{\pi \in \Pi(\mu_1,\ldots,\mu_N)} \int c \, d\pi + D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N), \tag{2.1}$$

where $\Pi(\mu_1, \ldots, \mu_N) \subset \mathcal{P}(X)$ denotes the set of couplings of the marginals μ_i. Note that $S(\mu_1, \ldots, \mu_N, c) < \infty$ by way of $\pi := \mu_1 \otimes \cdots \otimes \mu_N$. A standard argument of compactness and strict convexity then shows that (2.1) admits a unique optimizer $\pi^* \in \Pi(\mu_1, \ldots, \mu_N)$. When $p \in [1, \infty)$ is given, we always assume that c has growth of order p,

$$|c(x)| \le C(1 + d_{X,p}(x, \hat{x})^p) \tag{2.2}$$

for some $C > 0$ and $\hat{x} \in X$, whereas for $p = \infty$ the meaning is that c is bounded. For marginals $\mu_i \in \mathcal{P}_p(X_i)$, this ensures in particular that $c \in L^1(\pi)$ for any coupling π.

While some of our results below hold for general divergences, we use the notation S_{ent} in results specific to the entropic version, so that (2.1) becomes

$$S_{\mathrm{ent}}(\mu_1, \ldots, \mu_N, c) = \inf_{\pi \in \Pi(\mu_1,\ldots,\mu_N)} \int c \, d\pi + D_{\mathrm{KL}}(\pi, \mu_1 \otimes \cdots \otimes \mu_N). \tag{2.3}$$

Remark 2.1. A variation of (2.3) uses entropy relative to a reference measure $\hat{P}$ different from the product of the marginals,

$$\inf_{\pi \in \Pi(\mu_1,\ldots,\mu_N)} \int c \, d\pi + D_{\mathrm{KL}}(\pi, \hat{P}), \tag{2.4}$$

[1] The lower bound is easily relaxed in view of the behavior of (2.1) under shifts of c.

for instance (normalized) Lebesgue measure for problems with absolutely continuous marginals on $\mathbb{R}^d$. Of course, a compatibility condition between $\hat{P}$ and the marginals is necessary to guarantee that (2.4) is finite. As long as $\hat{P} = \hat{P}_1 \otimes \cdots \otimes \hat{P}_N$ is a product measure, a standard computation shows that the optimizer π^* of this problem is the same as the one of (2.3). Therefore, our stability results for (2.3) carry over to (2.4).

3 Results

3.1 Shadows and Preliminaries

Given $\pi \in \Pi(\mu_1, \ldots, \mu_N)$, we introduce a coupling $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$ of different marginals through a gluing construction. Intuitively, for $N = 2$, the transport $\tilde{\pi}$ 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. We think of $\tilde{\pi}$ as a coupling of $\tilde{\mu}_1, \tilde{\mu}_2$ that "shadows" $\pi \in \Pi(\mu_1, \mu_2)$ as closely as possible given the differing marginals. The formal definition reads as follows.

Definition 3.1 (Shadow)**.** 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* $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$ 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 $K(x) = K_1(x_1) \otimes \cdots \otimes K_N(x_N)$.

In general, the W_p-optimal kernel K_i need not be unique, so that there can in fact be more than one choice for the shadow. Any choice will do in what follows, and we shall speak of "the" shadow despite the abuse of language. As detailed in Remark 4.2, the shadow can also be understood as a particular choice of a W_p-projection of π onto $\Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$. The crucial additional property of the shadow is that its divergence is controlled by the one of π.

Lemma 3.2. *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*

$$W_p(\pi, \tilde{\pi}) = W_p(\mu_1, \ldots, \mu_N; \tilde{\mu}_1, \ldots, \tilde{\mu}_N),$$
$$D_f(\tilde{\pi}, \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N) \leq D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N).$$

To study the continuity properties of regularized optimal transport, we need to compare the cost of two couplings $\pi, \tilde{\pi}$ in the unregularized transport

problem. If c is L-Lipschitz, the following inequality holds for all probability measures $\pi, \tilde{\pi}$. We formulate an abstract condition to cover more general cases, especially Example 3.4 below.

Definition 3.3. Let $p \in [1, \infty]$ and $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$, $i = 1, \dots, N$. For a constant $L \geq 0$, we say that c satisfies (A_L) if

$$\left| \int c \, d(\pi - \tilde{\pi}) \right| \leq L W_p(\pi, \tilde{\pi}) \tag{A_L}$$

for all $\pi \in \Pi(\mu_1, \dots, \mu_N)$ and $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$.[2]

The most important application is quadratic cost.

Example 3.4. For $p = 2$ and cost $c(x_1, x_2) = \|x_1 - x_2\|^2$ on Euclidean space $\mathbb{R}^d \times \mathbb{R}^d$, we have that (A_L) holds with

$$L := \sqrt{2} \left[M(\mu_1) + M(\tilde{\mu}_1) + M(\mu_2) + M(\tilde{\mu}_2) \right]$$

where $M(\mu) := (\int \|x\|^2 \, \mu(dx))^{1/2}$ for $\mu \in \mathcal{P}(\mathbb{R}^d)$.

The example is a special case of the following observation.

Lemma 3.5. *Let $p \in [1, \infty)$. Let $c(x) = f(x)g(x)$ where f, g are Lipschitz and have growth of order at most $p - 1$. Then (A_L) holds with a constant L depending only on the Lipschitz and growth constants of f, g and the p-th moments of $\mu_i, \tilde{\mu}_i$, $i = 1, \dots, N$. For $p = \infty$, the analogue holds with dependence on the bounds of f, g instead of moments.*

This criterion generalizes to a product $c(x) = c_1(x) \cdots c_m(x)$ of m Lipschitz functions satisfying a suitable growth condition; cf. Remark 4.3.

The next example shows that (A_L) also holds for the p-th power as cost.

Example 3.6. For cost $c(x_1, x_2) = \|x_1 - x_2\|^p$ with $p \in (1, \infty)$ on Euclidean space $\mathbb{R}^d \times \mathbb{R}^d$, we have that (A_L) holds with

$$L := C_p \big[M_p(\mu_1) + M_p(\tilde{\mu}_1) + M_p(\mu_2) + M_p(\tilde{\mu}_2) \big]^{p-1},$$

where $M_p(\mu) := (\int \|x\|^p \, \mu(dx))^{1/p}$ for $\mu \in \mathcal{P}(\mathbb{R}^d)$ and C_p is a constant depending only on p.

The proof, detailed in Section 4, is similar to [52, Proposition 7.29] and proceeds by estimating the derivative of a curve connecting the integrals in question. The example generalizes to costs $c(x_1, x_2) = \bar{c}(x_1, x_2)^p$ with $\bar{c}$ being Lipschitz.

[2] In fact, (A_L) will only ever be used when one coupling is the shadow of the other, but that restriction does not seem to substantially enhance the applicability.

3.2 Stability through Shadows

We can now state our first result, establishing the continuity of (2.1) with respect to the marginals. The qualitative part (i) holds for general costs, the quantitative part (ii) applies, in particular, to quadratic costs under 2-Wasserstein distance.

Theorem 3.7 (Continuity of Value)**.** *Let $p \in [1, \infty]$.*

(i) Let $\mu_i, \mu_i^n \in \mathcal{P}_p(X_i)$ satisfy $\lim_n W_p(\mu_i, \mu_i^n) = 0$ for $i = 1, \dots, N$. Then $S(\mu_1^n, \dots, \mu_N^n, c) \to S(\mu_1, \dots, \mu_N, c)$ and the associated optimal couplings converge in W_p.

(ii) Let $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$ for $i = 1, \dots, N$ and let c satisfy (A_L). Then

$$|S(\mu_1, \dots, \mu_N, c) - S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)| \leq L W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N).$$

This result will be proved by comparing the cost of a coupling with the cost of its shadow. Using the same idea, we can show the convergence of the cost functionals as follows.

Remark 3.8 (Γ-Convergence)**.** Define $\mathcal{F} : \mathcal{P}_p(X) \to \mathbb{R} \cup \{\infty\}$ by

$$\mathcal{F}(\pi) = \begin{cases} \int c \, d\pi + D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N) & \text{if } \pi \in \Pi(\mu_1, \dots, \mu_N), \\ \infty, & \text{otherwise} \end{cases}$$

and similarly $\mathcal{F}_n$ for the marginals μ_i^n. If $\lim_n W_p(\mu_i, \mu_i^n) = 0$, then $\mathcal{F}_n$ Γ-converges to $\mathcal{F}$; that is, given $\pi \in \mathcal{P}_p(X)$,

(a) $\mathcal{F}(\pi) \leq \liminf \mathcal{F}_n(\pi_n)$ for any $(\pi_n)_{n \geq 1} \subset \mathcal{P}_p(X)$ with $W_p(\pi, \pi_n) \to 0$,

(b) there exists a sequence $(\pi_n)_{n \geq 1} \subset \mathcal{P}_p(X)$ with $W_p(\pi, \pi_n) \to 0$ and $\mathcal{F}(\pi) \geq \limsup \mathcal{F}_n(\pi_n)$.

For the recovery sequence in (b), we can choose $\pi_n \in \Pi(\mu_1^n, \dots, \mu_N^n)$ to be the shadow of $\pi \in \Pi(\mu_1, \dots, \mu_N)$.

Remark 3.9. Theorem 3.7 (i) and Remark 3.8 generalize to a sequence of cost functions c_n converging to c as long as the convergence is strong enough to imply $\int c_n \, d\pi_n \to \int c \, d\pi$ whenever $\pi_n \in \Pi(\mu_1^n, \dots, \mu_N^n)$ converge in W_p to some $\pi \in \Pi(\mu_1, \dots, \mu_N)$.

Our second aim is to bound the distance between the optimizers for different marginals. The line of argument requires controlling Wasserstein distance through entropy, hence it is natural to postulate a transport inequality. Given $q \in [1, \infty)$, we say that $\mu_i \in \mathcal{P}_q(X_i)$, $i = 1, \ldots, N$ satisfy (I_q) with constant C_q if

$$W_q(\pi, \theta) \leq C_q D_{\mathrm{KL}}(\theta, \pi)^{\frac{1}{2q}} \quad \text{for all} \quad \pi, \theta \in \Pi(\mu_1, \ldots, \mu_N). \tag{I_q}$$

Similarly, they satisfy (I_q') with constant C_q' if

$$W_q(\pi, \theta) \leq C_q' \left[D_{\mathrm{KL}}(\theta, \pi)^{\frac{1}{q}} + \left(\frac{D_{\mathrm{KL}}(\theta, \pi)}{2} \right)^{\frac{1}{2q}} \right] \tag{I_q'}$$

for all $\pi, \theta \in \Pi(\mu_1, \ldots, \mu_N)$. The two inequalities serve a similar purpose, but (I_q') is implied by a weaker integrability condition. Indeed, when X is bounded, (I_q) holds as a simple consequence of Pinsker's inequality. Using the weighted inequalities of [9], (I_q) and (I_q') also hold under much weaker exponential moment conditions on μ_i as detailed in (ii) and (iii) below. In (i), we obtain a different relaxation where all but one space X_i are bounded. Thus for the standard case $N = 2$, if one marginal is bounded, no condition at all is needed on the other marginal.

Lemma 3.10. *(i) Let $X' := X_2 \times \cdots \times X_N$ and suppose that*

$$\mathrm{diam}_q(X') := \sup_{x,y \in X'} d_{X',q}(x, y) < \infty.$$

Then (I_q) holds with $C_q = 2^{-\frac{1}{2q}} \mathrm{diam}_q(X')$ for all $\mu_i \in \mathcal{P}_q(X_i)$.

(ii) If $\mu_i \in \mathcal{P}(X_i)$ satisfy $\int \exp(\alpha\, d_{X_i}(\hat{x}_i, x_i)^{2q})\, \mu_i(dx_i) < \infty$ for some $\alpha \in (0, \infty)$ and $\hat{x}_i \in X_i$, then (I_q) holds with constant

$$C_q = 2 \inf_{\hat{x} \in X, \alpha > 0} \left(\frac{N}{2\alpha} \sum_{i=1}^{N} \left(1 + \log \int \exp(\alpha d_{X_i}(\hat{x}_i, x_i)^{2q})\, \mu_i(dx_i) \right) \right)^{\frac{1}{2q}}.$$

(iii) If $\mu_i \in \mathcal{P}(X_i)$ satisfy $\int \exp(\alpha\, d_{X_i}(\hat{x}_i, x_i)^{q})\, \mu_i(dx_i) < \infty$ for some $\alpha \in (0, \infty)$ and $\hat{x}_i \in X_i$, then (I_q') holds with constant

$$C_q' = 2 \inf_{\hat{x} \in X, \alpha > 0} \left(\frac{1}{\alpha} \sum_{i=1}^{N} \left(\frac{3}{2} + \log \int \exp(\alpha d_{X_i}(\hat{x}_i, x_i)^{q}) \mu_i(dx_i) \right) \right)^{\frac{1}{q}}.$$

Noting the logarithm in the formulas for C_q and C_q', we observe that these constants are typically much smaller than the exponential moment itself. We also note that the condition in (iii) covers subgaussian marginals for $q = 2$.

We can now state a quantitative result for the stability of the optimizer of (2.3) relative to the marginals. In view of the above, the assumptions cover quadratic cost under 2-Wasserstein distance and subgaussian marginals.

Theorem 3.11 (Stability of Optimizers)**.** *Let $p \in [1, \infty]$ and $q \in [1, \infty)$ with $q \le p$, let $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$, let $\mu_1, \dots, \mu_N$ satisfy (I_q) with constant C_q, and let c satisfy (A_L). Then the optimizers $\pi^*, \tilde{\pi}^*$ of $S_{ent}(\mu_1, \dots, \mu_N, c)$ and $S_{ent}(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$ satisfy*

$$W_q(\pi^*, \tilde{\pi}^*) \le N^{(\frac{1}{q} - \frac{1}{p})} \, \Delta + C_q \, (2L \, \Delta)^{\frac{1}{2q}} , \quad \Delta := W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N).$$

If $\mu_1, \dots, \mu_N$ satisfy (I_q') with constant C_q' instead of (I_q), then

$$W_q(\pi^*, \tilde{\pi}^*) \le N^{(\frac{1}{q} - \frac{1}{p})} \, \Delta + C_q' \left[(2L\Delta)^{\frac{1}{q}} + (L \, \Delta)^{\frac{1}{2q}} \right].$$

In particular, $(\mu_1, \dots, \mu_N) \mapsto \pi^*$ is $\frac{1}{2p}$-Hölder continuous in W_p when restricted to a bounded set of marginals satisfying (A_L) and (I_p) or (T_p') with given constants.

This result will be derived by comparing the optimizer with its shadow and applying a strong convexity argument, more specifically, a Pythagorean relation for relative entropy. In Theorem 3.11, only one set of marginals needs to satisfy (I_q) or (I_q'). If the assumption holds for both (μ_i) and $(\tilde{\mu}_i)$, the proof shows that L can be replaced by $L/2$ in the assertion.

3.3 Stability through Transformation

Next, we improve the Hölder exponent of Theorem 3.11 for the case of bounded cost. The general line of argument is to reduce a difference in marginals to a difference in cost functions. Thus, we first state a stability result for the cost function under fixed marginals; it may be of independent interest.

Proposition 3.12 (Stability wrt. Cost)**.** *Let $p \in [1, \infty]$, let $\mu_i \in \mathcal{P}_p(X_i)$, $i = 1, \dots, N$ and $P = \mu_1 \otimes \cdots \otimes \mu_N$. Let $c, \tilde{c} : X \to \mathbb{R}_+$ be bounded measurable, then the optimizers $\pi^*, \tilde{\pi}^*$ of $S_{ent}(\mu_1, \dots, \mu_N, c)$ and $S_{ent}(\mu_1, \dots, \mu_N, \tilde{c})$ satisfy*

$$\|\pi^* - \tilde{\pi}^*\|_{TV} \le \frac{1}{2} a^{\frac{1}{p+1}} \, \|c - \tilde{c}\|_{L^p(P)}^{\frac{p}{p+1}},$$

$$D_{\mathrm{KL}}(\pi^*, \tilde{\pi}^*) + D_{\mathrm{KL}}(\tilde{\pi}^*, \pi^*) \le a^{\frac{2}{p+1}} \|c - \tilde{c}\|_{L^p(P)}^{\frac{2p}{p+1}},$$

where $a := \exp(N\|c\|_\infty) + \exp(N\|\tilde{c}\|_\infty)$. *Let* $q \in [1,\infty)$. *If* $\mu_1,\dots,\mu_N$ *satisfy* (I_q) *with constant* C_q, *then also*

$$W_q(\pi^*,\tilde{\pi}^*) \le 2^{-\frac{1}{2q}} C_q \left(a^{\frac{1}{p}} \|c-\tilde{c}\|_{L^p(P)}\right)^{\frac{p}{(p+1)q}},$$

whereas if $\mu_1,\dots,\mu_N$ *satisfy* (I_q') *with constant* C_q', *then*

$$W_q(\pi^*,\tilde{\pi}^*) \le C_q' \left[\left(a^{\frac{1}{p}} \|c-\tilde{c}\|_{L^p(P)}\right)^{\frac{2p}{(p+1)q}} + 2^{-\frac{1}{2q}} \left(a^{\frac{1}{p}} \|c-\tilde{c}\|_{L^p(P)}\right)^{\frac{p}{(p+1)q}}\right].$$

(For $p=\infty$, the exponent $\frac{p}{(p+1)q}$ should be read as $\frac{1}{q}$.) Proposition 3.12 will be derived by comparing the optimizers in the sense of relative entropy $D_{\mathrm{KL}}(\pi^*,\tilde{\pi}^*)$. Of course, this is not possible in the other results where the marginals differ in a possibly singular way. We observe that the constant a deteriorates exponentially in $\|c\|_\infty$, however due to the $a^{\frac{1}{p}}$ in the formula this can be counteracted by using a stronger L^p norm. In particular, for $p=\infty$, the direct dependence on $\|c\|_\infty, \|\tilde{c}\|_\infty$ disappears completely, and moreover we obtain a Lipschitz estimate from L^∞ to W_1.

Those features are inherited by our final result on the stability with respect to marginals; it improves the Hölder exponent of Theorem 3.11 in the case of bounded costs. As above, the dependence of the constant on $\|c\|_\infty$ is avoided for $p=\infty$; we now obtain a Lipschitz result from W_∞ into W_1.

Theorem 3.13 (Stability of Optimizers for Bounded Cost)**.** *Let* $p \in [1,\infty]$ *and* $q \in [1,\infty)$ *with* $q \le p$, *let* $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$ *satisfy* (I_q) *with constant* C_q *and let* c *be bounded Lipschitz. Then the optimizers* $\pi^*, \tilde{\pi}^*$ *of* $S_{ent}(\mu_1,\dots,\mu_N,c)$ *and* $S_{ent}(\tilde{\mu}_1,\dots,\tilde{\mu}_N,c)$ *satisfy*

$$W_q(\pi^*,\tilde{\pi}^*) \le N^{(\frac{1}{q}-\frac{1}{p})}\,\Delta + 2^{-\frac{1}{2q}} C_q \left(a^{\frac{1}{p}}\ \mathrm{Lip}_p(c)\,\Delta\right)^{\frac{p}{(p+1)q}}$$

where $a := 2\exp(N\|c\|_\infty)$ *and* $\Delta := W_p(\mu_1,\dots,\mu_N;\tilde{\mu}_1,\dots,\tilde{\mu}_N)$. *If* $\mu_i, \tilde{\mu}_i$ *satisfy* (I_q') *with constant* C_q' *instead of* (I_q), *then*

$$\begin{aligned} W_q(\pi^*,\tilde{\pi}^*) &\le N^{(\frac{1}{q}-\frac{1}{p})}\,\Delta \\ &\quad + 2^{-\frac{1}{q}} C_q' \left[\left(a^{\frac{1}{p}}\,\mathrm{Lip}_p(c)\,\Delta\right)^{\frac{2p}{(p+1)q}} + 2^{-\frac{1}{2q}} \left(a^{\frac{1}{p}}\,\mathrm{Lip}_p(c)\,\Delta\right)^{\frac{p}{(p+1)q}}\right]. \end{aligned}$$

In particular, $(\mu_1,\dots,\mu_N) \mapsto \pi^*$ *is* $\frac{1}{p+1}$*-Hölder continuous in* W_p *when restricted to a bounded set of marginals satisfying* (I_p) *or* (T_p') *with a given constant. For* $q=1$ *and* $p=\infty$, *we have the Lipschitz estimate*

$$W_1(\pi^*,\tilde{\pi}^*) \le \ell\, W_\infty(\mu_1,\dots,\mu_N;\tilde{\mu}_1,\dots,\tilde{\mu}_N)$$

with constant $\ell := N + (C_1/\sqrt{2})\,\mathrm{Lip}_\infty(c)$ *independent of* $\|c\|_\infty$. *The constant* ℓ *is sharp.*

As discussed in the Introduction, this result is based on a transformation: instead of dealing with two sets of marginals, we use a change of coordinates to transform $\tilde{\mu}_i$ to μ_i, at the expense of also transforming the cost function. For the resulting problem, we can apply Proposition 3.12. The sharpness of the constant ℓ is discussed in Example 4.10.

Remark 3.14. For simplicity, we have stated our results in the traditional setting where W_p is defined through a metric compatible with the underlying Polish space. However, much of the above generalizes to any measurable metric. For instance, the discrete metric can be used to see that for $p = 1$, our results include the total variation distance (see also [46] for further results on continuity in total variation). The majority of our arguments extend without change to the more general setting. In Definition 3.1, it is no longer clear that there is a coupling attaining $W_p(\mu_i, \tilde{\mu}_i)$. However, we can use an ϵ-optimal coupling to define an "approximate shadow" for which the first part of Lemma 3.2 is replaced by $W_p(\pi, \tilde{\pi}) \le W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N) + \epsilon$, and then we can argue the main results as before. The extension to measurable metrics also applies to Proposition 3.12. Theorem 3.13 extends with the caveat that one needs to provide a substitute for the technical Lemma 4.9 (ii) in the specific metric under consideration, as its proof uses separability of the metric.

3.4 Application to Sinkhorn's Algorithm

In this section we focus on $N = 2$ marginals μ_1, μ_2. Sinkhorn's algorithm is initialized at $\pi^0 := P_c$, where $\frac{dP_c}{d(\mu_1 \otimes \mu_2)}(x) = \frac{\exp(-c(x))}{\int \exp(-c)\, d(\mu_1 \otimes \mu_2)}$ is the Gibbs kernel associated with the cost c. The Sinkhorn iterates $\pi^n \in \mathcal{P}(X)$, $n \ge 1$ can then be defined recursively via

$$\begin{aligned} \frac{d\pi^n}{d\pi^{n-1}}(x) &:= \frac{d\mu_1}{d\pi_1^{n-1}}(x_1) && \text{for } n \text{ odd}, \\ \frac{d\pi^n}{d\pi^{n-1}}(x) &:= \frac{d\mu_2}{d\pi_2^{n-1}}(x_2) && \text{for } n \text{ even}, \end{aligned}$$

where π_i^{n-1} is the i-th marginal of π^{n-1}. It follows that $\pi_1^n = \mu_1$ for n odd and $\pi_2^n = \mu_2$ for n even: for each iterate, one of the two marginals is the correct marginal. The other marginal does not match μ_i, but converges to it as $n \to \infty$. Importantly, each iterate π^n is the solution of an entropic

optimal transport problem between its own marginals. As these marginals converge to (μ_1, μ_2), the convergence of Sinkhorn's algorithm can be framed as a particular instance of stability with respect to the marginals. As above, we denote by π^* the optimizer of $S_{\mathrm{ent}}(\mu_1, \mu_2, c)$. Moreover, we write

$$\mathcal{F}(\pi) := \int c \, d\pi + D_{\mathrm{KL}}(\pi, \mu_1 \otimes \mu_2)$$

for the entropic cost of $\pi \in \mathcal{P}(X)$, similarly as in Remark 3.8 but without the penalty.

Theorem 3.15 (Sinkhorn Convergence)**.** *Let $p \in [1, \infty)$. For $i = 1, 2$, let $\mu_i \in \mathcal{P}(X_i)$ satisfy $\int \exp(\alpha \, d_{X_i}(\hat{x}_i, x_i)^p) \, \mu_i(dx_i) < \infty$ for some $\alpha \in (0, \infty)$ and $\hat{x}_i \in X_i$.*

(i) Let c be continuous with growth of order p. As $n \to \infty$, we have

$$\mathcal{F}(\pi^n) \to \mathcal{F}(\pi^*), \qquad \pi^n \to \pi^* \quad \text{in} \quad W_p.$$

(ii) Let $1 \le q \le p$ and $c(x) = f(x)g(x)$ where f, g are Lipschitz with growth of order $p - 1$. For all $n \ge 2$, with a constant c_0 detailed in the proof,

$$|\mathcal{F}(\pi^*) - \mathcal{F}(\pi^n)| \le c_0 n^{-\frac{1}{2p}}, \qquad W_q(\pi^*, \pi^n) \le c_0 n^{-\frac{1}{4pq}}.$$

Theorem 3.15 with $p = q = 2$ implies W_2-convergence for quadratic cost with subgaussian marginals. The form $c(x) = f(x)g(x)$ can be extended as in Remark 4.3, or more generally to any condition guaranteeing (A_L) uniformly over the marginals produced by the algorithm. In particular, using Example 3.6, the assertion of the theorem also holds for $c(x) = \|x_2 - x_1\|^p$. The more detailed estimate given in the proof of the theorem shows that the constant c_0 is at the same scale as c; in particular, it does not grow exponentially with c.

4 Proofs

4.1 Shadows and Preliminaries

For the convenience of the reader, we first recall the data processing inequality for our setting. Let Y_1 and Y_2 be Polish spaces. If $\mu \in \mathcal{P}(Y_1)$ and $K : Y_1 \to \mathcal{P}(Y_2)$ is a stochastic kernel, we

denote by $\mu K \in \mathcal{P}(Y_2)$ the second marginal of $\mu \otimes K \in \mathcal{P}(Y_1 \times Y_2)$. (4.1)

Lemma 4.1. *Let $\mu, \nu \in \mathcal{P}(Y_1)$ and $K : Y_1 \to \mathcal{P}(Y_2)$ a kernel. Then*

$$D_f(\mu K, \nu K) \le D_f(\mu, \nu).$$

Proof. We may assume that $\mu \ll \nu$. For any kernels $K_1 \ll K_2 : Y_1 \to \mathcal{P}(Y_2)$,

$$\frac{d(\mu \otimes K_1)}{d(\nu \otimes K_2)}(x, y) = \frac{d\mu}{d\nu}(x) \frac{dK_1(x)}{dK_2(x)}(y) \quad \nu \otimes K_2\text{-a.s.} \tag{4.2}$$

In particular, $\frac{d(\mu \otimes K)}{d(\nu \otimes K)}(x, y) = \frac{d\mu}{d\nu}(x)$ and thus

$$D_f(\mu, \nu) = D_f(\mu \otimes K, \nu \otimes K). \tag{4.3}$$

Whereas in general, (4.2) and Jensen's inequality for f yield

$$\begin{aligned} D_f(\mu \otimes K_1, \nu \otimes K_2) &= \iint f\left(\frac{d\mu}{d\nu}(x) \frac{dK_1(x)}{dK_2(x)}(y)\right) K_2(x, dy)\nu(dx) \\ &\ge \int f\left(\frac{d\mu}{d\nu}(x)\right) \nu(dx) = D_f(\mu, \nu). \end{aligned} \tag{4.4}$$

Denote by $\mu \otimes K = (\mu K) \otimes \tilde{K}_1$ and $\nu \otimes K = (\nu K) \otimes \tilde{K}_2$ the "reverse" disintegrations from the second marginal to the first. Applying (4.4) to $(\mu K) \otimes \tilde{K}_1$ and $(\nu K) \otimes \tilde{K}_2$,

$$D_f(\mu \otimes K, \nu \otimes K) = D_f((\mu K) \otimes \tilde{K}_1, (\nu K) \otimes \tilde{K}_2) \ge D_f(\mu K, \nu K).$$

In view of (4.3), this yields the claim. □

We can now show the two fundamental properties of the shadow.

Proof of Lemma 3.2. Let $\mu_i \otimes K_i \in \Pi(\mu_i, \tilde{\mu}_i)$ be a W_p-optimal coupling and define $\kappa = \pi \otimes K \in \mathcal{P}(X \times X)$ where $K(x) = K_1(x_1) \otimes \cdots \otimes K_N(x_N)$, so that $\tilde{\pi} := \pi K$ is the shadow of π. In view of $\kappa \in \Pi(\pi, \tilde{\pi})$, for $p < \infty$,

$$\begin{aligned} W_p(\pi, \tilde{\pi})^p &\le \int d_{X,p}(x, y)^p \, \kappa(dx, dy) \\ &= \int \sum_{i=1}^N d_{X_i}(x_i, y_i)^p \, \kappa(dx, dy) = \sum_{i=1}^N W_p(\mu_i, \tilde{\mu}_i)^p. \end{aligned}$$

On the other hand, given an arbitrary coupling $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \ldots, \tilde{\mu}_N)$, any coupling $\gamma \in \Pi(\pi, \tilde{\pi})$ induces couplings $\gamma_i \in \Pi(\pi_i, \tilde{\pi}_i) = \Pi(\mu_i, \tilde{\mu}_i)$ of the

individual marginals, hence

$$W_p(\pi,\tilde\pi)^p = \inf_{\gamma\in\Pi(\pi,\tilde\pi)} \int \sum_{i=1}^N d_{X_i}(x_i,y_i)^p\,\gamma(dx,dy)$$
$$\geq \sum_{i=1}^N \inf_{\gamma_i\in\Pi(\mu_i,\tilde\mu_i)} \int d_{X_i}(x_i,y_i)^p\,\gamma_i(dx_i,dy_i) = \sum_{i=1}^N W_p(\mu_i,\tilde\mu_i)^p.$$

The argument for $p=\infty$ is similar, completing the proof of the first claim. To show the bound on the divergence, note that $\tilde\mu_1\otimes\cdots\otimes\tilde\mu_N = (\mu_1\otimes\cdots\otimes\mu_N)K$. Therefore, the data processing inequality (Lemma 4.1) yields

$$D_f(\tilde\pi,\tilde\mu_1\otimes\cdots\otimes\tilde\mu_N) = D_f(\pi K,(\mu_1\otimes\cdots\otimes\mu_N)K) \leq D_f(\pi,\mu_1\otimes\cdots\otimes\mu_N).\ \square$$

Remark 4.2. The preceding proof shows that the shadow is a W_p-projection onto $\Pi(\tilde\mu_1,\dots,\tilde\mu_N)$; that is, $\tilde\pi \in \arg\min_{\Pi(\tilde\mu_1,\dots,\tilde\mu_N)} W_p(\pi,\cdot)$. In general, the argmin may have more than one element. A simple example on $\mathbb{R}\times\mathbb{R}$ is $\mu_1=\mu_2=\delta_0$ and $\tilde\mu_1=\tilde\mu_2=(\delta_{-1}+\delta_1)/2$; here any element of $\Pi(\tilde\mu_1,\tilde\mu_2)$ has the same distance to the singleton $\Pi(\mu_1,\mu_2)=\{\delta_{(0,0)}\}$. In this example, the shadow of $\pi := \delta_{(0,0)}$ is unique. Clearly, not any projection is a shadow, and most projections fail to satisfy the divergence bound in Lemma 3.2.

Next, we show the criteria for (A_L).

Proof of Lemma 3.5 and Example 3.4. To show the lemma, let $\kappa\in\Pi(\pi,\tilde\pi)$ be a coupling attaining $W_p(\pi,\tilde\pi)$. Then

$$\int c\,d(\pi-\tilde\pi) = \int c(x)-c(y)\,\kappa(dx,dy)$$
$$= \int f(x)(g(x)-g(y))\,\kappa(dx,dy) + \int g(y)(f(x)-f(y))\,\kappa(dx,dy). \quad (4.5)$$

We estimate the first integral; the second is treated analogously. Hölder's inequality with q such that $1/p+1/q=1$ yields

$$\int |f(x)(g(x)-g(y))|\,\kappa(dx,dy) \leq \|f\|_{L^q(\pi)}\|g(x)-g(y)\|_{L^p(\kappa)}.$$

As $|f(x)| \leq C_f[1+d_{X_1}(x_1,\bar x_1)^l+\cdots+d_{X_N}(x_N,\bar x_N)^l]$ with $l\leq p-1 = p(1-1/p) = p/q$ and hence $lq\leq p$, and as π has marginals $\mu_i\in\mathcal{P}_p(X_i)$, we see that $\|f\|_{L^q(\pi)}$ is finite with a bound depending only on the p-th moments of μ_i, $i=1,\dots,N$. On the other hand,

$$\|g(x)-g(y)\|_{L^p(\kappa)} \leq \mathrm{Lip}_p(g) W_p(\pi,\tilde\pi)$$

due to the fact that κ attains $W_p(\pi, \tilde{\pi})$. The lemma follows. Example 3.4 follows from the above estimate with $f(x) = g(x) = \|x_1 - x_2\|$ in which case $\mathrm{Lip}_2(f) = \mathrm{Lip}_2(g) = \sqrt{2}$. □

Remark 4.3. Lemma 3.5 can be generalized to a product of any finite number of Lipschitz functions. Let $c(x) = c_1(x) \cdots c_m(x)$ where c_j are Lipschitz and decompose $c(x) - c(y)$ as in (4.5) with $f(x) := c_1(x) \cdots c_{m-1}(x)$ and $g(x) := c_m(x)$. Proceeding inductively, we obtain that

$$c(x) - c(y) = \sum_{j=1}^{m} A_j(x, y)(c_j(x) - c_j(y))$$

where $A_j(x, y)$ is a product of $m-1$ factors of the form $c_k(x)$ or $c_l(y)$. If $c_j(x)$, $j = 1, \ldots, m$ satisfy a growth condition suitably coordinated with a moment condition on $\mu_i, \tilde{\mu}_i$, then $\|A_j(x, y)\|_{L^q(\pi)}$ and $\|A_j(x, y)\|_{L^q(\tilde{\pi})}$ can be bounded in terms of those moments and we deduce an analogue of Lemma 3.5.

Proof of Example 3.6. Let κ be a W_p-optimal coupling of π and $\tilde{\pi}$. Set $\psi(x) := \|x\|^p$ and define $\varphi : [0, 1] \to \mathbb{R}$ by

$$\varphi(t) := \int \psi((1-t)(x_2 - x_1) + t(y_2 - y_1))\, \kappa(dx, dy);$$

then $c(x) = \psi(x_2 - x_1)$ and the quantity to be estimated is

$$\left| \int c\, d\pi - \int c\, d\tilde{\pi} \right| = |\varphi(0) - \varphi(1)|. \tag{4.6}$$

Using differentiation under the integral (justified by [26, Theorem 2.27]), we see that φ is differentiable and

$$\frac{\partial \varphi}{\partial t}(t) = \int \big\langle \nabla\psi((1-t)(x_2 - x_1) + t(y_2 - y_1)), (y_2 - y_1 - x_2 + x_1) \big\rangle\, \kappa(dx, dy).$$

Noting $\|\nabla\psi(v)\| = p\|v\|^{p-1}$ and writing $v_t = (1-t)(x_2 - x_1) + t(y_2 - y_1)$, the inequalities of Cauchy–Schwarz, Hölder and $(a+b)^p \le 2^{p-1}(a^p + b^p)$ yield

$$\begin{aligned}
\left| \frac{\partial \varphi}{\partial t}(t) \right| &\le \int \|\nabla\psi(v_t)\| \|(y_2 - x_2) + (x_1 - y_1)\|\, \kappa(dx, dy) \\
&\le \left(\int \|\nabla\psi(v_t)\|^{\frac{p}{p-1}}\, \kappa(dx, dy) \right)^{\frac{p-1}{p}} \left(\int \|(y_2 - x_2) + (x_1 - y_1)\|^p\, \kappa(dx, dy) \right)^{\frac{1}{p}} \\
&\le C_p' \left(\int \|v_t\|^p\, \kappa(dx, dy) \right)^{\frac{p-1}{p}} W_p(\pi, \tilde{\pi}) \\
&\le C_p \big[M_p(\mu_1) + M_p(\tilde{\mu}_1) + M_p(\mu_2) + M_p(\tilde{\mu}_2) \big]^{p-1} W_p(\pi, \tilde{\pi})
\end{aligned}$$

where C_p, C_p' are constants depending only on p. In view of (4.6), the claim follows. □

4.2 Stability through Shadows

We can now show the continuity of the value.

Proof of Theorem 3.7. (i) Let π^*, π_n^* be the optimizers for $S(\mu_1, \dots, \mu_N, c)$ and $S(\mu_1^n, \dots, \mu_N^n, c)$, respectively. For brevity, set $P = \mu_1 \otimes \cdots \otimes \mu_N$ and $P_n = \mu_1^n \otimes \cdots \otimes \mu_N^n$. After passing to a subsequence, π_n converges in W_p to some $\pi \in \Pi(\mu_1, \dots, \mu_N)$, by weak compactness. We have

$$\liminf_{n\to\infty} \int c\, d\pi_n^* + D_f(\pi_n^*, P_n) \geq \int c\, d\pi + D_f(\pi, P) \geq \int c\, d\pi^* + D_f(\pi^*, P)$$

by lower semicontinuity of $\int c\, d(\cdot) + D_f(\cdot, \cdot)$ and optimality of π^*. On the other hand, let $\tilde{\pi}_n \in \Pi(\mu_1^n, \dots, \mu_N^n)$ be the shadow of π^*. Then Lemma 3.2 shows $\lim_n W_p(\tilde{\pi}_n, \pi^*) = 0$ and $D_f(\tilde{\pi}_n, P_n) \leq D_f(\pi^*, P)$, hence

$$\begin{aligned}\limsup_{n\to\infty} \int c\, d\pi_n^* + D_f(\pi_n^*, P_n) &\leq \limsup_{n\to\infty} \int c\, d\tilde{\pi}_n + D_f(\tilde{\pi}_n, P_n) \\ &\leq \int c\, d\pi^* + D_f(\pi^*, P).\end{aligned}$$

Together, $\lim_n \int c\, d\pi_n^* + D_f(\pi_n^*, P_n) = \int c\, d\pi^* + D_f(\pi^*, P)$ and π must be the (unique) optimizer π^* of $S(\mu_1, \dots, \mu_N, c)$. In particular, the original sequence (π_n^*) converges to π^*, as claimed.

(ii) Let π^* be the optimizer of $S(\mu_1, \dots, \mu_N, c)$ and $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ its shadow. Using (A_L) and Lemma 3.2,

$$\begin{aligned}S(\mu_1, \dots, \mu_N, c) &= \int c\, d\pi^* + D_f(\pi^*, \mu_1 \otimes \cdots \otimes \mu_N) \\ &\geq \int c\, d\tilde{\pi} - L W_p(\pi^*, \tilde{\pi}) + D_f(\tilde{\pi}, \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N) \\ &\geq S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) - L W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N).\end{aligned}$$

The claim follows by symmetry. □

The proof of Γ-convergence follows the same line of argument.

Proof of Remark 3.8. Similarly to the preceding proof, (a) follows from the lower semicontinuity of $\int c\, d(\cdot) + D_f(\cdot, \cdot)$. For (b), let π_n be the shadow of π and use Lemma 3.2 to obtain $\int c\, d\pi_n \to \int c\, d\pi$ and $D_f(\pi_n, \mu_1^n \otimes \cdots \otimes \mu_N^n) \leq D_f(\pi, \mu_1 \otimes \cdots \otimes \mu_N)$, again as in the preceding proof. □

The criteria for the transport inequality (I_q) are derived as follows.

Proof of Lemma 3.10. (i) For the convenience of the reader, we first recall the standard argument for bounded X: combine $d_{X,q}(x,y)^q \le \mathrm{diam}_q(X)^q\,\mathbf{1}_{x\neq y}$ with the transport representation of total variation distance [40, Lemma 2.20] and Pinsker's inequality [40, Theorem 2.16] to obtain

$$\begin{aligned} W_q(\pi,\theta)^q &= \inf_{\kappa\in\Pi(\pi,\theta)} \int d_{X,q}(x,y)^q\,\kappa(dx,dy) \\ &\le \mathrm{diam}_q(X)^q \inf_{\kappa\in\Pi(\pi,\theta)} \int \mathbf{1}_{x\neq y}\,\kappa(dx,dy) \\ &= \mathrm{diam}_q(X)^q \|\pi-\theta\|_{TV} \le \mathrm{diam}_q(X)^q \Big(\frac{1}{2} D_{\mathrm{KL}}(\theta,\pi)\Big)^{1/2}. \end{aligned}$$

The above holds for arbitrary probabilities π,θ. To prove the stronger estimate claimed in the lemma, we improve the above by exploiting that $\pi,\theta \in \Pi(\mu_1,\dots,\mu_N)$. Indeed, let $\Pi_1(\pi,\theta) \subset \Pi(\pi,\theta)$ denote the set of couplings $\kappa \in \Pi(\pi,\theta)$ not moving mass in the X_1-direction; i.e.,

$$\kappa\{(x_1,\dots,x_N,y_1,\dots,y_N) : x_1 = y_1\} = 1.$$

Note that $\Pi_1(\pi,\theta) \neq \emptyset$ due to the fact that π and θ have the same marginal μ_1 on X_1. Clearly

$$\begin{aligned} W_q(\pi,\theta)^q &= \inf_{\kappa\in\Pi(\pi,\theta)} \int d_{X,q}(x,y)^q\,\kappa(dx,dy) \\ &\le \inf_{\kappa\in\Pi_1} \int d_{X,q}(x,y)^q\,\kappa(dx,dy) \\ &\le M^q \inf_{\kappa\in\Pi_1(\pi,\theta)} \int \mathbf{1}_{x\neq y}\,\kappa(dx,dy), \quad M := \mathrm{diam}_q(X_2\times\cdots\times X_N). \end{aligned}$$

On the other hand, we claim that π,θ having the same marginal implies

$$\inf_{\kappa\in\Pi_1(\pi,\theta)} \int \mathbf{1}_{x\neq y}\,\kappa(dx,dy) \le \|\pi-\theta\|_{TV}; \tag{4.7}$$

in words, where mass needs to be moved, one might as well move only in the directions $X_2,\dots,X_N$. Granted (4.7), we can proceed as in the beginning and conclude the assertion of the lemma,

$$W_q(\pi,\theta)^q \le M^q \|\pi-\theta\|_{TV} \le M^q \Big(\frac{1}{2} D_{\mathrm{KL}}(\theta,\pi)\Big)^{1/2}.$$

To show (4.7), consider the mutually singular measures $\tilde{\pi} = \pi - (\pi \wedge \theta)$ and $\tilde{\theta} = \theta - (\pi \wedge \theta)$, where $\pi \wedge \theta$ is defined as usual via $d(\pi \wedge \theta)/d(\pi + \theta) = \min\{d\pi/d(\pi+\theta), d\theta/d(\pi+\theta)\}$. These measures again share a common first marginal, so that $\Pi_1(\tilde{\pi}, \tilde{\theta}) \neq \emptyset$. Let $\tilde{\kappa} \in \Pi_1(\tilde{\pi}, \tilde{\theta})$ be arbitrary and let $\kappa \in \Pi(\pi, \theta)$ be the coupling given by $\kappa = \tilde{\kappa} + i$ where i is the identical coupling of $\pi \wedge \theta$ with itself. Then

$$\|\pi - \theta\|_{TV} \leq \int \mathbf{1}_{x \neq y} \, \kappa(dx, dy) = \int \mathbf{1}_{x \neq y} \, \tilde{\kappa}(dx, dy) = \|\tilde{\pi} - \tilde{\theta}\|_{TV}$$

where the last equality follows from mutual singularity. As $\|\tilde{\pi} - \tilde{\theta}\|_{TV} = \|\pi - \theta\|_{TV}$, all expressions are equal and (4.7) follows.

(ii) It is shown in [9, Corollary 2.4] that the inequality (I_q) holds for a given measure $\pi \in \mathcal{P}(X)$ and all $\theta \in \mathcal{P}(X)$ whenever

$$\int \exp(\tilde{\alpha} \, d_{X,q}(x, \hat{x})^{2q}) \, \pi(dx) < \infty \tag{4.8}$$

for some $\tilde{\alpha} > 0$ and $\hat{x} \in X$, with constant

$$C_{\pi,q} = 2 \inf_{\hat{x} \in X, \tilde{\alpha} > 0} \left(\frac{1}{2\tilde{\alpha}} \Big(1 + \log \int \exp(\tilde{\alpha} d_{X,q}(\hat{x}, x)^{2q}) \pi(dx) \Big) \right)^{\frac{1}{2q}}. \tag{4.9}$$

To obtain the claim for a coupling π (and general $\theta \in \mathcal{P}(X)$), note that

$$d_{X,q}(\hat{x}, x)^{2q} \leq N \sum_{i=1}^{N} d_{X,i}(\hat{x}_i, x_i)^{2q} = \frac{1}{N} \sum_{i=1}^{N} N^2 d_{X,i}(\hat{x}_i, x_i)^{2q}$$

and that the functional $f \mapsto \log \int \exp(\tilde{\alpha} f(x)) \, \pi(dx)$, is convex (as can be seen from a variational representation, e.g. [28, Example 4.34, p. 201]). Hence

$$\log \int \exp(\tilde{\alpha} d_{X,q}(\hat{x}, x)^{2q}) \, \pi(dx) \leq \frac{1}{N} \sum_{i=1}^{N} \log \int \exp(\tilde{\alpha} N^2 d_{X_i}(\hat{x}_i, x_i)^{2q}) \, \mu_i(dx_i).$$

To obtain the claim for C_q, we plug this inequality into (4.9) and set $\tilde{\alpha} = \alpha/N^2$. Similarly, the integrability condition in the lemma implies (4.8).

(iii) The proof is similar to (ii) but refers to a different result of [9]. Indeed, by [9, Corollary 2.3], it suffices to bound

$$C'_{\pi,q} = 2 \inf_{\hat{x} \in X, \tilde{\alpha} > 0} \left(\frac{1}{\tilde{\alpha}} \Big(\frac{3}{2} + \log \int \exp(\tilde{\alpha} d_{X,q}(\hat{x}, x)^{q}) \pi(dx) \Big) \right)^{\frac{1}{q}}.$$

Here the term inside the exponential already factorizes and we can directly apply the convexity of $f \mapsto \log \int \exp(\tilde{\alpha} f(x)) \pi(dx)$, which yields the claim after the substitution $\tilde{\alpha} = \alpha/N$. $\square$

As a preparation for the proof of Theorem 3.11, we recall a Pythagorean relation for the entropic optimal transport problem. We denote

$$\mathcal{F}(\pi) = \int c\, d\pi + D_{\mathrm{KL}}(\pi, \pi_1 \otimes \cdots \otimes \pi_N)$$

where $\pi_1, \dots, \pi_N$ are the marginals of π.

Lemma 4.4. *If $\pi^* \in \Pi(\mu_1, \dots, \mu_N)$ is the optimizer of $S(\mu_1, \dots, \mu_N, c)$,*

$$D_{\mathrm{KL}}(\pi, \pi^*) \le \mathcal{F}(\pi) - \mathcal{F}(\pi^*) \quad \textit{for all} \quad \pi \in \Pi(\mu_1, \dots, \mu_N).$$

Proof. Set $P = \mu_1 \otimes \cdots \otimes \mu_N$ and define $P_c \in \mathcal{P}(X)$ by $dP_c = \alpha^{-1} e^{-c}\, dP$, where α is the normalizing constant. Then

$$\mathcal{F}(\pi) = D_{\mathrm{KL}}(\pi, P_c) - \log \alpha, \tag{4.10}$$

so that the entropic optimal transport problem (2.3) is equivalent to minimizing $D_{\mathrm{KL}}(\cdot, P_c)$. In particular, $\pi^* = \arg\min_{\Pi(\mu_1,\dots,\mu_N)} D_{\mathrm{KL}}(\cdot, P_c)$ and the Pythagorean theorem for relative entropy [19, Theorem 2.2] yields

$$D_{\mathrm{KL}}(\pi, P_c) \ge D_{\mathrm{KL}}(\pi^*, P_c) + D_{\mathrm{KL}}(\pi, \pi^*) \quad \text{for all} \quad \pi \in \Pi(\mu_1, \dots, \mu_N).$$

In view of (4.10), the claim follows. (In the case under consideration, the assertion holds even with equality. We do not need that fact here.) □

We can now show the stability of optimizers with respect to the marginals.

Proof of Theorem 3.11. We detail the proof for (I_q); the argument for (I'_q) is identical. For notational convenience, we treat the case where $\tilde{\mu}_i$ (rather than μ_i) satisfy (I_q). Consider the optimizers $\pi^* \in \Pi(\mu_1, \dots, \mu_N)$ and $\tilde{\pi}^* \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$. Let $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$ be the shadow of π^* for the p-Wasserstein distance. Using Lemma 3.2 and (A_L) as in the proof of Theorem 3.7 (ii),

$$\mathcal{F}(\tilde{\pi}) - \mathcal{F}(\pi^*) \le \int c\, d(\tilde{\pi} - \pi^*) \le L\, W_p(\tilde{\pi}, \pi^*) \le L\Delta.$$

We also have $\mathcal{F}(\pi^*) - \mathcal{F}(\tilde{\pi}^*) \le L\Delta$ by Theorem 3.7 (ii), and adding the inequalities yields

$$\mathcal{F}(\tilde{\pi}) - \mathcal{F}(\tilde{\pi}^*) \le 2L\Delta.$$

(If both marginals satisfy (I_q) with constant L, we can assume by symmetry that $\mathcal{F}(\pi^*) - \mathcal{F}(\tilde{\pi}^*) \le 0$, and obtain the estimate with L instead of $2L$.) By Lemma 4.4, it follows that $D_{\mathrm{KL}}(\tilde{\pi}, \tilde{\pi}^*) \le 2L\Delta$, and now (I_q) implies

$$W_q(\tilde{\pi}, \tilde{\pi}^*) \le C_q (2L\Delta)^{\frac{1}{2q}}.$$

Recalling that W_r on X was defined relative to the distance $d_{X,r}$, Jensen's inequality implies $W_q(\cdot,\cdot) \leq N^{(\frac{1}{q}-\frac{1}{p})} W_p(\cdot,\cdot)$. In view of Lemma 3.2, we deduce $W_q(\pi^*,\tilde{\pi}) \leq N^{(\frac{1}{q}-\frac{1}{p})} W_p(\pi^*,\tilde{\pi}) \leq N^{(\frac{1}{q}-\frac{1}{p})}\Delta$. We conclude the proof via the triangle inequality,

$$W_q(\pi^*,\tilde{\pi}^*) \leq W_q(\pi^*,\tilde{\pi}) + W_q(\tilde{\pi},\tilde{\pi}^*) \leq N^{(\frac{1}{q}-\frac{1}{p})}\Delta + C_q\,(2L\Delta)^{\frac{1}{2q}}\,. \qquad \square$$

4.3 Stability with respect to Cost

Throughout this section, we fix $p \in [1,\infty]$, $\mu_i \in \mathcal{P}_p(X_i)$ for $i = 1,\dots,N$ and $c, \tilde{c} : X \to [0,\infty)$ satisfying the growth condition (2.2). The following observation is the starting point for the stability with respect to the cost function. We recall that $P := \mu_1 \otimes \cdots \otimes \mu_N$.

Lemma 4.5. *Let $\pi^*, \tilde{\pi}^*$ be the respective optimizers of $S_{\mathrm{ent}}(\mu_1,\dots,\mu_N,c)$ and $S_{\mathrm{ent}}(\mu_1,\dots,\mu_N,\tilde{c})$. Then*

$$D_{\mathrm{KL}}(\pi^*,\tilde{\pi}^*) + D_{\mathrm{KL}}(\tilde{\pi}^*,\pi^*) \leq \int (c-\tilde{c})\, d(\tilde{\pi}^* - \pi^*).$$

Proof. Lemma 4.4 yields

$$\begin{aligned} D_{\mathrm{KL}}(\pi^*,\tilde{\pi}^*) + D_{\mathrm{KL}}(\tilde{\pi}^*,\pi^*) &\leq \int c\, d\tilde{\pi}^* + D_{\mathrm{KL}}(\tilde{\pi}^*,P) + \int \tilde{c}\, d\pi^* + D_{\mathrm{KL}}(\pi^*,P) \\ &\quad - \int c\, d\pi^* - D_{\mathrm{KL}}(\pi^*,P) - \int \tilde{c}\, d\tilde{\pi}^* - D_{\mathrm{KL}}(\tilde{\pi}^*,P) \\ &= \int (c-\tilde{c})\, d(\tilde{\pi}^* - \pi^*). \end{aligned}$$

$\square$

Lemma 4.5 clearly implies a Lipschitz estimate with respect to $\|c-\tilde{c}\|_\infty$ by using Pinsker's inequality on the left-hand side. The following proof is a variation on that observation.

Proof of Proposition 3.12. Combining

$$\int (\tilde{c}-c)\, d(\pi^* - \tilde{\pi}^*) \leq \int |\tilde{c}-c| \left| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right| dP$$

with Hölder's inequality as well as (in case $p \neq 1$), for $q := \frac{p}{p-1}$,

$$\left| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right|^q \leq \left\| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right\|_{L^\infty(P)}^{q-1} \left| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right|,$$

yields

$$\int (\tilde{c} - c)\, d(\pi^* - \tilde{\pi}^*) \le \|\tilde{c} - c\|_{L^p(P)} (2\|\pi^* - \tilde{\pi}^*\|_{TV})^{1-\frac{1}{p}} \left\| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right\|_\infty^{\frac{1}{p}}. \tag{4.11}$$

Next, we show

$$\left\| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right\|_\infty \le a := \exp(N\|c\|_\infty) + \exp(N\|\tilde{c}\|_\infty). \tag{4.12}$$

Recall that by duality (e.g., [23, 44]), for certain "potentials" $\varphi_i : X_i \to \mathbb{R}$,

$$\frac{d\pi^*}{dP}(x) = \exp\left(-c + \oplus_i \varphi_i\right) \tag{4.13}$$

where $(\oplus_i \varphi_i)(x) := \sum_{i=1}^N \varphi(x_i)$, and moreover

$$\int \oplus_i \varphi_i \, dP = S_{\text{ent}}(\mu_1, \dots, \mu_N, c) \ge 0 \tag{4.14}$$

where the inequality is due to $c \ge 0$. To estimate the right-hand side of (4.13), recall that (4.13) and the fact that π^* is a coupling imply a conjugacy relation between the potentials (e.g., [23, 44, 45]), namely

$$\begin{aligned} \varphi_i(x_i) &= -\log \int \exp(-c(x) + \oplus_{j\ne i}\varphi_j(x_j))\, P_{-i}(dx_{-i}) \\ &\le \|c\|_\infty - \int \oplus_{j\ne i}\varphi_j \, dP_{-i}, \end{aligned}$$

where $x_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ and $P_{-i} := \otimes_{j\ne i}\mu_j$. Thus by (4.14),

$$\oplus_i \varphi_i(x) \le N\|c\|_\infty - (N-1)\int \oplus_{j=1}^N \varphi_j \, dP \le N\|c\|_\infty.$$

Using this in (4.13), we conclude that

$$\left\| \frac{d\pi^*}{dP} \right\|_\infty \le \exp(N\|c\|_\infty).$$

The analogue holds for $\tilde{\pi}^*$, hence $\left\| \frac{d\pi^*}{dP} - \frac{d\tilde{\pi}^*}{dP} \right\|_\infty \le \left\| \frac{d\pi^*}{dP} \right\|_\infty + \left\| \frac{d\tilde{\pi}^*}{dP} \right\|_\infty \le a$ as claimed in (4.12).

Pinsker's inequality, Lemma 4.5, (4.11) and (4.12) imply

$$\begin{aligned} 4\|\pi^* - \tilde{\pi}^*\|_{TV}^2 &\le D_{\text{KL}}(\pi^*, \tilde{\pi}^*) + D_{\text{KL}}(\tilde{\pi}^*, \pi^*) \\ &\le \int (c - \tilde{c})\, d(\tilde{\pi}^* - \pi^*) \le a^{\frac{1}{p}} (2\|\pi^* - \tilde{\pi}^*\|_{TV})^{1-\frac{1}{p}} \|\tilde{c} - c\|_{L^p(P)}. \end{aligned}$$

Dividing by $4\|\pi^* - \tilde{\pi}^*\|_{TV}^{1-\frac{1}{p}}$ yields

$$\|\pi^* - \tilde{\pi}^*\|_{TV}^{1+\frac{1}{p}} \leq \left(\frac{1}{2}\right)^{1+\frac{1}{p}} a^{\frac{1}{p}} \|\tilde{c} - c\|_{L^p(P)} \tag{4.15}$$

which is the first claim of the proposition. On the other hand, using Lemma 4.5 and (4.11) together with (4.15) yields

$$D_{\mathrm{KL}}(\pi^*, \tilde{\pi}^*) + D_{\mathrm{KL}}(\tilde{\pi}^*, \pi^*) \leq a^{\frac{1}{p}} \|\tilde{c} - c\|_{L^p(P)} \left(a^{\frac{1}{p}} \|\tilde{c} - c\|_{L^p(P)}\right)^{\frac{p-1}{p+1}}. \tag{4.16}$$

As (I_q) implies $2C_q^{-2q} W_q(\pi^*, \tilde{\pi}^*)^{2q} \leq D_{\mathrm{KL}}(\pi^*, \tilde{\pi}^*) + D_{\mathrm{KL}}(\tilde{\pi}^*, \pi^*)$, this proves the second claim of the proposition. For the last claim, we drop the nonnegative term $D_{\mathrm{KL}}(\tilde{\pi}^*, \pi^*)$ on the left-hand side of (4.16) and use (I_q') with the remaining inequality. □

4.4 Stability through Transformation

Let $p \in [1, \infty]$, $\mu_i, \tilde{\mu}_i \in \mathcal{P}_p(X_i)$ for $i = 1, \ldots, N$ and let $c : X \to [0, \infty)$ satisfy the growth condition (2.2). We begin with preliminary results connecting stability with respect to the marginals and stability with respect to the cost function. As in Definition 3.1, K denotes the kernel $K(x) = K_1(x_1) \otimes \cdots \otimes K_N(x_N)$, where $\mu_i \otimes K_i \in \Pi(\mu_i, \tilde{\mu}_i)$ is an optimal coupling attaining $W_p(\mu_i, \tilde{\mu}_i)$. We use the notation $Kc(x) := \int c(y)\, K(x, dy)$ for the integral of c with respect to the kernel.

Lemma 4.6. *Let $p \in [1, \infty]$ and let c be $\mathrm{Lip}_p(c)$-Lipschitz. Then*

$$\|c - Kc\|_{L^p(\pi)} \leq \mathrm{Lip}_p(c) W_p(\mu_1, \ldots, \mu_N; \tilde{\mu}_1, \ldots, \tilde{\mu}_N), \quad \pi \in \Pi(\mu_1, \ldots, \mu_N).$$

Proof. We only detail the calculation for $p < \infty$,

$$\begin{aligned}
\|c - Kc\|_{L^p(\pi)}^p &= \int \left| c(x) - \int c(y) K(x, dy) \right|^p \pi(dx) \\
&\leq \iint |c(x) - c(y)|^p K(x, dy) \pi(dx) \\
&\leq \mathrm{Lip}_p(c)^p \iint \sum_{i=1}^N d_{X,i}(x_i, y_i)^p K(x, dy) \pi(dx) \\
&= \mathrm{Lip}_p(c)^p \sum_{i=1}^N W_p(\mu_i, \tilde{\mu}_i)^p.
\end{aligned}$$

□

Next, consider the kernel $\tilde{K}$ defined like K but with the marginals reversed; that is, $\tilde{K}(x) = \tilde{K}_1(x_1) \otimes \cdots \otimes \tilde{K}_N(x_N)$, where $\tilde{\mu}_i \otimes \tilde{K}_i \in \Pi(\tilde{\mu}_i, \mu_i)$ is an optimal coupling attaining $W_p(\tilde{\mu}_i, \mu_i)$. The double integral $\tilde{K}Kc := \tilde{K}(Kc)$ thus corresponds to a round-trip between the marginals. In general, this round-trip leads to a positive gap R in value, as shown in the next result. The result will not be used in the subsequent proofs but it may be useful to understand the steps below, where we look for situations where the gap is zero.

Lemma 4.7. *Let $p \in [1, \infty]$. We have*

$$S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) \le S(\mu_1, \dots, \mu_N, Kc) \le S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) + R,$$

where $R := \int (\tilde{K}Kc - c)\, d\tilde{\pi}^*$ and $\tilde{\pi}^*$ is the optimizer of $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$. Moreover, $R \le 2 \operatorname{Lip}_p(c) W_p(\mu_1, \dots, \mu_N; \tilde{\mu}_1, \dots, \tilde{\mu}_N)$.

Proof. Set $\tilde{P} = \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N$ and recall (4.1). Using Lemma 4.1 twice,

$$\begin{aligned}
S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) &= \inf_{\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)} \int c\, d\tilde{\pi} + D_f(\tilde{\pi}, \tilde{P}) \\
&\le \inf_{\pi \in \Pi(\mu_1, \dots, \mu_N)} \int c\, d(\pi K) + D_f(\pi K, PK) \\
&\le \inf_{\pi \in \Pi(\mu_1, \dots, \mu_N)} \int Kc\, d\pi + D_f(\pi, P) \\
&= S(\mu_1, \dots, \mu_N, Kc) \\
&\le \int Kc\, d(\tilde{\pi}^* \tilde{K}) + D_f(\tilde{\pi}^* \tilde{K}, \tilde{P}\tilde{K}) \\
&\le \int \tilde{K}Kc\, d\tilde{\pi}^* + D_f(\tilde{\pi}^*, \tilde{P}) = S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) + R.
\end{aligned}$$

The bound for R is similar to the proof of Lemma 4.6. □

In Lemma 4.7, there is a gap between the values of $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$ and $S(\mu_1, \dots, \mu_N, Kc)$. If however the kernels $K, \tilde{K}$ are given by maps inverse to one another (as will be the case in the proof of Lemma 4.9 below), the gap is zero and the problems $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$ and $S(\mu_1, \dots, \mu_N, Kc)$ become equivalent in the following sense. We write $T_\sharp$ for the pushforward under T.

Lemma 4.8. *For $i = 1, \dots, N$, let $T_i : X_i \to X_i$ satisfy $\tilde{\mu}_i = (T_i)_\sharp \mu_i$ and admit a (measurable) a.s. inverse $T_i^{-1} : X_i \to X_i$; that is, $T_i^{-1} \circ T_i = \mathrm{id}$ μ_i-a.s. and $T_i \circ T_i^{-1} = \mathrm{id}$ $\tilde{\mu}_i$-a.s. Define*

$$T(x) = (T_1(x_1), \dots, T_N(x_N)), \quad T^{-1}(x) = (T_1^{-1}(x_1), \dots, T_N^{-1}(x_N)).$$

Then $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) = S(\mu_1, \dots, \mu_N, c \circ T)$ *and the optimizers* $\tilde{\pi}^*$, π^* *of the two problems are related by* $\tilde{\pi}^* = T_\sharp \pi^*$ *and* $\pi^* = T_\sharp^{-1} \tilde{\pi}^*$.

Proof. Set $P = \mu_1 \otimes \cdots \otimes \mu_N$ and $\tilde{P} = \tilde{\mu}_1 \otimes \cdots \otimes \tilde{\mu}_N$. We have

$$\begin{aligned} \int c \circ T \, d\pi + D_f(\pi, P) &= \int c \circ T \, d(T_\sharp^{-1}(T_\sharp \pi)) + D_f(T_\sharp^{-1}(T_\sharp \pi), T_\sharp^{-1} \tilde{P}) \\ &= \int c \, d(T_\sharp \pi) + D_f(T_\sharp \pi, \tilde{P}) \end{aligned}$$

for any $\pi \in \Pi(\mu_1, \dots, \mu_N)$, hence taking infimum over $\pi \in \Pi(\mu_1, \dots, \mu_N)$ yields $S(\mu_1, \dots, \mu_N, c \circ T) \geq S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$. Symmetric results hold starting from $\tilde{\pi} \in \Pi(\tilde{\mu}_1, \dots, \tilde{\mu}_N)$. Thus $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c) = S(\mu_1, \dots, \mu_N, c \circ T)$, and now the formulas for the optimizers follow as well. □

In the simplest case, the optimal couplings for $W_p(\mu_i, \tilde{\mu}_i)$ are given by invertible maps, and then we can apply Lemma 4.8 directly to prove Theorem 3.13. In general, we approximate the marginals with measures having that property as detailed next, passing to an augmented space to guarantee that the setting is sufficiently rich. We write δ_x for the Dirac measure at x.

Lemma 4.9. *Let* $p \in [1, \infty]$. *Let* $\bar{X}_i = X_i \times (-1, 1)$ *and embed the marginals as* $\nu_i := \mu_i \otimes \delta_0$ *and* $\tilde{\nu}_i := \tilde{\mu}_i \otimes \delta_0$ *for* $i = 1, \dots, N$. *Set* $\bar{X} = \prod_{i=1}^N \bar{X}_i$ *and define* $\bar{c} : \bar{X} \to \mathbb{R}$ *by* $\bar{c}(x, u) := c(x)$ *for* $x \in X$ *and* $u \in (-1, 1)^N$.

(i) We have $S(\mu_1, \dots, \mu_N, c) = S(\nu_1, \dots, \nu_N, \bar{c})$ *and the corresponding optimizers* π, θ *are related by* $\theta = \pi \otimes \delta_0^N$.

If $\tilde{\pi}, \tilde{\theta}$ *are the optimizers for* $S(\tilde{\mu}_1, \dots, \tilde{\mu}_N, c)$ *and* $S(\tilde{\nu}_i, \dots, \tilde{\nu}_N, \bar{c})$, *then*

$$W_p(\pi, \tilde{\pi}) = W_p(\theta, \tilde{\theta}).$$

(ii) Given $0 < \epsilon < 1$ *and* $i = 1, \dots, N$, *there exist* $\nu_i^\epsilon, \tilde{\nu}_i^\epsilon \in \mathcal{P}(\bar{X}_i)$ *with*

$$W_p(\nu_i, \nu_i^\epsilon) \leq \epsilon, \qquad W_p(\tilde{\nu}_i, \tilde{\nu}_i^\epsilon) \leq \epsilon \tag{4.17}$$

and an a.s. invertible map $T_i^\epsilon : \bar{X}_i \to \bar{X}_i$ *such that* $\tilde{\nu}_i^\epsilon = (T_i^\epsilon)_\sharp \nu_i^\epsilon$ *and the corresponding coupling attains* $W_p(\nu_i^\epsilon, \tilde{\nu}_i^\epsilon)$.

Proof. (i) follows immediately from the definitions; we prove (ii). The case $p < \infty$ is standard: for n large enough, there exist $\rho_i, \tilde{\rho}_i \in \mathcal{P}(\bar{X}_i)$ of the form

$$\rho_i = \frac{1}{n} \sum_{k=1}^n \delta_{(x_k, 0)}, \qquad \tilde{\rho}_i = \frac{1}{n} \sum_{k=1}^n \delta_{(\tilde{x}_k, 0)}$$

such that $W_p(\nu_i, \rho_i) \leq \frac{\epsilon}{2}$ and $W_p(\tilde{\nu}_i, \tilde{\rho}_i) \leq \frac{\epsilon}{2}$; for instance, one can use suitable realizations of i.i.d. samples (e.g., [35, Corollary 1.1]). Next, choose distinct $u_1, \ldots, u_n \in (0,1)$ small enough such that the measures

$$\nu_i^\epsilon = \frac{1}{n}\sum_{k=1}^n \delta_{(x_k,u_k)}, \qquad \tilde{\nu}_i^\epsilon = \frac{1}{n}\sum_{k=1}^n \delta_{(\tilde{x}_k,u_k)}$$

satisfy $W_p(\rho_i, \nu_i^\epsilon) \leq \frac{\epsilon}{2}$ and $W_p(\tilde{\rho}_i, \tilde{\nu}_i^\epsilon) \leq \frac{\epsilon}{2}$. Then (4.17) holds and $\nu_i^\epsilon, \tilde{\nu}_i^\epsilon$ are empirical measures on n distinct points due to the choice of $u_1, \ldots, u_n$. As a result, there is an optimal transport map that is one-to-one on the supports.

Let $p = \infty$. Here a different argument is necessary. (The following also gives an alternate proof for $p < \infty$.) As X is Polish, we can find a dense sequence $(q_k) \subset X$ and a countable measurable partition (Q_k) of X with $q_k \in Q_k$ and $\operatorname{diam} Q_k \leq \frac{\epsilon}{4}$. Consider the approximations

$$\rho_i := \sum_{k=1}^\infty \nu_i(Q_k)\, \delta_{q_k} \otimes \delta_0, \qquad \tilde{\rho}_i := \sum_{k=1}^\infty \tilde{\nu}_i(Q_k)\, \delta_{q_k} \otimes \delta_0$$

which clearly satisfy $W_\infty(\rho_i, \nu_i) < \frac{\epsilon}{2}$ and $W_\infty(\tilde{\rho}_i, \tilde{\nu}_i) < \frac{\epsilon}{2}$, but may have atoms of unequal mass. Let $\rho_i \otimes U_i \in \Pi(\rho_i, \tilde{\rho}_i)$ be a W_∞-optimal coupling, then $U_i : \bar{X}_i \to \mathcal{P}(\bar{X}_i)$ is a stochastic kernel such that for each k,

$$U_i((q_k, 0)) = \sum_{j=1}^\infty w_{j,k}\, \delta_{q_j} \otimes \delta_0,$$

for some weights $w_{j,k} \geq 0$ with $\sum_{j=1}^\infty w_{j,k} = 1$. Let $\epsilon_0 > 0$ and pick disjoint numbers $u_{j,k} \in (0, \epsilon_0)$, define

$$\nu_i^\epsilon := \sum_{j,k=1}^\infty \nu_i(Q_k) w_{j,k} \delta_{q_k} \otimes \delta_{u_{j,k}}, \qquad \tilde{\nu}_i^\epsilon := \sum_{j,k=1}^\infty \nu_i(Q_k) w_{j,k}\, \delta_{q_j} \otimes \delta_{u_{j,k}}$$

and observe that $W_\infty(\nu_i^\epsilon, \rho_i) < \frac{\epsilon}{2}$ and $W_\infty(\tilde{\nu}_i^\epsilon, \tilde{\rho}_i) < \frac{\epsilon}{2}$ for ϵ_0 sufficiently small (note that $u_{j,k} := 0$ would lead to $\nu_i^\epsilon = \rho_i$ and $\tilde{\nu}_i^\epsilon = \nu_i^\epsilon U_i = \tilde{\rho}_i$). Now (4.17) holds by the triangle inequality. Define

$$T_i^\epsilon : \{q_k : k \in \mathbb{N}\} \times \{u_{j,k} : j,k \in \mathbb{N}\} \to \{q_k : k \in \mathbb{N}\} \times \{u_{j,k} : j,k \in \mathbb{N}\},$$

$$T_i^\epsilon(q_k, u_{j,k}) := (q_j, u_{j,k})$$

which is one-to-one as the $u_{j,k}$ are distinct. Moreover, $\rho_i \otimes U_i \in \Pi(\rho_i, \tilde{\rho}_i)$ implies $\tilde{\nu}_i^\epsilon = (T_i^\epsilon)_\sharp \nu_i^\epsilon$, and since $\rho_i \otimes U_i$ attains $W_\infty(\rho_i, \tilde{\rho}_i) = W_\infty(\nu_i^\epsilon, \tilde{\nu}_i^\epsilon)$, the coupling induced by T_i^ϵ attains $W_\infty(\nu_i^\epsilon, \tilde{\nu}_i^\epsilon)$. □

After these preparations, we are ready to prove Theorem 3.13.

Proof of Theorem 3.13. We detail the proof for (I_q); the argument for (I'_q) is identical. We shall apply Proposition 3.12 though the equivalence outlined in Lemma 4.8. To this end, we extend the spaces X_i by the interval $(-1,1)$ and introduce $\nu_i, \tilde{\nu}_i, \bar{c}$ as in Lemma 4.9. In view of Lemma 4.9 (i), it suffices to prove the claim for these data instead of $\mu_i, \tilde{\mu}_i, c$.

Let $\epsilon > 0$, choose $\nu_i^\epsilon, \tilde{\nu}_i^\epsilon, T^\epsilon$ as in Lemma 4.9 (ii) and denote by $\theta^\epsilon, \tilde{\theta}^\epsilon, \hat{\theta}^\epsilon$ the respective optimizers of $S_{\mathrm{ent}}(\nu_1^\epsilon, \dots, \nu_N^\epsilon, \bar{c})$ and $S_{\mathrm{ent}}(\tilde{\nu}_1^\epsilon, \dots, \tilde{\nu}_N^\epsilon, \bar{c})$ and $S_{\mathrm{ent}}(\nu_1^\epsilon, \dots, \nu_N^\epsilon, \bar{c} \circ T^\epsilon)$, respectively. Noting that $\mathrm{Lip}_p(\bar{c}) = \mathrm{Lip}_p(c)$ and setting $\Delta(\epsilon) := W_p(\nu_1^\epsilon, \dots, \nu_N^\epsilon; \tilde{\nu}_1^\epsilon, \dots, \nu_N^\epsilon)$, Lemma 4.6 yields

$$\|\bar{c} - \bar{c} \circ T^\epsilon\|_{L^p(P)} \le \mathrm{Lip}_p(c)\, \Delta(\epsilon)$$

and thus Proposition 3.12 shows that

$$W_q(\hat{\theta}^\epsilon, \theta^\epsilon) \le C_q \left(\frac{1}{2}\right)^{\frac{1}{2q}} \left(a^{\frac{1}{p}} \mathrm{Lip}_p(c)\, \Delta(\epsilon)\right)^{\frac{p}{(p+1)q}}. \tag{4.18}$$

As $\tilde{\theta}^\epsilon = T^\epsilon{}_\sharp \hat{\theta}^\epsilon$ by Lemma 4.8 and T_i^ϵ attains $W_p(\nu_i^\epsilon, \tilde{\nu}_i^\epsilon)$, it follows by the same calculation as in the proof of Theorem 3.11 that

$$W_q(\tilde{\theta}^\epsilon, \hat{\theta}^\epsilon) \le N^{(\frac{1}{q} - \frac{1}{p})} W_p(\tilde{\theta}^\epsilon, \hat{\theta}^\epsilon) \le N^{(\frac{1}{q} - \frac{1}{p})} \Delta(\epsilon).$$

Combining the two estimates, we find that

$$\begin{aligned} W_q(\theta^\epsilon, \tilde{\theta}^\epsilon) &\le W_q(\theta^\epsilon, \hat{\theta}^\epsilon) + W_q(\hat{\theta}^\epsilon, \tilde{\theta}^\epsilon) \\ &\le N^{(\frac{1}{q} - \frac{1}{p})} \Delta(\epsilon) + C_q \left(\frac{1}{2}\right)^{\frac{1}{2q}} \left(a^{\frac{1}{p}} \mathrm{Lip}_p(c)\, \Delta(\epsilon)\right)^{\frac{p}{(p+1)q}}. \end{aligned}$$

Letting $\epsilon \to 0$, the left-hand side converges to $W_q(\pi^*, \tilde{\pi}^*)$ by Theorem 3.11 and Lemma 4.9 (ii), while $\Delta(\epsilon) \to \Delta$ by construction. The claim on sharpness is discussed in Example 4.10 below. □

Finally, we exhibit a family of examples for which the constant ℓ of Theorem 3.13 is optimal.

Example 4.10 (Sharpness of ℓ in Theorem 3.13.)**.** On $X = [-1,1]^2$, let

$$\mu_1 = \mu_2 = \frac{1}{2}(\delta_{-1} + \delta_1), \qquad \tilde{\mu}_1 = \tilde{\mu}_2 = \frac{1}{2}(\delta_{-1+\varepsilon} + \delta_{1-\varepsilon}),$$

where $\varepsilon \in (0, 1/2)$ is a parameter. We define the cost function $c = c(\varepsilon)$ by

$$\begin{aligned} c(-1,-1) &= c(1,1) = c(-1+\varepsilon, 1-\varepsilon) = c(1-\varepsilon, -1+\varepsilon) = 0, \\ c(1,-1) &= c(-1,1) = c(-1+\varepsilon, -1+\varepsilon) = c(1-\varepsilon, 1-\varepsilon) = \varepsilon, \end{aligned}$$

then c is Lipschitz with constant $\mathrm{Lip}_\infty(c) = 1$. Setting $\alpha(\varepsilon) := \frac{\exp(\varepsilon)}{1+\exp(\varepsilon)}$, we calculate the optimizers $\pi^*, \tilde{\pi}^*$ of $S_{\mathrm{ent}}(\mu_1, \mu_2, c)$ and $S_{\mathrm{ent}}(\tilde{\mu}_1, \tilde{\mu}_2, c)$ to be

$$\begin{aligned} \pi^* &= \frac{\alpha(\epsilon)}{2}\big(\delta_{(-1,-1)} + \delta_{(1,1)}\big) + \frac{1-\alpha(\epsilon)}{2}\big(\delta_{(-1,1)} + \delta_{(1,-1)}\big), \\ \tilde{\pi}^* &= \frac{1-\alpha(\epsilon)}{2}\big(\delta_{(-1+\varepsilon,-1+\varepsilon)} + \delta_{(1-\varepsilon,1-\varepsilon)}\big) + \frac{\alpha(\epsilon)}{2}\big(\delta_{(1-\varepsilon,-1+\varepsilon)} + \delta_{(-1+\varepsilon,1-\varepsilon)}\big). \end{aligned}$$

Next, we find

$$W_1(\pi^*, \tilde{\pi}^*) = 2(1-\alpha(\varepsilon))2\varepsilon + (2\alpha(\varepsilon) - 1)2$$

by observing that an optimal coupling $\kappa \in \Pi(\pi^*, \tilde{\pi}^*)$ is to move a total mass of $2(1-\alpha(\varepsilon))$ over a $d_{X,1}$-distance of 2ε and mass $2\alpha(\varepsilon) - 1$ over distance $(2-\varepsilon) + \varepsilon = 2$. In view of $\alpha(\varepsilon) = \frac{1}{2} + \frac{\varepsilon}{4} + \mathcal{O}(\varepsilon^3)$ as $\varepsilon \to 0$, we deduce

$$W_1(\pi^*, \tilde{\pi}^*) = 3\varepsilon + \mathcal{O}(\varepsilon^2).$$

On the other hand, clearly

$$W_\infty(\mu_1, \mu_2; \tilde{\mu}_1, \tilde{\mu}_2) = \varepsilon.$$

In summary, any constant ℓ such that $W_1(\pi^*, \tilde{\pi}^*) \le \ell W_\infty(\mu_1, \mu_2; \tilde{\mu}_1, \tilde{\mu}_2)$ holds in the above example for all ε, has to satisfy $\ell \ge 3$.

It remains to see that we attain $\ell = 3$ in the last assertion of Theorem 3.13. For $q = 1$, Lemma 3.10 (i) with $\mathrm{diam}_1(X_2) = \mathrm{diam}([-1,1]) = 2$ shows that (I_q) is satisfied with $C_1 = \sqrt{2}$. Hence, the formula in Theorem 3.13 reads

$$\ell = N + (C_1/\sqrt{2})\,\mathrm{Lip}_\infty(c) = 2 + 1 = 3$$

as desired.

We remark that this example can be extended to more general parameters. Replacing c by Lc for some $L > 0$ leads to a different Lipschitz constant in the definition of l. Replacing $\alpha(\varepsilon)$ by $\alpha(L\varepsilon)$ in the formula for $W_1(\pi^*, \tilde{\pi}^*)$, one finds that the constant l is again sharp. Similarly, replacing $[-1,1]$ by $[-K, K]$ for some $K > 0$ and replacing 1 by K in the definition of the marginals, we find that only the constant C_1 changes in the definition of l, while for $W_1(\pi^*, \tilde{\pi}^*)$ one replaces the final 2 by $2K$. Again, the constant l remains sharp.

4.5 Application to Sinkhorn's Algorithm

Proof of Theorem 3.15. We first observe that π^n is the optimizer of the problem $S_{\mathrm{ent}}(\pi_1^n, \pi_2^n, c)$:

$$\begin{aligned}\pi^n &= \operatorname*{arg\,min}_{\pi\in\Pi(\pi_1^n,\pi_2^n)} D_{\mathrm{KL}}(\pi, \pi^0)\\ &= \operatorname*{arg\,min}_{\pi\in\Pi(\pi_1^n,\pi_2^n)} \int c\, d\pi + D_{\mathrm{KL}}(\pi, \mu_1\otimes\mu_2)\\ &= \operatorname*{arg\,min}_{\pi\in\Pi(\pi_1^n,\pi_2^n)} \int c\, d\pi + D_{\mathrm{KL}}(\pi, \pi_1^n\otimes\pi_2^n),\end{aligned}$$

where the last step used Remark 2.1. (The first identity is well known; e.g., it follows from the fact that by construction, $d\pi^n/d\pi^0$ admits a factorization $a(x_1)b(x_2)$.) To apply our stability results, we require the convergence of the marginals in W_p. Indeed, $D_{\mathrm{KL}}(\pi_i^n, \mu_i) \to 0$ holds by a standard entropy calculation, see for instance [51]. More precisely, we have

$$D_{\mathrm{KL}}(\pi_i^n, \mu_i) \le 2\frac{D_{\mathrm{KL}}(\pi^*, P_c)}{n} \tag{4.19}$$

according to [36, Corollary 1]. By the exponential moment condition on μ_i and [9, Corollary 2.3], (4.19) yields

$$W_p(\pi_i^n, \mu_i) \le C_0 C_{\mu_i}(n^{-\frac{1}{p}} + n^{-\frac{1}{2p}}) \qquad \text{where}$$

$$\begin{aligned}C_0 &:= \max\left\{(2D_{\mathrm{KL}}(\pi^*, P_c))^{\frac{1}{p}}, (2D_{\mathrm{KL}}(\pi^*, P_c))^{\frac{1}{2p}}\right\},\\ C_{\mu_i} &:= 2\inf_{x_0\in X_i, \alpha>0}\left(\frac{1}{\alpha}\Big(\frac{3}{2} + \log\int \exp(\alpha d_{X_i}(x_0, x_i))\,\mu_i(dx_i)\Big)\right)^{\frac{1}{p}}.\end{aligned}$$

As a result,

$$\Delta := \max_{i=1,2} W_p(\pi_i^n, \mu_i) \le C_0 \max\{C_{\mu_1}, C_{\mu_2}\}(k^{-\frac{1}{p}} + k^{-\frac{1}{2p}}). \tag{4.20}$$

We remark that $\Delta = W_p(\pi_1^n, \pi_2^n; \mu_1, \mu_2)$ as $W_p(\pi_1^n, \mu_1) = 0$ or $W_p(\pi_2^n, \mu_2) = 0$ for each n, consistent with our previous notation. By (4.20), π_i^n has a finite p-th moment. Noting also that

$$D_{\mathrm{KL}}(\pi^n, \mu_1\otimes\mu_2) = D_{\mathrm{KL}}(\pi^n, \pi_1^n\otimes\pi_2^n) + \sum_{i=1}^2 D_{\mathrm{KL}}(\pi_i^n, \mu_i), \tag{4.21}$$

assertion (i) thus follows directly from Theorem 3.7 (i).

Regarding (ii), note that the p-th moments of π_i^n are bounded uniformly in n due to (4.20). In view of Lemma 3.5, the cost function c thus satisfies (A_L) with a uniform constant L for the marginals $(\pi_1^n, \pi_2^n)_n$ as well as (μ_1, μ_2). Using also (4.19) and (4.21), Theorem 3.7 (ii) yields

$$|\mathcal{F}(\pi^*) - \mathcal{F}(\pi^n)| \leq L\,\Delta + 2D_{\mathrm{KL}}(\pi^*, P_c)n^{-1}.$$

In view of (4.20), the claimed rate for $|\mathcal{F}(\pi^*) - \mathcal{F}(\pi^n)|$ follows. Finally, (I_q') holds with constant C_q' by Lemma 3.10 (iii) and thus Theorem 3.11 yields

$$W_q(\pi^*, \pi^n) \leq 2^{(\frac{1}{q}-\frac{1}{p})}\,\Delta + C_q'(2L)^{1/q}\,\Delta^{\frac{1}{q}} + C_q' L^{\frac{1}{2q}}\,\Delta^{\frac{1}{2q}},$$

so that the claimed rate for $W_q(\pi^*, \pi^n)$ follows via (4.20). □

References

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

[2] 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.

[3] D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, 2017.

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

[6] E. Bernton, P. Ghosal, and M. Nutz. Entropic optimal transport: Geometry and large deviations. *Duke Math. J., to appear*, 2021. arXiv:2102.04397.

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

[8] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, 2018.

[9] F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math. (6)*, 14(3):331–352, 2005.

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

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

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

[13] Y. Chen, T. Georgiou, and M. Pavon. Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM J. Appl. Math.*, 76(6):2375–2396, 2016.

[14] 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.

[15] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45(1):223–256, 2017.

[16] 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.

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

[18] A. Corenflos, J. Thornton, G. Deligiannidis, and A. Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, pages 2100–2111. PMLR, 2021.

[19] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975.

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

[21] M. Cuturi, O. Teboul, and J.-P. Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[22] G. Deligiannidis, V. De Bortoli, and A. Doucet. Quantitative uniform stability of the iterative proportional fitting procedure. *Preprint arXiv:2108.08129v1*, 2021.

[23] S. Di Marino and A. Gerolin. An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *J. Sci. Comput.*, 85(2):Paper No. 27, 28, 2020.

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

[25] M. Essid and J. Solomon. Quadratically regularized optimal transport on graphs. *SIAM J. Sci. Comput.*, 40(4):A1961–A1986, 2018.

[26] G. B. Folland. *Real analysis*. Pure and Applied Mathematics. John Wiley & Sons, New York, second edition, 1999.

[27] H. Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 101–203. Springer, Berlin, 1988.

[28] H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. W. de Gruyter, Berlin, 3rd edition, 2011.

[29] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735, 1989.

[30] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity

of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[31] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448. 2016.

[32] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, PMLR, pages 1608–1617, 2018.

[33] P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and Schrödinger bridges. *Preprint arXiv:2106.03670v1*, 2021.

[34] 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.

[35] D. Lacker. A non-exponential extension of Sanov's theorem via convex duality. *Adv. in Appl. Probab.*, 52(1):61–101, 2020.

[36] F. Léger. A gradient descent perspective on Sinkhorn. *Appl. Math. Optim.*, 84(2):1843–1855, 2021.

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

[38] 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.

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

[40] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

[41] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32*, pages 4541–4551. 2019.

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

[43] 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.

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

[45] M. Nutz and J. Wiesel. Entropic optimal transport: Convergence of potentials. *Probab. Theory Related Fields, to appear*. arXiv:2104.11720v2.

[46] M. Nutz and J. Wiesel. Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *Preprint arXiv:2201.10059v1*, 2022.

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

[48] 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.

[49] A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein two-sample

testing and related families of nonparametric tests. *Entropy*, 19(2):Paper No. 47, 15, 2017.

[50] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40:99–121, 2000.

[51] L. Rüschendorf. Convergence of the iterative proportional fitting procedure. *Ann. Statist.*, 23(4):1160–1174, 1995.

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

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