On the Convergence Rate of Sinkhorn's Algorithm[*]

Promit Ghosal[†] Marcel Nutz[‡]

April 5, 2025

Abstract

We study Sinkhorn's algorithm for solving the entropically regularized optimal transport problem. Its iterate π_t is shown to satisfy $H(\pi_t|\pi_*)+H(\pi_*|\pi_t) = O(t^{-1})$ where H denotes relative entropy and π_* the optimal coupling. This holds for a large class of cost functions and marginals, including quadratic cost with subgaussian marginals. We also obtain the rate $O(t^{-1})$ for the dual suboptimality and $O(t^{-2})$ for the marginal entropies. More precisely, we derive non-asymptotic bounds, and in contrast to previous results on linear convergence that are limited to bounded costs, our estimates do not deteriorate exponentially with the regularization parameter. We also obtain a stability result for π_* as a function of the marginals, quantified in relative entropy.

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

1 Introduction

Let (X,μ) and (Y,ν) be Polish probability spaces and $c:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}$ a measurable and nonnegative (or suitably integrable) cost function. The entropic optimal transport problem with regularization parameter $\varepsilon\in(0,\infty)$ is

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathsf{X}\times\mathsf{Y}} c(x,y)\,\pi(dx,dy)+\varepsilon H(\pi|\mu\otimes\nu), \tag{1.1}$$

where $H(\,\cdot\,|\mu\otimes\nu)$ denotes relative entropy (Kullback–Leibler divergence) with respect to the product measure $\mu\otimes\nu$ of the marginals,

$$H(\pi|\mu\otimes\nu):=\begin{cases}\int\log\frac{d\pi}{d(\mu\otimes\nu)}\,d\pi, & \pi\ll\mu\otimes\nu,\\ \infty, & \pi\not\ll\mu\otimes\nu,\end{cases}$$

[*] The authors thank Stephan Eckstein and Flavien Léger for helpful discussions.

[†] Department of Statistics, University of Chicago, promit@uchicago.edu. Research supported by NSF Grant DMS-2153661.

[‡] Depts. of Statistics and Mathematics, Columbia University, mnutz@columbia.edu. Research supported by NSF Grants DMS-1812661, DMS-2106056, DMS-2407074.

and $\Pi(\mu, \nu)$ the set of all couplings; i.e., probability measures π on $\mathsf{X}\times\mathsf{Y}$ with marginals (μ, ν). Entropic optimal transport traces back to Schrödinger's thought experiment [52] on the most likely evolution of a particle system and its mathematical formalization as the Schrödinger bridge (see [29, 39] for surveys). Nowadays, following [22], entropic optimal transport is widely used as an approximation of the (unregularized) Monge–Kantorovich optimal transport problem corresponding to $\varepsilon = 0$, especially for computing the Wasserstein distance in high-dimensional applications such as machine learning, statistics, image and language processing (e.g., [2, 4, 16, 50]). As a result, the cost function of principal interest is $c(x, y) = |x - y|^p$ on $\mathbb{R}^d \times \mathbb{R}^d$, especially for $p = 2$, and the convergence properties as $\varepsilon \to 0$ are an active area of research (early contributions are [15, 19, 38, 42, 43], current ones are [1, 6, 8, 13, 21, 28, 34, 44, 46, 48, 57]). The main appeal of (1.1) in this applied context is that it can be solved efficiently by Sinkhorn's algorithm, also called iterative proportional fitting procedure or IPFP; see [49] and its numerous references.

Sinkhorn's algorithm can be stated equivalently in primal or dual terms. In the primal formulation, it is initialized at the probability measure $\pi_{-1} \propto e^{-c/\varepsilon} d(\mu \otimes \nu)$ and the iterates are defined for $t \geq 0$ by

$$\pi_{2t} := \operatorname*{arg\,min}_{\Pi(*,\nu)} H(\,\cdot\,|\pi_{2t-1}), \qquad \pi_{2t+1} := \operatorname*{arg\,min}_{\Pi(\mu,*)} H(\,\cdot\,|\pi_{2t}), \tag{1.2}$$

where $\Pi(*, \nu)$ is the set of measures on $\mathsf{X} \times \mathsf{Y}$ with second marginal ν (and arbitrary first marginal), and $\Pi(\mu, *)$ is defined analogously. The dual formulation produces two sequences of functions $\varphi_t : \mathsf{X} \to \mathbb{R}$ and $\psi_t : \mathsf{Y} \to \mathbb{R}$. Namely, we initialize at $\varphi_0 := 0$ and define for $t \geq 0$ the iterates

$$\begin{aligned} \psi_t(y) &:= -\log \int_{\mathsf{X}} e^{\varphi_t(x) - c(x,y)/\varepsilon}\, \mu(dx), \\ \varphi_{t+1}(x) &:= -\log \int_{\mathsf{Y}} e^{\psi_t(y) - c(x,y)/\varepsilon}\, \nu(dy). \end{aligned} \tag{1.3}$$

The primal and dual formulations are related by $d\pi_{2t} = e^{\varphi_t \oplus \psi_t - c/\varepsilon}\, d(\mu \otimes \nu)$ where $(\varphi \oplus \psi)(x, y) := \varphi(x) + \psi(y)$, and similarly for π_{2t-1} with ψ_{t-1}. To wit, the minimization in (1.2) is solved in closed form by (1.3). The latter boils down to a simple matrix-vector multiplication in a discretized setting, whence the suitability for high-dimensional problems.

Our main result is the convergence of the algorithm in the sense of relative entropy and its rate, under general conditions including unbounded costs. Methodologically, we proceed in two steps. First, we show that certain (exponential) moment estimates for the iterates φ_t, ψ_t imply a convergence rate. Second, we show how to obtain the necessary estimates for important classes of cost functions c and marginals (μ, ν). The following illustrates the main findings by giving a simplified statement for powers of the distance cost, covering in particular quadratic cost with subgaussian marginals (μ, ν). We denote by $\pi_* \in \Pi(\mu, \nu)$ the unique solution of (1.1) and by (φ_*, ψ_*) the associated dual potentials; see Section 2 for the pertinent definitions.

Example 1.1. Let $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$ and $c(x, y) = |x - y|^p$ where $p \in [0, \infty)$. Suppose that

$$\int e^{\lambda |x|^p}\, \mu(dx) + \int e^{\lambda |y|^p}\, \nu(dy) < \infty \quad \text{for some } \lambda > 0.$$

Let π_t and (φ_t, ψ_t) be the Sinkhorn iterates (1.2)–(1.3), and let (μ_t, ν_t) denote the marginals of π_t. We have

$$H(\pi_*|\pi_t) + H(\pi_t|\pi_*) = O(t^{-1}), \tag{1.4}$$

$$\int (\varphi_* - \varphi_t)\, d\mu + \int (\psi_* - \psi_t)\, d\nu = O(t^{-1}), \tag{1.5}$$

$$H(\mu_t|\mu) + H(\mu|\mu_t) + H(\nu_t|\nu) + H(\nu|\nu_t) = O(t^{-2}). \tag{1.6}$$

The constants implicit in $O(\cdot)$ above are detailed in Sections 4 and 5; their dependence on the regularization parameter ε is a scaling by ε^{-2p}.

Note that convergence rates for total variation follow immediately from the above rates via Pinsker's inequality, with constants scaling like ε^{-p}. More generally, we also cover costs $c(x, y) = d(x, y)^p$ for an arbitrary (measurable) distance d on a Polish space and differentiable costs c with $|Dc(x, y)| \leq C(1 + |x| + |y|)^{p-1}$ on a Banach space, or any measurable costs with similar growth properties. Indeed, our rates depend only on an a priori estimate of the form

$$C_1 := \sup_{t \geq 0} \inf_{\alpha > 0} \frac{2}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha |\varphi_t - \varphi_*|}\, d\mu \right) < \infty. \tag{1.7}$$

Our key result (Theorem 4.4) is a *non-asymptotic* bound of the form

$$H(\pi_*|\pi_{2t}) \leq \frac{C}{t - t_1}, \quad t > t_1$$

where $t_1 \in \mathbb{N}$ is bounded explicitly and C depends at most quadratically on ε^{-p}. By a bootstrap-type argument, this turns out to imply similar non-asymptotic bounds for the other quantities in (1.4)–(1.6); cf. Corollaries 4.6 and 4.7.

To provide (1.7), we establish that $|\varphi_t(x)| \leq C(1+|x|^p)$ for a general class of costs whenever $\int e^{\lambda|y|^p}\nu(dy) < \infty$ for some $\lambda > 0$. This implies (1.7) under the symmetric condition $\int e^{\lambda|x|^p}\mu(dx) < \infty$. Moreover, we obtain a linear dependence of (1.7) on ε^{-p}. Mirroring a priori estimates for c-convex functions in optimal transport theory [32, 56], our approach applies to arbitrary "biconjugate" functions, including Sinkhorn iterates and dual potentials (cf. Section 5). In particular, it provides a reasonably general solution to the vexing problem of bounding potentials from below.

A key innovation for our convergence analysis is to use the "weighted Csiszár–Kullback–Pinsker" inequality of Bolley–Villani [9]. On the strength of bounds like (1.7), the Bolley–Villani inequality allows us to estimate the relative entropy between certain couplings by the (symmetric) relative entropy of their marginals, with constants depending on the moments of the potentials of the couplings. This also gives rise to a new stability result (Theorem 3.3) for the optimal coupling π_* of (1.1) wrt. the marginals (μ,ν) which is of independent interest. For the convergence results, the second key innovation uses the finer details of Sinkhorn's algorithm and its monotonicity properties: we relate the improvement $H(\pi_*|\pi_{2t}) - H(\pi_*|\pi_{2t+2})$ to a marginal entropy and deduce a difference equation enabling us to analyze $H(\pi_*|\pi_{2t})$ through elementary arguments.

Sinkhorn's algorithm dates back as far as [24]; see also [30, 40] for Fortet's (related but different) iteration. Early contributions to the analysis include [35, 36, 53, 54]. Introduced by [31] in the discrete (matrix scaling) case and generalized to the continuous setting by [14], a contraction argument can be used to show linear (i.e., geometric) convergence in the Hilbert–Birkhoff metric when the cost c is uniformly bounded. A different avenue, viable also in the multi-marginal problem, is taken in [10] where the algorithm is seen as a dual block-coordinate ascent. Here, boundedness of c is responsible for the strong concavity which yields linear convergence of the iterates φ_t, ψ_t in L^2 and of their suboptimality gap. A common feature of all previous (but see the end of this section) arguments for linear convergence is that (a) bounded cost is crucial and (b) the constants deteriorate *exponentially* in the regularization parameter ε. For instance, the suboptimality gap decays like β^t in [10], where $\beta = 1 - e^{-24\|c\|_\infty/\varepsilon}$. As ε is taken to be small in practice, β is extremely close to 1, arguably failing to explain the fast convergence observed in computational practice.

Using weak* compactness, [51] obtained a qualitative convergence result (in relative entropy) relaxing the boundedness condition on the cost. However, several other conditions are introduced, including one (see (B1) in [51]) that essentially forces c to be bounded from above in one variable and thus excludes quadratic cost with unbounded marginal supports. A different form of compactness, in the space of dual potentials, was used in [47] to show convergence $\pi_t \to \pi_*$ in total variation under the condition $\int e^{\lambda c}\, d(\mu \otimes \nu) < \infty$ for some $\lambda > 0$. Yet this approach does not yield convergence in relative entropy, and more importantly, does not yield a rate. Another qualitative result, on weak convergence $\pi_t \to \pi_*$, was stated in [45] for a very general continuous cost c, where it was observed that Sinkhorn convergence can be deduced from the weak stability of (1.1) wrt. the marginals (μ, ν). The latter was established in [33] using the geometric approach of [7] which avoids integrability conditions on c. Again, the result is purely qualitative.

The same link with stability was exploited in [27] whose main result is the quantitative stability of the optimal coupling π_* wrt. the marginals (μ, ν) in the p-Wasserstein distance $\mathcal{W}_p$, under certain continuity and integrability conditions on the cost and marginals. By a general result of [37], the marginals (μ_t, ν_t) of the Sinkhorn iterate π_t converge to (μ, ν) in relative entropy at rate $O(t^{-1})$. Combining the two results, [27] shows in particular that for subgaussian marginals (μ, ν) and costs c which are the product of two Lipschitz functions, $\mathcal{W}_2(\pi_t, \pi_*) = O(t^{-1/16})$. While the approach of [27] does not yield strong convergence (total variation or even relative entropy), it is the closest convergence result in that it covers (some) unbounded costs and is quantitative. We note that the rate in (1.4) is significantly faster and that the marginal rate of [37] is improved to $O(t^{-2})$ in (1.6) in our context.

Stability of entropic optimal transport wrt. the marginals has also been studied independently of Sinkhorn's algorithm, in addition to the aforementioned works. The first stability result is due to [12], for a setting with bounded cost and marginals equivalent to a common reference measure with densities uniformly bounded above and below. The authors showed by a differential approach that the potentials are continuous in L^p relative to the marginal densities. Very recently, a differential approach was also used in [11] to show uniform continuity of the potentials and their derivatives in Wasserstein distance $\mathcal{W}_2$, in a compact (possibly multi-marginal) setting. For two marginals, uniform continuity of the potentials in $\mathcal{W}_1$ was previously obtained by [23] using the Hilbert–Birkhoff projective metric. Also very recently, [5] showed a stability result in Wasserstein distance for unbounded marginals which extends to divergences other than relative entropy. Finally, in the setting of dynamic Schrödinger bridges satisfying a logarithmic Sobolev inequality for the underlying dynamics and marginal distributions with finite Fisher information, [17] used a negative order weighted homogeneous Sobolev norm to give a stability result for the relative entropy of the Schrödinger bridges wrt. their marginals. While in the dynamic setting, this result is closest to the spirit of our Theorem 3.3 as the estimate is in relative entropy and costs may be unbounded.

Since this paper was completed, several results on Sinkhorn's algorithm have been obtained, tackling some of the aforementioned issues. For quadratic cost and unbounded continuous marginals satisfying a log-concavity condition, [20] proves linear convergence based on a fine analysis of the gradients of Schrödinger potentials and Sinkhorn iterates. Meanwhile, [26] adapts Hilbert's projective metric to unbounded functions and obtains linear convergence for a general class of unbounded costs functions and sufficiently integrable marginals. For bounded and sufficiently regular cost and marginals, [18] shows that linear convergence holds with a constant depending polynomially (rather than exponentially) on the regularization parameter.

The remainder of this paper is organized as follows. Section 2 introduces the necessary background and notation. Section 3 provides basic estimates for the relative entropy between certain couplings in terms of their marginals, including the stability result for the optimal couplings. Section 4 establishes the main results on Sinkhorn's algorithm, taking for granted an a priori estimate for the dual iterates. Finally, Section 5 shows how to obtain those estimates for arbitrary biconjugate functions, for a large class of costs and marginals.

2 Background and Notation

In the entropic optimal transport problem (1.1), we may divide by ε to reduce to the case $\varepsilon = 1$, at the expense of changing the cost function to c/ε. In what follows, we will thus assume that $\varepsilon = 1$, yet keep an eye on the scaling behavior of the results. Our problem then reads

$$\mathcal{C}_1(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi + H(\pi|\mu\otimes\nu). \tag{2.1}$$

Throughout, (X,μ) and (Y,ν) are Polish probability spaces and $c : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}$ is a measurable function such that $e^{-c} \in L^1(\mu\otimes\nu)$ and (2.1) is finite. We write $\sim$ for measure-theoretic equivalence (having the same nullsets) and $a^{\pm} = \max\{\pm a, 0\}$ for $a \in \mathbb{R}$.

The following facts can be found, e.g., in [45, Section 4]. The problem (2.1) admits a unique minimizer $\pi_* \in \Pi(\mu,\nu)$; moreover, $\pi_* \sim \mu\otimes\nu$ with density

$$d\pi_* = e^{\varphi_*\oplus\psi_*-c}\, d(\mu\otimes\nu) \tag{2.2}$$

for some measurable functions $\varphi_* : \mathsf{X} \to \mathbb{R}$ and $\psi_* : \mathsf{Y} \to \mathbb{R}$ called *potentials.* The potentials are a.s. unique up an additive constant; i.e., $(\varphi_* - \alpha, \psi_* + \alpha)$ is another pair of potentials for any $\alpha \in \mathbb{R}$. The fact that $\pi_* \in \Pi(\mu, \nu)$ is equivalent to the *Schrödinger system*

$$\begin{aligned} \psi_*(y) &= -\log \int e^{\varphi_*(x) - c(x,y)}\,\mu(dx), \\ \varphi_*(x) &= -\log \int e^{\psi_*(y) - c(x,y)}\,\nu(dy). \end{aligned} \tag{2.3}$$

Occasionally it is convenient to absorb the cost into the "reference" measure $R \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$,

$$dR = \xi^{-1} e^{-c} d(\mu \otimes \nu), \qquad \xi := \int e^{-c}\, d(\mu \otimes \nu) \tag{2.4}$$

which allows us to state (2.1) as the entropy minimization

$$\mathcal{C}_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} H(\pi|R) - \log \xi. \tag{2.5}$$

Here $\mathcal{P}(\mathsf{Z})$ denotes the set of Borel probability measures on a topological space Z.

Recall that the primal iterates $\pi_t \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ of Sinkhorn's algorithm were stated in (1.2) and the dual iterates (φ_t, ψ_t) in (1.3), where we now have $\varepsilon = 1$. We remark in passing that (1.3) can be seen as the Gauss–Seidel algorithm for the system (2.3); this will be relevant in Section 5. Recall also that $\varphi_0 := 0$. With the conventions that $\pi_{-1} := R$ and $\psi_{-1} := -\log \xi$ as defined in (2.4), the primal and dual iterates are related for all $t \geq 0$ by

$$d\pi_{2t} = e^{\varphi_t \oplus \psi_t - c}\, d(\mu \otimes \nu), \qquad d\pi_{2t-1} = e^{\varphi_t \oplus \psi_{t-1} - c}\, d(\mu \otimes \nu); \tag{2.6}$$

see [45, Algorithm 6.2]. By construction, π_{2t} has marginals (μ_{2t}, ν) while π_{2t-1} has marginals (μ, ν_{2t-1}), where

$$\frac{d\mu_{2t}}{d\mu} = e^{\varphi_t - \varphi_{t+1}}, \qquad \frac{d\nu_{2t-1}}{d\nu} = e^{\psi_{t-1} - \psi_t}, \qquad t \geq 0. \tag{2.7}$$

We recall from [45, Lemma 6.4 (i)] that $\varphi_t \in L^1(\mu)$ and $\psi_t \in L^1(\nu)$. Moreover, writing $\mu(\varphi) := \int \varphi\, d\mu$ for brevity,

$$0 \leq \mu(\varphi_t) \leq \mu(\varphi_{t+1}) \leq \mathcal{C}_1(\mu, \nu) + \log \xi, \qquad t \geq 0, \tag{2.8}$$

$$-\log \xi \leq \nu(\psi_t) \leq \nu(\psi_{t+1}) \leq \mathcal{C}_1(\mu, \nu), \qquad t \geq -1. \tag{2.9}$$

Here the first two inequalities in each line hold by [45, Lemma 6.4 (iii)] while the last inequality follows from $\mu(\varphi_t) + \nu(\psi_t) \leq \mathcal{C}_1(\mu, \nu)$ via the first inequality in the other line. The inequality $\mu(\varphi_t) + \nu(\psi_t) \leq \mathcal{C}_1(\mu, \nu)$, in turn, holds by [45, Lemma 6.4 (ii) and Proposition 6.5] or, more directly, by duality. In most applications, c is nonnegative and thus $\log \xi \leq 0$, so that the inequalities also hold with that term omitted.

3 Auxiliary Entropy Estimates

In this section, we derive several auxiliary results that bound the relative entropy of couplings in terms of their marginals and potentials. For our analysis of Sinkhorn's algorithm in Section 4, Lemma 3.2 below will be used to estimate the relative entropy between the Sinkhorn iterate π_t and the optimal coupling π_* (cf. (4.2)), whereas Lemma 3.6 will be used to bound the dual suboptimality of π_t (cf. Corollary 4.7). As a by-product, our considerations yield a stability result for the optimal coupling with respect to the marginals, reported in Theorem 3.3.

Lemma 3.1. *Consider $\mu, \mu' \in \mathcal{P}(\mathsf{X})$ with $H(\mu'|\mu) < \infty$ and let $F : \mathsf{X} \to \mathbb{R}$ be measurable. Then*

$$\left| \int F \, d(\mu' - \mu) \right| \leq C(F) \left(\sqrt{H(\mu'|\mu)} + \frac{1}{2} H(\mu'|\mu) \right)$$

where

$$C(F) := \inf_{\alpha > 0} \frac{2}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha |F|} \, d\mu \right).$$

Proof. The weighted CKP inequality of Bolley–Villani [9, Theorem 2.1] states that for any measurable $\phi : \mathsf{X} \to [0, \infty]$,

$$\|\phi \, d(\mu' - \mu)\|_{TV} \leq C_\phi \left(\sqrt{H(\mu'|\mu)} + \frac{1}{2} H(\mu'|\mu) \right)$$

where $C_\phi = \frac{3}{2} + \log \int e^{2\phi} \, d\mu$. The claim follows by choosing $\phi = \frac{\alpha}{2}|F|$ and optimizing over $\alpha > 0$. □

Lemma 3.2. *Consider $\mu, \mu' \in \mathcal{P}(\mathsf{X})$ with $H(\mu'|\mu) < \infty$ and $\nu, \nu' \in \mathcal{P}(\mathsf{Y})$ with $H(\nu'|\nu) < \infty$, as well as measurable functions $f, f' : \mathsf{X} \to \mathbb{R}$ and $g, g' : \mathsf{Y} \to \mathbb{R}$ with*

$$C_1 := \inf_{\alpha > 0} \frac{2}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha |f' - f|} \, d\mu \right) < \infty,$$

$$C_2 := \inf_{\alpha > 0} \frac{2}{\alpha} \left(\frac{3}{2} + \log \int e^{\alpha |g' - g|} \, d\nu \right) < \infty.$$

Let $\pi \in \Pi(\mu, \nu)$ be of the form $d\pi = e^{f \oplus g - c} d(\mu \otimes \nu)$.

(a) If $\pi' \in \Pi(\mu', \nu')$ and $d\pi' = e^{f' \oplus g' - c} d(\mu' \otimes \nu')$, then

$$\begin{aligned} H(\pi'|\pi) + H(\pi|\pi') \leq C_1 \sqrt{H(\mu'|\mu)} + (1 + C_1/2) H(\mu'|\mu) + H(\mu|\mu') \\ + C_2 \sqrt{H(\nu'|\nu)} + (1 + C_2/2) H(\nu'|\nu) + H(\nu|\nu'). \end{aligned}$$

(b) If $\pi' \in \Pi(\mu', \nu')$ and $d\pi' = e^{f' \oplus g' - c} d(\mu \otimes \nu)$, then

$$\begin{aligned} H(\pi'|\pi) + H(\pi|\pi') \leq C_1 \left(\sqrt{H(\mu'|\mu)} + \frac{1}{2} H(\mu'|\mu) \right) \\ + C_2 \left(\sqrt{H(\nu'|\nu)} + \frac{1}{2} H(\nu'|\nu) \right). \end{aligned}$$

If $\nu = \nu'$, the requirement that $C_2 < \infty$ can be dropped.

Proof. (a) In view of $C_1 < \infty$ and $H(\mu'|\mu) < \infty$, we have $|f' - f| \in L^1(\mu')$ by the variational representation of $H(\mu'|\mu)$; cf. [45, Equation (1.4)]. Similarly, $|g' - g| \in L^1(\nu')$. Using the definition of relative entropy and $\pi' \in \Pi(\mu', \nu')$,

$$\begin{aligned} & H(\pi'|\pi) \\ & = \int \log \left(\frac{d\pi'}{d(\mu' \otimes \nu')} \frac{d(\mu \otimes \nu)}{d\pi} \frac{d(\mu' \otimes \nu')}{d(\mu \otimes \nu)} \right) d\pi' \\ & = \int (f' - f)\, d\pi' + \int (g' - g)\, d\pi' + \int \log \left(\frac{d\mu'}{d\mu} \right) d\pi' + \int \log \left(\frac{d\nu'}{d\nu} \right) d\pi' \\ & = \int (f' - f)\, d\mu' + \int (g' - g)\, d\nu' + H(\mu'|\mu) + H(\nu'|\nu). \end{aligned}$$

A symmetric expression holds for $H(\pi|\pi')$; note that $|f' - f| \in L^1(\mu)$ and $|g' - g| \in L^1(\nu)$ due to $C_1 + C_2 < \infty$ and thus $H(\mu|\mu') + H(\nu|\nu') = \infty$ if and only if $H(\pi|\pi') = \infty$. Adding the two expressions and applying Lemma 3.1,

$$\begin{aligned} H(\pi'|\pi) + H(\pi|\pi') &= H(\mu'|\mu) + H(\mu|\mu') + H(\nu'|\nu) + H(\nu|\nu') \\ & \quad + \int (f' - f)\, d(\mu' - \mu) + \int (g' - g)\, d(\nu' - \nu) \\ & \leq C_1 \sqrt{H(\mu'|\mu)} + (1 + C_1/2) H(\mu'|\mu) + H(\mu|\mu') \\ & \quad + C_2 \sqrt{H(\nu'|\nu)} + (1 + C_2/2) H(\nu'|\nu) + H(\nu|\nu'). \end{aligned}$$

(b) Similarly as in (a),

$$\begin{aligned}H(\pi'|\pi) &= \int \log\left(\frac{d\pi'}{d(\mu\otimes\nu)}\frac{d(\mu\otimes\nu)}{d\pi}\right) d\pi' \\ &= \int (f'-f)\,d\pi' + \int (g'-g)\,d\pi' \\ &= \int (f'-f)\,d\mu' + \int (g'-g)\,d\nu'\end{aligned}$$

leads to

$$H(\pi'|\pi) + H(\pi|\pi') = \int (f'-f)\,d(\mu'-\mu) + \int (g'-g)\,d(\nu'-\nu),$$

and now the claim again follows via Lemma 3.1.

Regarding the last assertion, suppose that $\nu=\nu'$. The above calculations go through (with $0*\infty:=0$) if we can ensure that $g'-g\in L^1(\nu)$. For a function G depending only on $y\in\mathsf{Y}$, clearly $G\in L^1(\nu)$ is equivalent to $G\in L^1(\pi)$ and to $G\in L^1(\pi')$. In general, note that if $\pi'\ll\pi$, then $\log(d\pi'/d\pi)^-\in L^1(\pi')$ as $x\mapsto x\log x$ is bounded from below. Consider first (b). Here $\pi'\sim\pi$ and $\log(d\pi'/d\pi)=(f'-f)\oplus(g'-g)$, and as $(f'-f)\in L^1(\mu+\mu')$ thanks to $C_1<\infty$, it follows that $(g'-g)^-\in L^1(\pi')$ and hence $(g'-g)^-\in L^1(\nu)$ as g',g depend only on $y\in\mathsf{Y}$. Interchanging the roles of π,π' yields $(g-g')^-\in L^1(\pi)$ and hence $(g'-g)^+\in L^1(\nu)$. In summary, $g'-g'\in L^1(\nu)$ as desired. The argument for (a) is similar after noting that we may assume $H(\mu|\mu')<\infty$ without loss of generality. □

In view of (2.2), Lemma 3.2 entails the following stability result for the optimal coupling which is of independent interest.

Theorem 3.3 (Stability)**.** *Consider $(\mu,\nu),(\mu',\nu')\in\mathcal{P}(\mathsf{X})\times\mathcal{P}(\mathsf{Y})$ such that the associated EOT problems* (2.1) *are finite. Let $\pi\in\Pi(\mu,\nu)$, $\pi'\in\Pi(\mu',\nu')$ be the respective solutions and (φ,ψ), (φ',ψ') associated potentials. Then*

$$\begin{aligned}H(\pi'|\pi)+H(\pi|\pi') \le{}& C_1\sqrt{H(\mu'|\mu)} + (1+C_1/2)H(\mu'|\mu) + H(\mu|\mu') \\ &+ C_2\sqrt{H(\nu'|\nu)} + (1+C_2/2)H(\nu'|\nu) + H(\nu|\nu')\end{aligned}$$

where

$$C_1 := \inf_{\alpha>0}\frac{2}{\alpha}\left(\frac{3}{2}+\log\int e^{\alpha|\varphi'-\varphi|}\,d\mu\right),\quad C_2 := \inf_{\alpha>0}\frac{2}{\alpha}\left(\frac{3}{2}+\log\int e^{\alpha|\psi'-\psi|}\,d\nu\right).$$

Remark 3.4. Lemma 3.2 and Theorem 3.3 generalize to the multi-marginal context where (μ,ν), (μ',ν') are replaced by $(\mu_1,\dots,\mu_N)$, $(\mu'_1,\dots,\mu'_N)$ with corresponding constants $C_1,\dots,C_N$.

Remark 3.5. This is an informal remark about the general approach. Neglecting higher-order terms, Theorem 3.3 is a Hölder-type estimate where the left-hand side is a relative entropy and the right-hand is the square-root of a relative entropy with some constant in front. To obtain a Lipschitz estimate along these lines, one would need to replace the constant by a term involving another square-root of relative entropy. That additional step is indeed possible in the bounded setting of [23]. There, it is shown that the potentials are Lipschitz continuous with respect to the marginals as a map from 1-Wasserstein space to the space of continuous functions with uniform metric, and by Pinsker's inequality, the 1-Wasserstein distance is further bounded by a square-root of relative entropy. In our more general setting, that step is not available, and we shall merely estimate the constant to be finite. At a high level, that is why we obtain Hölder stability and (in the next section) sublinear convergence, rather than Lipschitz stability and linear convergence.

The following lemma is based on a similar calculation as Lemma 3.2; it will be used to bound the dual suboptimality in Corollary 4.7. We recall the notation (2.4).

Lemma 3.6. *Consider $\mu, \mu' \in \mathcal{P}(\mathsf{X})$ with $H(\mu'|\mu) < \infty$ and $\nu, \nu' \in \mathcal{P}(\mathsf{Y})$ with $H(\nu'|\nu) < \infty$, as well as measurable functions $f, f' : \mathsf{X} \to \mathbb{R}$ and $g, g' : \mathsf{Y} \to \mathbb{R}$ with*

$$\begin{aligned} \tilde{C}_1 &:= \inf_{\alpha>0} \frac{2}{\alpha}\left(\frac{3}{2} + \log \int e^{\alpha|f'|}\, d\mu\right) < \infty, \\ \tilde{C}_2 &:= \inf_{\alpha>0} \frac{2}{\alpha}\left(\frac{3}{2} + \log \int e^{\alpha|g'|}\, d\nu\right) < \infty. \end{aligned}$$

Let $\pi \in \Pi(\mu, \nu)$ and $\pi' \in \Pi(\mu', \nu')$ be of the form

$$d\pi = e^{f \oplus g - c}\, d(\mu \otimes \nu), \quad d\pi' = e^{f' \oplus g' - c}\, d(\mu \otimes \nu).$$

Then we have $H(\pi'|R) < \infty$ and

$$\begin{aligned} H(\pi|R) - H(\pi'|R) \leq H(\pi|\pi') &+ \tilde{C}_1\left(\sqrt{H(\mu'|\mu)} + \frac{1}{2}H(\mu'|\mu)\right) \\ &+ \tilde{C}_2\left(\sqrt{H(\nu'|\nu)} + \frac{1}{2}H(\nu'|\nu)\right). \end{aligned}$$

If $\nu = \nu'$ and $(f, g) \in L^1(\mu) \times L^1(\nu)$, the requirement that $\tilde{C}_2 < \infty$ can be dropped.

Proof. In view of $H(\mu'|\mu) < \infty$ and $H(\nu'|\nu) < \infty$, the finiteness of $\tilde{C}_1, \tilde{C}_2$ implies $f' \in L^1(\mu + \mu')$ and $g' \in L^1(\nu + \nu')$, and in particular $H(\pi'|R) < \infty$. Using the definition of relative entropy, (2.4), the stated form of $d\pi$ and $d\pi'$, and $\pi' \in \Pi(\mu', \nu')$,

$$\begin{aligned} H(\pi'|R) + H(\pi|\pi') &= \log \xi + \int f' \oplus g' \, d\pi' + \int (f - f') \oplus (g - g') \, d\pi \\ &= \int f' \oplus g' \, d(\pi' - \pi) + \log \xi + \int f \oplus g \, d\pi \\ &= \int f' \, d(\mu' - \mu) + \int g' \, d(\nu' - \nu) + H(\pi|R). \end{aligned}$$

We see that $H(\pi|\pi') = \infty$ if and only if $H(\pi|R) = \infty$, and in that case the claim is trivial. In the finite case, the claim follows by rearranging the above and using Lemma 3.1 to estimate the integrals.

Turning to the last assertion, let $\nu = \nu'$ and $(f, g) \in L^1(\mu) \times L^1(\nu)$. As $f' \in L^1(\mu + \mu')$ due to $\tilde{C}_1 < \infty$, the fact that $\log(d\pi'/dR)^- \in L^1(\pi')$ implies $(g')^- \in L^1(\pi')$ and hence $(g')^- \in L^1(\nu)$. Whereas $\log(d\pi/d\pi')^- \in L^1(\pi)$ implies $(-g')^- \in L^1(\pi)$ and hence $(g')^+ \in L^1(\nu)$. Thus $g' \in L^1(\nu)$ and now the above calculation goes through as stated. □

4 Analysis of Sinkhorn's Algorithm

Recall that the entropic optimal transport problem (2.1) was assumed to be finite for the given marginals (μ, ν), that $\pi_* \in \Pi(\mu, \nu)$ denotes its unique optimizer and that (φ_*, ψ_*) are associated potentials (2.2). Recall also the primal Sinkhorn iterates π_t defined in (1.2), especially that $\pi_{2t} \in \Pi(\mu_{2t}, \nu)$ for $t \geq 0$, and the definitions of the dual iterates (φ_t, ψ_t) in (1.3).

The following lemma records two important monotonicity properties; the first one well known and the second due to [37]. See [45, Proposition 6.5] and [45, Proposition 6.10] for proofs in our setting.

Lemma 4.1. *(i) The sequence $\{H(\pi_*|\pi_t)\}_{t \geq -1}$ is decreasing.*

(ii) For all $t \geq 0$,

$$\begin{aligned} H(\mu_{2t}|\mu) \geq H(\nu|\nu_{2t+1}) \geq H(\mu_{2t+2}|\mu) \geq H(\nu|\nu_{2t+3}) \geq \dots, \\ H(\mu|\mu_{2t}) \geq H(\nu_{2t+1}|\nu) \geq H(\mu|\mu_{2t+2}) \geq H(\nu_{2t+3}|\nu) \geq \dots. \end{aligned}$$

In particular, $\{H(\mu_{2t}|\mu)\}_{t \geq 0}$ and $\{H(\mu|\mu_{2t})\}_{t \geq 0}$ are decreasing.

Our main results hinge on the following simple yet crucial observation.

Lemma 4.2. *For all $t \geq 0$,*

$$\begin{aligned}
H(\pi_*|\pi_{2t+2}) - H(\pi_*|\pi_{2t}) &= -[H(\mu|\mu_{2t}) + H(\nu|\nu_{2t+1})] \\
&\leq -[H(\mu_{2t+2}|\mu) + H(\mu|\mu_{2t+2})] \\
&\leq -H(\mu_{2t+2}|\mu).
\end{aligned}$$

Proof. Using the definitions of π_{2t+2}, π_{2t} followed by $\pi^* \in \Pi(\mu,\nu)$ and (2.7),

$$\begin{aligned}
&-H(\pi_*|\pi_{2t+2}) + H(\pi_*|\pi_{2t}) \\
&\quad = \int (\varphi_{t+1} \oplus \psi_{t+1}) - (\varphi_* \oplus \psi_*)\, d\pi_* + \int (\varphi_* \oplus \psi_*) - (\varphi_t \oplus \psi_t)\, d\pi_* \\
&\quad = \mu(\varphi_{t+1} - \varphi_t) + \nu(\psi_{t+1} - \psi_t) = H(\mu|\mu_{2t}) + H(\nu|\nu_{2t+1}),
\end{aligned}$$

which is the claimed identity. By Lemma 4.1 we have $H(\mu|\mu_{2t}) \geq H(\mu|\mu_{2t+2})$ and $H(\nu|\nu_{2t+1}) \geq H(\mu_{2t+2}|\mu)$, showing the first inequality. The second inequality is trivial. □

While we have not addressed the convergence of $H(\pi_*|\pi_{2t})$ yet, Lemma 4.2 entails that *the convergence rate of the marginals is at least one order faster.* This remarkable fact holds in full generality, without the moment conditions to be imposed below for our main results.

Proposition 4.3. *For all $t \geq 0$,*

$$H(\mu_{2t}|\mu) + H(\mu|\mu_{2t}) \leq \frac{2H(\pi_*|\pi_{2\lfloor t/2 \rfloor})}{t}.$$

Proof. Lemma 4.2 shows that $a_{2s} := H(\mu_{2s}|\mu) + H(\mu|\mu_{2s})$ satisfies

$$a_{2s} \leq H(\pi_*|\pi_{2s-2}) - H(\pi_*|\pi_{2s}), \quad s \geq 1.$$

As (a_{2s}) is decreasing by Lemma 4.1, it follows for any $m \geq 0$ that

$$m a_{4m} \leq \sum_{s=m+1}^{2m} a_{2s} \leq H(\pi_*|\pi_{2m}) - H(\pi_*|\pi_{4m}) \leq H(\pi_*|\pi_{2m}),$$

showing the claim for $t = 2m$. Similarly,

$$(m+1)a_{4m+2} \leq \sum_{s=m+1}^{2m+1} a_{2s} \leq H(\pi_*|\pi_{2m}) - H(\pi_*|\pi_{4m+2}) \leq H(\pi_*|\pi_{2m})$$

implies the claim for $t = 2m+1$. □

In view of $H(\pi_*|\pi_{2\lfloor t/2\rfloor}) \leq H(\pi_*|R) < \infty$, Proposition 4.3 already includes that the marginal entropies converge at least with rate $O(t^{-1})$. This will be improved to $O(t^{-2})$ in Corollary 4.6 below.

From now on, suppose that

$$C_1 := \sup_{t\geq 0} \inf_{\alpha>0} \frac{2}{\alpha}\left(\frac{3}{2} + \log \int e^{\alpha|\varphi_t-\varphi_*|}\, d\mu\right) < \infty; \tag{4.1}$$

sufficient conditions will be given in Section 5 (especially Corollary 5.8). As π_{2t} and π_* have the common second marginal ν, Lemma 3.2 then yields

$$H(\pi_{2t}|\pi_*) + H(\pi_*|\pi_{2t}) \leq C_1\left(\sqrt{H(\mu_{2t}|\mu)} + \frac{1}{2}H(\mu_{2t}|\mu)\right), \quad t \geq 0. \tag{4.2}$$

We can now state our main result. Its presentation includes a slight nuisance: to state a non-asymptotic bound, we need to account for both terms in (4.2), even though the square-root is clearly the asymptotically dominating one. This is the reason for the time t_1 below. While the geometric convergence (a) holds in the first few iterations before t_1, the key result is (b) establishing the $O(t^{-1})$ rate. See Remark 4.5 for more details on the constants and in particular their dependence on the regularization parameter ε.

Theorem 4.4. *Let* $t_0 = \inf\{t \geq 0 : H(\mu_{2t}|\mu) \leq 1\}$ *and* $t_1 = (t_0 - 1) \vee 0$ *and*

$$\kappa = (\tfrac{3}{2}C_1)^{-1} \wedge (2H(\pi_*|R))^{-1/2}.$$

Then t_1 *admits the bound* $t_1 \leq ((t_2 \wedge t_3) - 1)^+$, *where*

$$t_2 := \lceil H(\pi_*|R) - H(\pi_0|R)\rceil,$$
$$t_3 := \inf\{t \in \mathbb{N} : \lfloor t/2\rfloor \log(1+\kappa) + \log t \geq \log(2H(\pi_*|\pi_0))\}.$$

(a) For $0 \leq t < t_1$,

$$\begin{aligned} H(\pi_*|\pi_{2t}) &\leq H(\pi_*|\pi_0)(1+\kappa)^{-t} \\ &\leq H(\pi_*|R)(1+\kappa)^{-t}. \end{aligned}$$

(b) For $t \geq t_1$,

$$\begin{aligned} H(\pi_*|\pi_{2t}) &\leq \frac{1}{H(\pi_*|\pi_{2t_1})^{-1} + \frac{1}{2}\kappa^2(t-t_1)} \\ &\leq \frac{1}{H(\pi_*|R)^{-1} + \frac{1}{2}\kappa^2(t-t_1)} \\ &\leq 5\frac{C_1^2 \vee H(\pi_*|R)}{t-t_1}. \end{aligned} \tag{4.3}$$

Remark 4.5 (On the Constants)**.** Recall that the general problem (1.1) involves a regularization parameter $\varepsilon > 0$. While we have assumed $\varepsilon = 1$ in the above results, the general case is readily recovered by applying the results to the cost c/ε instead of c. In particular, we can track the dependence of the constants on ε as follows.

(a) Suppose for simplicity that $c \geq 0$. Then (2.5) yields the explicit upper bound

$$H(\pi_*|R) = \mathcal{C}_1(\mu, \nu) + \log \xi \leq \mathcal{C}_1(\mu, \nu) \leq \int c \, d(\mu \otimes \nu)$$

as $\pi := \mu \otimes \nu$ is an admissible control in (2.1). Translating back to (1.1), we observe that this bound scales linearly in the regularization ε^{-1}.

(b) When using the bounds for φ_t, φ_* from Section 5, the constant C_1 of (4.1) scales at most linearly in ε^{-p} for typical examples of costs; e.g., $c(x, y) = d(x, y)^p$ with $p \geq 1$ (see Theorem 5.7). In particular, when ε is small, C_1 is the dominating term in $C_1^2 \vee H(\pi_*|R)$ and in the definition of κ. And most importantly, *the constant in* (4.3) *grows at most like* ε^{-2p}*.*

Proof of Theorem 4.4. We first prove the upper bounds for $H(\pi_*|\pi_{2t})$. As $H(\mu_{2t}|\mu)$ is monotone (Lemma 4.1), we have $H(\mu_{2t}|\mu) \leq 1$ for $t \geq t_0$ and $H(\mu_{2t}|\mu) \geq 1$ for $0 \leq t < t_0$. Consequently, (4.2) yields

$$H(\pi_{2t}|\pi_*) + H(\pi_*|\pi_{2t}) \leq \begin{cases} \frac{3}{2} C_1 H(\mu_{2t}|\mu), & 0 \leq t < t_0, \\ \frac{3}{2} C_1 \sqrt{H(\mu_{2t}|\mu)}, & t \geq t_0. \end{cases}$$

With $t_1 := (t_0 - 1) \vee 0$ it follows that for any $0 < \kappa \leq (\frac{3}{2} C_1)^{-1}$,

$$H(\mu_{2t+2}|\mu) \geq \begin{cases} \kappa H(\pi_*|\pi_{2t+2}), & 0 \leq t < t_1, \\ \kappa^2 H(\pi_*|\pi_{2t+2})^2, & t \geq t_1. \end{cases}$$

Combining this with the last inequality in Lemma 4.2 yields

$$H(\pi_*|\pi_{2t+2}) - H(\pi_*|\pi_{2t}) \leq - \begin{cases} \kappa H(\pi_*|\pi_{2t+2}), & 0 \leq t < t_1, \\ \kappa^2 H(\pi_*|\pi_{2t+2})^2, & t \geq t_1 \end{cases} \tag{4.4}$$

which we can analyze through the lens of difference equations.

(a) Consider $0 \leq t < t_1$ and set $F(t) := H(\pi_*|\pi_{2t})$. As $\kappa \leq (\frac{3}{2} C_1)^{-1}$, (4.4) shows

$$F(t + 1) - F(t) \leq -\kappa F(t + 1).$$

We see by induction that F is dominated by the unique solution G of the difference equation

$$G(t+1) - G(t) = -\kappa G(t+1), \quad 0 \le t < t_1; \quad G(0) = F(0).$$

This solution is explicitly given by

$$G(t) = F(0)(1+\kappa)^{-t}, \quad 0 \le t < t_1.$$

The second inequality follows as $F(0) = H(\pi_*|\pi_0) \le H(\pi_*|\pi_{-1}) = H(\pi_*|R)$ by Lemma 4.1.

(b) Let $t \ge t_1$. Now (4.4) reads

$$F(t+1) - F(t) \le -\kappa^2 F(t+1)^2, \quad t \ge t_1.$$

By an elementary analysis and induction, we see that $F(t)$, $t \ge t_1$ is dominated by the unique nonnegative solution G of the difference equation

$$G(t+1) - G(t) = -\kappa^2 G(t+1)^2, \quad t \ge t_1; \quad G(t_1) = F(t_1).$$

Clearly G is decreasing: $G(t+1) \le G(t) \le G(t_1) = F(t_1)$ and hence

$$G(t+1) = G(t) - \kappa^2 G(t+1)^2 \ge G(t) - \kappa^2 F(t_1) G(t) = \beta G(t)$$

where $\beta := 1 - \kappa^2 F(t_1) > 0$ due to $\kappa^2 \le [2H(\pi_*|R)]^{-1} \le [2F(t_1)]^{-1}$. It follows that G is in turn dominated by the unique solution S of

$$S(t+1) - S(t) = -\gamma S(t) S(t+1), \quad t \ge t_1; \quad S(t_1) = F(t_1)$$

with $\gamma := \kappa^2 \beta = \kappa^2(1 - \kappa^2 F(t_1))$, namely

$$S(t) = \frac{1}{F(t_1)^{-1} + \gamma(t - t_1)}, \quad t \ge t_1.$$

Using again $\kappa^2 \le [2F(t_1)]^{-1}$, we have $\gamma = \kappa^2(1 - \kappa^2 F(t_1)) \ge \kappa^2/2$, yielding the first claim. The other inequalities follow with $F(t_1)^{-1} \ge H(\pi_*|R)^{-1} \ge 0$, again by Lemma 4.1.

(c) It remains to prove the bounds for t_1. With $A := H(\pi_*|R) - H(\pi_0|R)$, it holds that $H(\mu_{2t}|\mu) \le A/t$ for all $t \ge 1$; cf. [45, Corollary 6.12]. Hence $t_0 \le \lceil A \rceil = t_2$ and thus $t_1 \le (t_2 - 1)^+$. On the other hand, suppose for contradiction that $t_3 < t_0$. As $t_3 \ge 2$, this implies $\lfloor t_3/2 \rfloor < t_0 - 1 \le t_1$. Using Proposition 4.3, the bound $H(\pi_*|\pi_{2t}) \le H(\pi_*|\pi_0)(1+\kappa)^{-t}$ for $0 \le t \le t_0$ as proved in (a), and the definition of t_3, we deduce

$$H(\mu_{2t_3}|\mu) \le 2t_3^{-1} H(\pi_*|\pi_{2\lfloor t_3/2 \rfloor}) \le 2t_3^{-1} H(\pi_*|\pi_0)(1+\kappa)^{-\lfloor t_3/2 \rfloor} \le 1,$$

contradicting $t_3 < t_0$. □

While the above analysis crucially depended on the particular monotonicity properties of $H(\pi_*|\pi_{2t})$, we can now deduce rates for the other expressions: the obtained rate $O(t^{-1})$ for $H(\pi_*|\pi_{2t})$ implies the rate $O(t^{-2})$ for the marginals entropies $H(\mu_{2t}|\mu)$ and $H(\mu|\mu_{2t})$, and through the latter, we can further deduce that the reverse entropy $H(\pi_{2t}|\pi_*)$ admits the same rate $O(t^{-1})$ as $H(\pi_*|\pi_{2t})$.

Corollary 4.6. *For $t \geq 2t_1$,*

$$H(\mu_{2t}|\mu) + H(\mu|\mu_{2t}) \leq 10\frac{C_1^2 \vee H(\pi_*|R)}{(\lfloor t/2 \rfloor - t_1)t} = O(t^{-2}),$$
$$H(\pi_{2t}|\pi_*) + H(\pi_*|\pi_{2t}) \leq 5\frac{C_1^2 \vee (H(\pi_*|R)^{1/2}C_1)}{\sqrt{(\lfloor t/2 \rfloor - t_1)t}} = O(t^{-1}).$$

Proof. The first claim follows by combining Proposition 4.3 and Theorem 4.4, and then the second claim follows via (4.2). □

We can state a similar result for $\int(\varphi_* \oplus \psi_* - \varphi_t \oplus \psi_t)\, d(\mu \otimes \nu)$, measuring the "suboptimality" of (φ_t, ψ_t) in the dual problem of (2.1).

Corollary 4.7. *Let*

$$\tilde{C}_1 := \sup_{t \geq 0} \inf_{\alpha > 0} \frac{2}{\alpha}\left(\frac{3}{2} + \log \int e^{\alpha|\varphi_t|}\, d\mu\right) < \infty \tag{4.5}$$

and $\bar{C}_1 := C_1 + \tilde{C}_1$. Then for all $t \geq 2t_1$,

$$0 \leq \int(\varphi_* \oplus \psi_* - \varphi_t \oplus \psi_t)\, d(\mu \otimes \nu) \leq 5\frac{C_1\bar{C}_1 \wedge (H(\pi_*|R)^{1/2}\bar{C}_1)}{\sqrt{(\lfloor t/2 \rfloor - t_1)t}} = O(t^{-1}).$$

Proof. Let $t \geq 0$. Using (2.2), (2.6) and the fact that $\varphi_* \oplus \psi_*$ is the maximizer of the dual problem (e.g., [45]), we have

$$H(\pi_*|R) - H(\pi_{2t}|R) = \int \varphi_* \oplus \psi_*\, d(\mu \otimes \nu) - \int \varphi_t \oplus \psi_t\, d(\mu \otimes \nu) \geq 0.$$

On the other hand, note that $C_1 < \infty$ implies $\varphi_* \in L^1(\mu)$, and then also $\psi_* \in L^1(\nu)$ as $H(\pi_*|R) < \infty$. Lemma 3.6 and (4.2) thus yield

$$\begin{aligned} H(\pi_*|R) - H(\pi_{2t}|R) &\leq H(\pi_*|\pi_{2t}) + \tilde{C}_1\left(\sqrt{H(\mu_{2t}|\mu)} + \frac{1}{2}H(\mu_{2t}|\mu)\right) \\ &\leq \bar{C}_1\left(\sqrt{H(\mu_{2t}|\mu)} + \frac{1}{2}H(\mu_{2t}|\mu)\right). \end{aligned}$$

The claim now follows from Corollary 4.6. □

Recall that the potentials (φ_*, ψ_*) are unique only up to an additive constant. To state any result on the convergence $\varphi_t \to \varphi_*$, we need to choose a particular constant. Indeed, by the monotonicity (2.8), the limit $m := \lim_{t\to\infty} \mu(\varphi_t)$ exists, and we choose the constant such that $\mu(\varphi_*) = m$. Then, the following holds.

Corollary 4.8. *We have $\varphi_t \to \varphi_*$ in $L^p(\mu)$ for every $p \in [1,\infty)$.*

Proof. We have $H(\pi_*|\pi_{2t}) \to 0$ by Theorem 4.4 and hence $\|\pi_{2t} - \pi_*\|_{TV} \to 0$ by Pinsker's inequality. The latter is equivalent to the convergence of the densities, $e^{\varphi_t \oplus \psi_t - c} \to e^{\varphi_* \oplus \psi_* - c}$ in $L^1(\mu \otimes \nu)$. As a consequence, $\varphi_t \oplus \psi_t \to \varphi_* \oplus \psi_*$ in measure $\mu \otimes \nu$. After passing to a subsequence, we deduce that $\varphi_t \oplus \psi_t \to \varphi_* \oplus \psi_*$ pointwise on a Borel set $A \subset \mathsf{X} \times \mathsf{Y}$ with $(\mu \otimes \nu)(A) = 1$. By Fubini's theorem, $\mu\{x \in \mathsf{X} : (x, y_0) \in A\} = 1$ for ν-a.a. $y_0 \in \mathsf{Y}$. Fix any such y_0, then we deduce $\varphi_t(x) + \psi_t(y_0) \to \varphi_*(x) + \psi_*(y_0)$ for μ-a.a. $x \in \mathsf{X}$. To wit, there are finite constants a_t such that $\varphi_t + a_t \to \varphi_*$ μ-a.s. In view of the uniform integrability implied by $C_1 < \infty$, this convergence also holds in $L^p(\mu)$ for any $p \in [1,\infty)$. In particular, $\mu(\varphi_t) + a_t \to \mu(\varphi_*)$. Recalling that φ_* was chosen such that $\mu(\varphi_*) = \lim_t \mu(\varphi_t)$, we see that $\lim_t a_t = 0$ and $\varphi_t \to \varphi_*$ in $L^p(\mu)$ for every p. As φ_* was fixed from the beginning, this must hold for the original sequence $(\varphi_t)_{t\geq 0}$. □

5 Estimates for Conjugate Functions

The main goal of this section is to provide uniform (in t) bounds for the Sinkhorn iterates φ_t, and thus enable the application of Theorem 4.4 to a broad class of cost functions. Our bounds are based on general considerations around (bi)conjugate functions, which will simultaneously yield bounds for Sinkhorn iterates and potentials. While upper bounds are direct from the integrability of the cost function $c(x,y)$, lower bounds use more details about the interaction between the two variables (x,y). More precisely, our results are based on its asymptotic properties, hence are robust with respect to measurable perturbations that are bounded or of lower order growth.

A priori estimates for conjugate (c-convex) functions are very familiar in optimal transport (e.g., [56]). In the context of entropic optimal transport, such ideas have been used in the context of bounded or Lipschitz costs (e.g., [10, 25]); that line of argument is generalized by Lemma 5.9 below but does not apply in the setting of main interest to us. For quadratic cost, [41] bounds the dual potentials in order to bound certain empirical processes. While the quadratic cost is special as it is separable up to an inner product, those bounds are the closest precursors that we are aware of.

Given a measurable function $f : \mathsf{X} \to [-\infty, \infty]$, we define its conjugate $f^c : \mathsf{Y} \to [-\infty, \infty]$ and its biconjugate $f^{cc} : \mathsf{X} \to [-\infty, \infty]$ by

$$f^c(y) = -\log \int e^{f(x)-c(x,y)}\,\mu(dx), \quad y \in \mathsf{Y},$$
$$f^{cc}(x) = -\log \int e^{f^c(y)-c(x,y)}\,\nu(dy), \quad x \in \mathsf{X},$$

provided that the integrals are well defined. We note the abuse of notation: the first conjugation involves $c(\cdot, y)$ and μ while the second involves $c(x, \cdot)$ and ν.[1] The results of this section will be applied in two situations:

- The Sinkhorn iterates satisfy $\psi_t = \varphi_t^c$ and $\varphi_{t+1} = \varphi_t^{cc}$ for $t \geq 0$, by their definition (1.3).
- If (φ_*, ψ_*) are potentials, they solve[2] the Schrödinger system (2.3) which can be restated as $\psi_* = \varphi_*^c$ and $\varphi_* = \varphi_*^{cc}$.

In both situations, the involved integrals are well defined. The common feature is that φ_{t+1}, φ_* are biconjugates of some function f (either φ_t or φ_*). Moreover, we have a priori bounds on quantities such as $\mu(f)$ and $\nu(f^c)$, cf. (2.8)–(2.9), whence we do not mind having such expressions in our estimates below.

Throughout this section, $(\mathsf{X}, d_{\mathsf{X}})$ and $(\mathsf{Y}, d_{\mathsf{Y}})$ are metric spaces with arbitrary but fixed reference points $x_0 \in \mathsf{X}$, $y_0 \in \mathsf{Y}$. In the case of normed spaces, those are usually taken to be the origins. For brevity, we denote $|x| := d_{\mathsf{X}}(x, x_0)$ for $x \in \mathsf{X}$ and $|y| := d_{\mathsf{Y}}(y, y_0)$ for $y \in \mathsf{Y}$, even in the metric case.

Our estimates are based on the interplay of growth properties of c with integrability properties of the marginals. Our main interest is with cost functions of superlinear growth; most practical examples with (sub)linear growth can be covered by a more direct argument detailed in Section 5.1 below.

To fix ideas, consider the example that $c(x,y) = |y-x|^2$ is the quadratic cost on $\mathbb{R} \times \mathbb{R}$ and ν is Gaussian. Then $\int e^{\lambda |y|^2} \nu(dy)$ is infinite for some $\lambda > 0$ yet finite for small enough λ. Thus, the natural growth of the cost may fail to be exponentially integrable (especially once we recall that the cost may be scaled by a large constant ε^{-1} corresponding to the regularization

[1] Moreover, an abuse of terminology: to distinguish from optimal transport theory, one might want to call f^c a soft conjugate or soft c-transform, but we have opted for brevity.

[2] To be precise, while the potentials are only defined up to nullsets, we choose versions satisfying (2.3) everywhere.

parameter in (1.1)). In the example, $c(x,y) = y^2 - 2xy + x^2$, so that we can easily isolate the term y^2 having the critical growth. In general, we will assume that $\int e^{\lambda|y|^p}\nu(dy) < \infty$ for some $\lambda > 0$, and while natural costs then have growth of order p, we will aim to isolate a function $c_2(y)$ such that $c(x,y) - c_2(y)$ has growth of order $< p$ in y. The following is a representative example of what we aim to prove.

Example 5.1. Let $p \in [1,\infty)$ and $\int e^{\lambda|y|^p}\,\nu(dy) < \infty$ for some $\lambda > 0$. Suppose there exists a measurable function $c_2 : \mathsf{Y} \to \mathbb{R}$ such that

$$\begin{aligned} |c(x,y) - c_2(y)| &\leq C(1 + |x||y|^{p-1} + |x|^p), \\ |c_2(y)| &\leq C(1 + |y|^p) \end{aligned} \tag{5.1}$$

for some $C > 0$. Then, with a constant K independent of f,

$$|f^{cc}(x)| \leq |\mu(f)| + |\nu(f^c)| + K(1 + |x|^p).$$

Condition (5.1) is only a special case of the bounds to be handled below, but it already covers the most important examples, as shown by the next two lemmas.

Lemma 5.2. *Let $p \in [1,\infty)$, let $(\mathsf{X}, |\cdot|_{\mathsf{X}})$, $(\mathsf{Y}, |\cdot|_{\mathsf{Y}})$ be Banach spaces, and consider $\mathsf{X} \times \mathsf{Y}$ with a compatible norm $|\cdot|$. Let $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ be Gateaux differentiable with Gateaux derivative satisfying $|D_u c(z)| \leq C(1 + |z|^{p-1})$ for all $u, z \in \mathsf{X} \times \mathsf{Y}$. Then for all $(x,y) \in \mathsf{X} \times \mathsf{Y}$,*

$$\begin{aligned} |c(x,y) - c(0,y)| &\leq 2^{p-1}C(1 + |x|_{\mathsf{X}}|y|_{\mathsf{Y}}^{p-1} + |x|_{\mathsf{X}}^p), \\ |c(0,y)| &\leq |c(0,0)| + C(1 + |y|_{\mathsf{Y}}^p). \end{aligned}$$

Proof. Fix $(x,y) \in \mathsf{X} \times \mathsf{Y}$ and write $x' := (x,0)$, $y' := (0,y)$. A version of the mean value theorem (cf. the proof of [3, Theorem 1.8, p. 13]) shows that for some $\lambda \in [0,1]$,

$$|c(x,y) - c(0,y)| = |c(y' + x') - c(y')| \leq |D_{x'}c(y' + \lambda x')||x'|.$$

As $|y' + \lambda x'| \leq |y'| + |x'| = |y|_{\mathsf{Y}} + |x|_{\mathsf{X}}$, the assumption then yields

$$\begin{aligned} |c(x,y) - c(0,y)| &\leq C(1 + (|y|_{\mathsf{Y}} + |x|_{\mathsf{X}})^{p-1})|x|_{\mathsf{X}} \\ &\leq 2^{p-1}C(1 + |x|_{\mathsf{X}}|y|_{\mathsf{Y}}^{p-1} + |x|_{\mathsf{X}}^p). \end{aligned}$$

Similarly, $|c(0,y) - c(0,0)| \leq |D_{y'}c(\lambda y')||y'| \leq C(1 + |y|_{\mathsf{Y}}^p)$. □

The second example concerns the case $\mathsf{X} = \mathsf{Y}$; exponents $p \leq 1$ are handled in Section 5.1 below.

Lemma 5.3. *Let $p \in [1, \infty)$, let d be a metric on X and $|x| := d(x, x_0)$ for some $x_0 \in \mathsf{X}$. Then for all $x, y \in \mathsf{X}$,*

$$\left|d(x,y)^p - |y|^p\right| \leq p2^{p-1}(|x||y|^{p-1} + |x|^p).$$

Proof. The mean value theorem on $\mathbb{R}$ shows that for any $a, h \in \mathbb{R}$ with $a \geq 0$ and $a + h \geq 0$ there exists $\lambda \in [0, 1]$ with

$$|(a+h)^p - a^p| \leq |p(a+\lambda h)^{p-1}h| \leq (Ca^{p-1} + C|h|^{p-1})|h| \leq Ca^{p-1}|h| + C|h|^p$$

where $C = p2^{p-1}$. Noting that $d(x, y) \leq |x| + |y|$, the above with $a = |y|$ and $h = |x|$ yields

$$d(x,y)^p - |y|^p \leq (|y| + |x|)^p - |y|^p \leq C|x||y|^{p-1} + C|x|^p,$$

which is the desired upper bound. For the lower bound, suppose first that $|x| \leq |y|$ and note that $d(x, y) + |x| \geq d(y, x_0) = |y|$. Then we can apply the above with $a = |y|$ and $h = -|x|$ to find

$$d(x,y)^p - |y|^p \geq (|y| - |x|)^p - |y|^p \geq -C|x||y|^{p-1} - C|x|^p.$$

Whereas if $|x| \geq |y|$, then trivially $d(x, y)^p - |y|^p \geq -|y|^p \geq -|x|^p$. □

Having justified a condition like (5.1), let us now focus on how to obtain the bounds for f^{cc}. We first record an auxiliary result on the equivalence between certain exponential moments and a bound on the moment-generating function, extending a familiar relationship for subgaussian measures (corresponding to $p = 2$ and $q = 1$).

Lemma 5.4. *Let $0 < q < p$. The existence of $\lambda > 0$ such that*

$$K := \int e^{\lambda |y|^p} \nu(dy) < \infty \tag{5.2}$$

is equivalent to the existence of $C_0, C > 0$ such that

$$\int e^{t|y|^q} \nu(dy) \leq C_0\, e^{Ct^{p/(p-q)}}, \quad t \geq 0. \tag{5.3}$$

Moreover, C and C_0 depend only on λ, K, p, q.

Proof. Assuming (5.2), Hölder's inequality yields

$$\int e^{t|y|^q}\nu(dy) \leq \Big(\int e^{2t|y|^q-\lambda|y|^p}\nu(dy)\Big)^{1/2}\Big(\int e^{\lambda|y|^p}\nu(dy)\Big)^{1/2}$$
$$\leq \sqrt{K}\Big(\int e^{2t|y|^q-\lambda|y|^p}\nu(dy)\Big)^{1/2}$$

and calculus shows that the exponent under the integral satisfies

$$\sup_{\xi\geq 0}(2t\xi^q-\lambda\xi^p) = 2Ct^{\frac{p}{p-q}} \quad \text{for} \quad C := (2q/\lambda p)^{\frac{q}{p-q}}(1-q/p).$$

Setting $C_0 = \sqrt{K}$, it follows that (5.3) holds. Conversely, if (5.3) holds and a random variable Y has distribution ν under P, then

$$P\{|Y|\geq s\} = P\{e^{t|Y|^q}\geq e^{ts^q}\} \leq e^{-ts^q}E[e^{t|Y|^q}]$$
$$\leq e^{-ts^q}C_0e^{Ct^{p/(p-q)}} = C_0e^{Ct^{p/(p-q)}-ts^q}.$$

Optimizing the choice of t yields

$$\inf_{t\geq 0}(Ct^{p/(p-q)}-ts^q) = -C's^p \quad \text{for} \quad C' := (p/q)((p-q)/Cp)^{p/q-1}.$$

Thus $P\{|Y|\geq s\}\leq C_0e^{-C's^p}$, or equivalently $P\{|Y|^{p/2}\geq u\}\leq C_0e^{-C'u^2}$. That is, $Y' := |Y|^{p/2}$ is subgaussian, which is equivalent to (5.2); see for instance [55, Proposition 2.5.2, p. 22]. □

We can now state the main tool. As we have opted for a symmetric presentation, its condition involves functions $c_1, a_\pm$ that are partially redundant. The main idea is, still, to isolate a well-chosen function $c_2(y)$ from the cost such as to improve the integrability.

Lemma 5.5. *Fix $p>0$ and suppose that*

$$\int e^{\lambda|y|^p}\,\nu(dy) < \infty \tag{5.4}$$

for some $\lambda>0$. Fix $N\geq 0$ and let $\alpha_i\in[0,p]$, $\beta_i\in[0,p)$, $1\leq i\leq N$ be such that $\alpha_i+\beta_i\leq p$. Set $\tilde\alpha_i := p-\beta_i\geq\alpha_i$. On the strength of Lemma 5.4, there are $C, C_0>0$ with

$$\int e^{t|y|^{\beta_i}}\nu(dy) \leq C_0\,e^{Ct^{p/\tilde\alpha_i}}, \quad t\geq 0 \tag{5.5}$$

for all $1 \le i \le N$. Let c be of the form

$$c(x,y) = c_1(x) + c_2(y) + \hat{c}(x,y)$$

where $c_1 \in L^1(\mu)$, $c_2 \in L^1(\nu)$ and

$$\hat{c}(x,y) \le a_+(x) + \sum_{i=1}^N K_+^i |x|^{\alpha_i} |y|^{\beta_i}, \tag{5.6}$$

$$\hat{c}(x,y) \ge -a_-(x) - \sum_{i=1}^N K_-^i |x|^{\alpha_i} |y|^{\beta_i} \tag{5.7}$$

with $K_\pm^i \in \mathbb{R}_+$ and $a_\pm$ nonnegative measurable functions such that

$$A_+ = \int a_+(x)\, \mu(dx) < \infty. \tag{5.8}$$

Finally, set $A_{\alpha_i} := \int |x|^{\alpha_i}\, \mu(dx)$. Let $f \in L^1(\mu)$ be such that f^c, f^{cc} are well defined. Then

$$f^c(y) - c_2(y) \le A_+ - \mu(f - c_1) + \sum_{i=1}^N K_+^i A_{\alpha_i} |y|^{\beta_i}, \tag{5.9}$$

$$\begin{aligned} f^{cc}(x) - c_1(x) \ge{}& \mu(f - c_1) - A_+ - \log C_0 - a_-(x) \\ &- C \sum_{i=1}^N N^{\frac{p}{\tilde{\alpha}_i} - 1} (K_+^i A_{\alpha_i} + K_-^i |x|^{\alpha_i})^{p/\tilde{\alpha}_i}. \end{aligned} \tag{5.10}$$

Proof. We may assume that $c_1 = 0 = c_2$ and $c = \hat{c}$. By Jensen's inequality,

$$\begin{aligned} f^c(y) &\le -\log \int e^{f(x) - c(x,y)}\, \mu(dx) \\ &\le \int (c(x,y) - f(x))\, \mu(dx) \\ &\le \int \Big(a_+(x) + \sum_{i=1}^N K_+^i |x|^{\alpha_i} |y|^{\beta_i} \Big)\, \mu(dx) - \mu(f) \\ &\le A_+ + \sum_{i=1}^N K_+^i A_{\alpha_i} |y|^{\beta_i} - \mu(f), \end{aligned}$$

proving the upper bound. Using it in the definition of f^{cc} then yields

$$\begin{aligned}
-f^{cc}(x) &= \log \int e^{f^c(y)-c(x,y)}\,\nu(dy) \\
&\leq A_+ - \mu(f) + a_-(x) \\
&\qquad + \log \int \exp\Big(\sum_{i=1}^N K_+^i A_{\alpha_i} |y|^{\beta_i} + \sum_{i=1}^N K_-^i |x|^{\alpha_i} |y|^{\beta_i}\Big)\,\nu(dy).
\end{aligned}$$

We then apply Hölder's inequality and (5.5) to obtain

$$\begin{aligned}
&\int e^{\sum_{i=1}^N (K_+^i A_{\alpha_i} + K_-^i |x|^{\alpha_i})|y|^{\beta_i}}\,\nu(dy) \\
&\qquad\qquad \leq \prod_{i=1}^N \Big(\int e^{N(K_+^i A_{\alpha_i} + K_-^i |x|^{\alpha_i})|y|^{\beta_i}}\,\nu(dy)\Big)^{\frac{1}{N}} \\
&\qquad\qquad \leq C_0 \prod_{i=1}^N e^{CN^{p/\tilde{\alpha}_i - 1}(K_+^i A_{\alpha_i} + K_-^i |x|^{\alpha_i})^{p/\tilde{\alpha}_i}},
\end{aligned}$$

completing the proof. □

Remark 5.6. One can generalize Lemma 5.5 by allowing additional functions $b_\pm(y)$ in (5.6)–(5.7). The upper bound (5.9) then has an additional term $+b_+(y)$. In the proof of the lower bound, an application of Hölder's inequality adds a term $-\log[(\int e^{2(b_-(y)+b_+(y))}\,\nu(dy))^{1/2}]$ in (5.10) and also changes N to $2N$. We have chosen the simpler statement above as the examples of our main interest do not necessitate these additional functions.

The next theorem summarizes our bounds for biconjugate functions.

Theorem 5.7. *Let c and ν satisfy* (5.4)–(5.8)*. Suppose in addition that $c_1, c_2, a_\pm$ have growth of order at most p and that $\int |x|^p\,\mu(dx) < \infty$. Then*

$$|f^{cc}(x)| \leq \mu(f^-) + \nu((f^c)^-) + K(1 + |x|^p)$$

where K is independent of f. If c is replaced by $\tilde{c} := c/\varepsilon$ with $\varepsilon \in (0,1)$, we have a corresponding bound with constant

$$\tilde{K} := \varepsilon^{-\max\{p/\alpha,1\}} K, \quad \textit{where} \quad \alpha := \min_i \tilde{\alpha}_i = p - \max_i \beta_i. \tag{5.11}$$

Proof. The lower bound (5.10) already states

$$f^{cc}(x) - c_1(x) \geq \mu(f - c_1) - K(1 + |x|^p).$$

If we replace c with $\tilde{c} := c/\varepsilon$ for $\varepsilon \in (0,1)$, the constants in (5.10) change by a factor ε^{-1} (except $C, C_0, A_+, A_{\alpha_i}$ that do not depend on the cost function). Thus by inspection of (5.10), an analogous bound holds with the constant $\tilde{K} := \varepsilon^{-\max\{p/\alpha,1\}}K$.

Set $B_+ = \int b_+(x)\,\mu(dx)$ and $B_{\beta_i} = \int |y|^{\beta_i}\,\nu(dy)$, then the upper bound in Lemma 5.5 (applied to f^c instead of f) translates to

$$f^{cc}(x) - c_1(x) \leq B_+ - \nu(f^c - c_2) + a_+(x) + \sum_{i=1}^{N} K_+^i B_{\beta_i} |x|^{\alpha_i}.$$

This can be stated as

$$f^{cc}(x) - c_1(x) \leq -\nu(f^c - c_2) + K'(1 + |x|^p)$$

and here K' scales linearly in ε^{-1}. As c_1, c_2 have growth of order at most p, the theorem follows. □

We note that Example 5.1 is a special case of Theorem 5.7; here $\alpha = 1$ and $a_\pm(x) = C(1 + |x|^p)$ while $c_1 = 0$.

In view of the a priori bounds (2.8)–(2.9), we deduce the following.

Corollary 5.8. *Under the conditions of Theorem 5.7, the Sinkhorn iterates $(\varphi_t)_{t\geq 0}$ and the dual potential φ_* satisfy*

$$|\varphi_t(x)| + |\varphi_*(x)| \leq K(1 + |x|^p)$$

with K independent of t. In particular, if $\int e^{\lambda |x|^p}\,\mu(dx) < \infty$ for some $\lambda > 0$, then the constants C_1 of (4.1) *and $\tilde{C}_1$ of* (4.5) *are finite. If $c \geq 0$, then $K, C_1, \tilde{C}_1$ all admit the scaling behavior* (5.11).[3]

5.1 Results for Linear and Sublinear Growth

While the above results hold for general $p \geq 0$, an easier route is often available for $p \in [0,1]$, at least with some additional structure. The following is a generalization of a familiar result for Lipschitz costs. In contrast to the above, a single conjugation is sufficient to obtain upper and lower bounds in this setting. No exponential integrability is necessary, and the bounds scale linearly in the regularization ε^{-1}. We recall that $\mathcal{W}_1$ denotes the 1-Wasserstein distance.

[3] The condition $c \geq 0$ guarantees that $\log \xi \leq 0$ in (2.8)–(2.9). If costs can be negative, the scaling behavior of ξ needs to be examined separately.

Lemma 5.9. *Let $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ be concave and nondecreasing.*[4] *Suppose*

$$|c(x,y_1) - c(x,y_2)| \leq \omega(d_{\mathsf{Y}}(y_1,y_2)) \quad \text{for all} \quad y_1, y_2 \in \mathsf{Y}, \quad x \in \mathsf{X}. \tag{5.12}$$

Then $|f^c(y_1) - f^c(y_2)| \leq \omega(d_{\mathsf{Y}}(y_1,y_2))$ and, with $B_1 := \int |y| \, \nu(dy)$,

$$|f^c(y) - \nu(f^c)| \leq \omega(\mathcal{W}_1(\delta_y, \nu)) \leq \omega(B_1 + |y|) \leq \omega(B_1) + \omega(|y|).$$

Proof. Let $y_1, y_2 \in \mathsf{Y}$. As $|c(x,y_1) - c(x,y_2)| \leq \omega(d_{\mathsf{Y}}(y_1,y_2))$,

$$\begin{aligned} -f^c(y_2) &= \log \int e^{f(x) - c(x,y_2)} \, \mu(dx) \\ &\leq \log \int e^{f(x) - c(x,y_1) + \omega(d_{\mathsf{Y}}(y_1,y_2))} \, \mu(dx) \\ &= \omega(d_{\mathsf{Y}}(y_1,y_2)) - f^c(y_1). \end{aligned}$$

The first claim follows as y_1 and y_2 are interchangeable. Using Jensen's inequality, we deduce that

$$\left| f^c(y) - \int f^c \, d\nu \right| \leq \int |f^c(y) - f^c(y')| \, \nu(dy') \leq \int \omega(d_{\mathsf{Y}}(y,y')) \, \nu(dy')$$

which, by concavity and monotonicity of ω, is bounded by

$$\omega\left(\int d_{\mathsf{Y}}(y,y') \, \nu(dy') \right) = \omega(\mathcal{W}_1(\delta_y, \nu)) \leq \omega(B_1 + |y|).$$

Note that ω is subadditive by concavity and $\omega(0) \geq 0$. □

Corollary 5.10. *Let μ, ν have finite first moments and let c satisfy* (5.12). *Then the Sinkhorn iterates $(\varphi_t)_{t \geq 0}$ and the dual potential φ_* satisfy*

$$|\varphi_t(x)| + |\varphi_*(x)| \leq K + \omega(|x|)$$

with K independent of t. In particular, if $\int e^{\lambda \omega(|x|)} \, \mu(dx) < \infty$ for some $\lambda > 0$, then the constants C_1 of (4.1) *and $\tilde{C}_1$ of* (4.5) *are finite. If $c \geq 0$, then $K, \omega, C_1, \tilde{C}_1$ scale at most linearly with the cost c.*

The following complements Lemma 5.3 for $p \leq 1$.

[4] Note that ω need not be a modulus of continuity: it is not assumed that $\omega(0) = 0$. For instance, all bounded costs satisfy the condition (5.12).

Example 5.11. Let d be a measurable metric on $\mathsf{X} = \mathsf{Y}$ and $|x| := d(x, x_0)$ for some $x_0 \in \mathsf{X}$. Let $A_1 := \int |x| \, \mu(dx) < \infty$, $B_1 := \int |y| \, \nu(dy) < \infty$ and

$$c(x, y) = d(x, y)^p, \quad \text{where} \quad p \in [0, 1].$$

By Lemma 5.9, any (bi)conjugate function admits the modulus of continuity $\omega(s) = s^p$ and

$$|f^c(y) - \nu(f^c)| \leq B_1^p + |y|^p, \qquad |f^{cc}(x) - \mu(f^{cc})| \leq A_1^p + |x|^p.$$

If $\int e^{\lambda |x|^p} \, \mu(dx) < \infty$ for some $\lambda > 0$, Corollary 5.10 implies that C_1 of (4.1) and $\tilde{C}_1$ of (4.5) are finite with bounds growing at most linearly in the regularization ε^{-1}.

Funding, Conflicts and Data Availability. Promit Ghosal was supported by NSF Grant DMS-2153661. Marcel Nutz was supported by NSF Grants DMS-1812661 and DMS-2106056. The authors have no relevant financial or non-financial interests to disclose. Data and code sharing is not applicable to this article as no data or code were generated or analyzed.

References

[1] 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.

[2] 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.

[3] A. Ambrosetti and G. Prodi. *A primer of nonlinear analysis*, volume 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1995.

[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] E. Bayraktar, S. Eckstein, and X. Zhang. Stability and sample complexity of divergence regularized optimal transport. *Bernoulli*, 31(1):213–239, 2025.

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

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

[8] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. Towards optimal running times for optimal transport. *Oper. Res. Lett.*, 52:Paper No. 107054, 8, 2024.

[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, L. Chizat, and M. Laborde. Displacement smoothness of entropic optimal transport. *ESAIM Control Optim. Calc. Var.*, 30:Paper No. 25, 24, 2024.

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

[14] 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.

[15] 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.

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

[17] A. Chiarini, G. Conforti, G. Greco, and L. Tamanini. Gradient estimates for the Schrödinger potentials: convergence to the Brenier map and quantitative stability. *Comm. Partial Differential Equations*, 48(6):895–943, 2023.

[18] L. Chizat, A. Delalande, and T. Vaškevičius. Sharper exponential convergence rates for Sinkhorn's algorithm in continuous settings. *Preprint arXiv:2407.01202v1*, 2024.

[19] 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.

[20] G. Conforti, A. Durmus, and G. Greco. Quantitative contraction rates for Sinkhorn algorithm: beyond bounded costs and compact marginals. *Preprint arXiv:2304.04451v1*, 2023.

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

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

[23] G. Deligiannidis, V. de Bortoli, and A. Doucet. Quantitative uniform stability of the iterative proportional fitting procedure. *Ann. Appl. Probab.*, 34(1A):501–516, 2024.

[24] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statistics*, 11:427–444, 1940.

[25] 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.

[26] S. Eckstein. Hilbert's projective metric for functions of bounded growth and exponential convergence of Sinkhorn's algorithm. *Preprint arXiv:2311.04041v1*, 2023.

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

[28] S. Eckstein and M. Nutz. Convergence rates for regularized optimal transport via quantization. *Math. Oper. Res.*, 49(2):1223–1240, 2024.

[29] 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.

[30] R. Fortet. Résolution d'un système d'équations de M. Schrödinger. *J. Math. Pures Appl.*, 19:83–105, 1940.

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

[32] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

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

[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] C. T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

[36] S. Kullback. Probability densities with given marginals. *Ann. Math. Statist.*, 39:1236–1243, 1968.

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

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

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

[40] C. Léonard. Revisiting Fortet's proof of existence of a solution to the Schrödinger system. *Preprint arXiv:1904.13211v1*, 2019.

[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] L. Nenna and P. Pegon. Convergence rate of entropy-regularized multimarginal optimal transport costs. *Canad. J. Math.*, pages 1–21, 2024.

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

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

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

[48] S. Pal. On the difference between entropic cost and the optimal transport cost. *Ann. Appl. Probab.*, 34(1B):1003–1028, 2024.

[49] 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.

[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] E. Schrödinger. Über die Umkehrung der Naturgesetze. Sitzungsberichte Preuss. Akad. Wiss. *Akad. Wiss., Berlin. Phys. Math.*, 144:144–153, 1931.

[53] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

[54] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

[55] R. Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018.

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

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