Introduction to Entropic Optimal Transport

Marcel Nutz[*]

This version: December 5, 2022

Abstract

This text develops mathematical foundations for entropic optimal transport and Sinkhorn's algorithm in a self-contained yet general way. It is a revised version of lecture notes from a course given in Paris during the fall of 2021; some parts date back to an earlier course at Columbia University in 2020.

Contents

[*] Columbia University, Depts. of Statistics and Mathematics, mnutz@columbia.edu. Research partially supported by an Alfred P. Sloan Fellowship and NSF Grants DMS-1812661, DMS-2106056. Much of this work was completed while the author was serving as Columbia–Ecole Polytechnique Alliance Professor; support from the Columbia Alliance program is gratefully acknowledged. For their kind hospitality, encouragement and numerous suggestions that helped to improve these notes, the author would like to thank Guillaume Carlier, Giovanni Conforti, Stephan Eckstein, AYeong Lee, Flavien Léger, Christian Léonard, Gilles Pagès, Nizar Touzi, Johannes Wiesel, Long Zhao.

Introduction

Applications of optimal transport are thriving in areas such as machine learning, statistics, economics or image processing. Regularization plays a key role in enabling efficient algorithms with provable convergence (see [29] for a recent monograph with numerous references). Entropic regularization is the most popular choice as it allows for Sinkhorn's algorithm (also called iterative proportional fitting procedure, IPFP) that can be implemented at large scale and is analytically tractable. The entropically regularized transport problem may be formulated as

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{Y}} c(x,y)\,\pi(dx,dy) + \varepsilon H(\pi|\mu\otimes\nu). \tag{1}$$

Here $\Pi(\mu,\nu)$ is the set of couplings of the given marginal probability measures μ,ν on spaces X,Y. In the first term, $c:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}$ is the cost function; the most important example is quadratic cost $c(x,y)=\|x-y\|^2$ on $\mathbb{R}^d\times\mathbb{R}^d$. In the penalization term, $H(\,\cdot\,|\mu\otimes\nu)$ denotes the relative entropy with respect to the product measure $\mu\otimes\nu$ and $\varepsilon>0$ is the regularization parameter. The basic idea is to solve this "entropic" optimal transport problem for small $\varepsilon>0$ to obtain an approximation of the (unregularized) optimal transport problem that corresponds to $\varepsilon=0$.

The problem (1) is also of its own interest; i.e., without letting $\varepsilon\to 0$. On the one hand, applied researchers have started to exploit numerous benefits resulting from regularization (such as smoothness, statistical properties, etc.), so that the regularization is sometimes seen as an advantage rather than an approximation error (e.g., [21, 30]). On the other hand, entropic optimal transport can be seen as a special case of the (static) Schrödinger bridge problem that has a long history in physics. Indeed, (1) can be translated into an equivalent *static Schrödinger bridge problem*

$$\inf_{\pi\in\Pi(\mu,\nu)} H(\pi|R) \tag{2}$$

by introducing the auxiliary reference measure R with

$$dR \propto e^{-c/\varepsilon} d(\mu\otimes\nu). \tag{3}$$

We do not discuss Schrödinger's original dynamic problem in this text but refer the interested reader to the surveys [17, 26].

The convex minimization (2) falls into the class of *entropy minimization problems* of the general form

$$\inf_{Q\in\mathcal{Q}} H(Q|R)$$

with a convex set $\mathcal{Q}$. Following [10], such problems admit a simple and elegant general theory which we introduce in Section 1: existence and uniqueness of a minimizer, characterization by a first-order condition, and other properties. In Section 2 we apply this theory to the Schrödinger bridge problem (2) where $\mathcal{Q}$ is the set of couplings and work out the corresponding characterization—the optimal density is given by so-called Schödinger potentials; roughly speaking, these are the Lagrange multipliers for the marginal constraints. The potentials can be characterized as the solution of a system of two equations, the so-called *Schrödinger system.* We focus on the case (3) of interest to us, where R is equivalent to $\mu \otimes \nu$ and moreover often $c \in L^1(\mu \otimes \nu)$. This allows us to achieve fairly general results while avoiding some of the difficulties known in the theory of more general Schrödinger bridges (see [3, 4, 18, 33, 34], among others). The potentials can also be seen as the solution to a *dual problem* in the sense of convex analysis; this is detailed in Section 3. Section 4 translates the results from Schrödinger bridges to entropic optimal transport via (3) and adds another basic observation: the potentials inherit regularity from the cost function c. Indeed, the Schrödinger system can be seen as a conjugacy relation reminiscent of the notion of c-convexity in optimal transport theory and allows for various types of a priori estimates. Section 5 studies the convergence of entropic optimal transport (1) to (standard) optimal transport as the regularization parameter $\varepsilon \to 0$. Specifically, we are interested in the convergence of the optimal value (1), the optimal couplings and the Schrödinger potentials. We conclude with Section 6 on Sinkhorn's algorithm. Here we first derive the convergence of the marginal distributions and a general bound for their convergence rate, then continue with select results on the convergence of the couplings and potentials. While the convergence properties are reasonably well understood for bounded cost functions, the unbounded case is only partially understood.

1 Entropy

Consider a measurable space $(\Omega, \mathcal{F})$ and denote by $\mathcal{P}(\Omega)$ its collection of probability measures. In what follows, whenever a probability named R is specified, $E[\cdot] = E^R[\cdot] = \int \cdot \, dR$ denotes the corresponding expectation, whereas other measures are indicated explicitly.

Definition 1.1. Given $Q, R \in \mathcal{P}(\Omega)$, the *entropy of Q relative to R* is

$$H(Q|R) = \begin{cases} E^Q[\log \frac{dQ}{dR}], & Q \ll R, \\ \infty, & Q \not\ll R. \end{cases}$$

Relative entropy is also called *Kullback–Leibler divergence.* If $Q \ll R$, we can write

$$H(Q|R) = E[h(dQ/dR)], \quad h(z) := z \log z$$

as an integral under the reference measure R. Here and below, the convention $0 \times (\pm\infty) := 0$ is used. Noting that $h : [0,\infty] \to [-e^{-1}, \infty]$ is strictly convex and using Jensen's inequality,

$$Q \mapsto H(Q|R) \quad \text{is nonnegative and convex,}$$

and strictly convex on the set where it is finite. Clearly $H(Q|R) = 0$ if and only if $Q = R$.

1.1 Basic Properties

We define the total variation distance between $P, Q \in \mathcal{P}(\Omega)$ as

$$\|P - Q\|_{TV} := \int \left| \frac{dP}{dR} - \frac{dQ}{dR} \right| dR$$

for an arbitrary measure $R \gg P, Q$; that is, the L^1-distance of their densities. Two other representations are

$$\|P - Q\|_{TV} = \sup_{|\phi| \leq 1 \text{ mbl.}} \int \phi \, d(P - Q) = 2 \sup_{A \subset \Omega \text{ mbl.}} \big(P(A) - Q(A)\big).$$

(An equally popular definition is to divide the right-hand side by 2, to normalize $\|P - Q\|_{TV} \leq 1$.)

Lemma 1.2 (Pinsker's Inequality)**.** *The total variation distance of $Q, R \in \mathcal{P}(\Omega)$ satisfies*

$$\|Q - R\|_{TV} \leq \sqrt{2H(Q|R)}.$$

Proof. We may assume that $H(Q|R) < \infty$. We have $3(x-1)^2 \leq f(x)g(x)$ for $f(x) = 4 + 2x$ and $g(x) = x \log x - x + 1$, or equivalently $\sqrt{3}|x - 1| \leq f(x)^{1/2} g(x)^{1/2}$. Denoting $Z = dQ/dR$, this yields

$$\begin{aligned} 3\|Q - R\|_{TV}^2 &= E[\sqrt{3}|Z - 1|]^2 \leq E[f(Z)^{1/2} g(Z)^{1/2}]^2 \\ &\leq E[f(Z)]E[g(Z)] = (4 + 2E[Z])E[Z \log Z] = 6H(Q|R), \end{aligned}$$

where Hölder's inequality was used to pass to the second line. □

When Ω is Polish (i.e., a completely metrizable topological space), we always assume that $\mathcal{F}$ is the Borel σ-field and endow $\mathcal{P}(\Omega)$ with the associated weak convergence topology (in the probabilistic sense): $Q_n \to Q$ weakly means that $E^{Q_n}[f] \to E^Q[f]$ for all $f \in C_b(\Omega)$, where $C_b(\Omega)$ denotes the space of bounded continuous functions.

Lemma 1.3. *Given $Q, R \in \mathcal{P}(\Omega)$, we have the variational representations*

$$H(Q|R) = \sup_{\phi\colon \Omega \to \mathbb{R} \text{ bdd. mbl.}} \big(E^Q[\phi] - \log E^R[e^\phi]\big) \tag{1.1}$$

$$= \sup_{\phi\colon \Omega \to \mathbb{R} \text{ mbl., } E^R[e^\phi]<\infty} \big(E^Q[\phi] - \log E^R[e^\phi]\big). \tag{1.2}$$

In particular, the function $(Q,R) \mapsto H(Q|R)$ is jointly convex and jointly lower semicontinuous wrt. convergence in variation. If Ω is Polish, this also holds wrt. weak convergence.

Proof. Fix $Q \ll R$. Consider $0 < \psi \in L^1(R)$ and define $dR' = \alpha^{-1}\psi\, dR$, where $\alpha := E^R[\psi] > 0$ is the normalizing constant. Then $\frac{dQ}{dR} = \frac{dQ}{dR'}\alpha^{-1}\psi$ and hence

$$H(Q|R) = H(Q|R') + E^Q[\log\psi] - \log\alpha \geq E^Q[\log\psi] - \log E^R[\psi].$$

This shows in particular that

$$H(Q|R) \geq \sup_{\psi} \big(E^Q[\log\psi] - \log E^R[\psi]\big) \tag{1.3}$$

where the supremum is taken over functions $\psi > 0$ that are bounded and bounded away from zero. Consider $\psi_0 := dQ/dR \in L^1(R)$. Then $\psi_n := (1/n) \vee \psi_0 \wedge n$ is such a function and dominated convergence yields

$$E^Q[\log\psi_n] - \log E^R[\psi_n] \to E^Q[\log\psi_0] - \log E^R[\psi_0] = H(Q|R),$$

showing that equality holds in (1.3). Writing $\phi = \log\psi$, we have (1.1). Note that the set over which the supremum is taken no longer depends on Q and R, and that (1.1) also holds when $Q \not\ll R$. Moreover, the function $(Q,R) \mapsto (E^Q[\phi] - \log E^R[e^\phi])$ is convex and continuous wrt. convergence in variation.[1] It follows that the supremum $H(Q|R)$ is convex and lower semicontinuous as claimed.

When Ω is a metric space, $C_b(\Omega)$ is dense in $L^1(\mu)$ for any Borel measure μ. Hence the supremum in (1.1) can be further restricted to continuous ϕ, and then the weak lower semicontinuity follows in the same way.

[1] As the function ϕ is fixed, this also holds under a significantly weaker convergence: we could weaken "convergence in variation" to "set-wise convergence" in Lemma 1.3.

Finally, we show (1.2). Here "$\leq$" is trivial. For the proof of "$\geq$" we may assume that $H(Q|R) < \infty$ and hence $Q \ll R$, but in that case we may use (1.3) with $0 < \psi := e^{\phi} \in L^1(R)$. $\square$

We observe that (1.2) implies a way to infer integrability under Q from exponential integrability under R: given a real function φ and $\beta > 0$ with $e^{\beta\varphi} \in L^1(R)$,

$$E^Q[\varphi] \leq \frac{1}{\beta}\Big(\log E^R[e^{\beta\phi}] + H(Q|R)\Big) \tag{1.4}$$

by using $\phi := \beta\varphi$ in (1.2).

For the remainder of the section, we fix the reference $R \in \mathcal{P}(\Omega)$. Moreover, $x^+ := \max\{x, 0\}$ and $x^- := \max\{-x, 0\}$ denote the positive and negative parts of x.

Lemma 1.4. *Let $Q, Q' \in \mathcal{P}(\Omega)$ and $Q' \ll R$.*

(a) If $H(Q|R) < \infty$, then $(\log \frac{dQ'}{dR})^+ \in L^1(Q)$ and $E^Q[\log \frac{dQ'}{dR}] \leq H(Q|R)$.

(b) If either $H(Q|R) < \infty$ or $H(Q|Q') < \infty$, then

$$H(Q|R) - H(Q|Q') = E^Q[\log \tfrac{dQ'}{dR}].$$

Proof. (a) Let $H(Q|R) < \infty$, then in particular $Q \ll R$. Consider the Lebesgue decomposition $Q' = Q'_1 + Q'_2$ into $Q'_1 \ll Q$ and $Q'_2 \perp Q$. Using the inequality $\log x \leq x - 1$ with $x = z'/z$ for $z' \geq 0$ and $z > 0$ yields $\log z' \leq \log z + z'/z - 1$. Evaluating this at $z' = \frac{dQ'_1}{dR}$ and $z = \frac{dQ}{dR}$ then shows

$$Q\text{-a.s.,} \quad \log \tfrac{dQ'}{dR} = \log \tfrac{dQ'_1}{dR} \leq \log \tfrac{dQ}{dR} + \tfrac{dQ'_1}{dR} / \tfrac{dQ}{dR} - 1 = \log \tfrac{dQ}{dR} + \tfrac{dQ'_1}{dQ} - 1.$$

It follows that $(\log \frac{dQ'}{dR})^+ \in L^1(Q)$ and

$$E^Q[\log \tfrac{dQ'}{dR}] \leq H(Q|R) + Q'_1(\Omega) - 1 \leq H(Q|R).$$

(b) *Case 1:* $H(Q|Q') < \infty$. Then $\log \frac{dQ}{dQ'} \in L^1(Q)$, justifying $(*)$ in

$$\begin{aligned} H(Q|R) = E^Q[\log \tfrac{dQ}{dR}] &= E^Q[\log(\tfrac{dQ}{dQ'} \tfrac{dQ'}{dR})] = E^Q[\log(\tfrac{dQ}{dQ'}) + \log(\tfrac{dQ'}{dR})] \\ &\overset{(*)}{=} E^Q[\log \tfrac{dQ}{dQ'}] + E^Q[\log \tfrac{dQ'}{dR}] = H(Q|Q') + E^Q[\log \tfrac{dQ'}{dR}]. \end{aligned} \tag{1.5}$$

Case 2: $H(Q|Q') = \infty$. Then $H(Q|R) < \infty$ and hence $(\log \frac{dQ'}{dR})^+ \in L^1(Q)$ by (a), so that $E^Q[\log \frac{dQ'}{dR}]$ is well-defined. We need to show that $E^Q[\log \frac{dQ'}{dR}] = -\infty$, or equivalently that $E^Q[(\log \frac{dQ'}{dR})^-] = \infty$.

Case (i): $Q \ll Q'$. Suppose for contradiction that $(\log \frac{dQ'}{dR})^- \in L^1(Q)$. Then $\log \frac{dQ'}{dR} \in L^1(Q)$ and $(*)$ in (1.5) is again justified. Now (1.5) implies $H(Q|R) = \infty$, a contradiction.

Case (ii): $Q \not\ll Q'$. Then $Q\{\frac{dQ'}{dR} = 0\} > 0$ and thus $E^Q[(\log \frac{dQ'}{dR})^-] = \infty$ as desired. □

Lemma 1.5. *Let $Q, Q' \in \mathcal{P}(\Omega)$ satisfy $Q \ll Q' \ll R$. Then*

$$\left\| \log \frac{dQ'}{dR} - \log \frac{dQ}{dR} \right\|_{L^1(Q)} \leq \sqrt{8H(Q|Q')} + H(Q|Q').$$

Proof. Let $Z' = dQ'/dR$ and $Z = dQ/dR$. Then

$$\begin{aligned} E^Q|\log Z' - \log Z| = E^Q \left| \log \frac{Z'}{Z} \right| &= 2E^Q \left[\log^+ \frac{Z'}{Z} \right] - E^Q \left[\log \frac{Z'}{Z} \right] \\ &= 2E^Q \left[\log^+ \frac{Z'}{Z} \right] + H(Q|Q'). \end{aligned}$$

In view of the elementary inequality $\log^+ x \leq |x - 1|$,

$$E^Q \left[\log^+ \frac{Z'}{Z} \right] \leq E^Q \left| \frac{Z'}{Z} - 1 \right| = E^R|Z' - Z| = \|Q' - Q\|_{TV}.$$

Combining this with the above, we find that

$$E^Q|\log Z' - \log Z| \leq 2\|Q' - Q\|_{TV} + H(Q|Q').$$

It remains to apply Pinsker's inequality $\|Q' - Q\|_{TV} \leq \sqrt{2H(Q|Q')}$ from Lemma 1.2. □

The next inequality, sometimes called data processing inequality or contraction of relative entropy, expresses the idea that transforming data does not increase the amount of information contained in it, or more precisely, increase the ability to distinguish between two probability distributions.

Lemma 1.6 (Data Processing)**.** *Let $P, Q \in \mathcal{P}(\Omega)$ and $K : \Omega \to \mathcal{P}(\Omega')$ a kernel.*[2] *Let $P' \in \mathcal{P}(\Omega')$ be the second marginal of $P \otimes K \in \mathcal{P}(\Omega \times \Omega')$ and similarly Q' for Q. Then*

$$H(P'|Q') \leq H(P|Q).$$

[2] Throughout this text, kernel stands for Markov kernel.

Proof. We may assume that $P \ll Q$. For any kernels $K_1 \ll K_2 : \Omega \to \mathcal{P}(\Omega')$,

$$\frac{d(P \otimes K_1)}{d(Q \otimes K_2)}(\omega, \omega') = \frac{dP}{dQ}(\omega)\frac{dK_1(\omega)}{dK_2(\omega)}(\omega') \quad Q \otimes K_2\text{-a.s.} \tag{1.6}$$

In particular, $\frac{d(P \otimes K)}{d(Q \otimes K)}(\omega, \omega') = \frac{dP}{dQ}(\omega)$ and thus the definition of H implies

$$H(P|Q) = H(P \otimes K|Q \otimes K).$$

It remains to show that $H(P \otimes K|Q \otimes K) \geq H(P'|Q')$.

1. Proof via Jensen's Inequality. Jensen's inequality for $h(x) = x \log x$ and (1.6) and yield

$$\begin{aligned} H(P \otimes K_1|Q \otimes K_2) &= \iint h\left(\frac{dP}{dQ}(\omega)\frac{dK_1(\omega)}{dK_2(\omega)}(\omega')\right) K_2(\omega, d\omega')Q(d\omega) \\ &\geq \int h\left(\frac{dP}{dQ}(\omega)\right) Q(d\omega) = H(P|Q). \end{aligned} \tag{1.7}$$

Denote by $P' \otimes K_1'$ the "reverse" disintegration of $P \otimes K$ from the second marginal to the first, meaning that $K_1' : \Omega' \to \mathcal{P}(\Omega)$ is the conditional distribution of the first marginal given the second (which we tacitly assume to exist). Similarly, $Q \otimes K = Q' \otimes K_2'$. Applying (1.7) to $P' \otimes K_1'$ and $Q' \otimes K_2'$,

$$H(P \otimes K|Q \otimes K) = H(P' \otimes K_1'|Q' \otimes K_2') \geq H(P'|Q').$$

The only drawback of this (standard) argument is that existence of disintegrations is a nontrivial issue that requires some assumptions in the probability space. The following is less intuitive but more elementary.

2. Proof via Variational Representation. The definition of P' means that

$$\int_{\Omega'} \psi(\omega')\, P'(d\omega') = \int_{\Omega \times \Omega'} \psi(\omega')\, P \otimes K(d\omega, d\omega')$$

for any bounded measurable function $\psi : \Omega' \to \mathbb{R}$, where on the right-hand side we see ψ as a special case of a function $\phi : \Omega \times \Omega' \to \mathbb{R}$ that depends only on one variable, $\psi(\omega') = \phi(\omega, \omega')$. We use this for both P', ψ and Q', e^{ψ} in the variational representation (1.1):

$$\begin{aligned} H(P \otimes K|Q \otimes K) &= \sup_{\phi:\, \Omega \times \Omega' \to \mathbb{R} \text{ bdd. mbl.}} \left(E^{P \otimes K}[\phi] - \log E^{Q \otimes K}[e^{\phi}]\right) \\ &\geq \sup_{\psi:\, \Omega' \to \mathbb{R} \text{ bdd. mbl.}} \left(E^{P \otimes K}[\psi] - \log E^{Q \otimes K}[e^{\psi}]\right) \\ &= \sup_{\psi:\, \Omega' \to \mathbb{R} \text{ bdd. mbl.}} \left(E^{P'}[\psi] - \log E^{Q'}[e^{\psi}]\right) \\ &= H(P'|Q'). \end{aligned}$$

□

One special case of Lemma 1.6 occurs when the kernel is deterministic: given a measurable map $T : \Omega \to \Omega'$, we can consider the kernel $K(\omega, d\omega') = \delta_{T(\omega)}(d\omega')$. Then P' is sometimes called the *push-forward* of P under T and denoted $T_{\#}P$ (or $T_{*}P$), especially in analysis. (In probabilistic terms, $T_{\#}P$ is the *law* of the random variable T under P). In this case, the data processing inequality reads

$$H\big(T_{\#}P\big|T_{\#}Q\big) \leq H(P|Q). \tag{1.8}$$

Example 1.7. If P, Q are probabilities on a product space $\mathsf{X} \times \mathsf{Y}$ and P', Q' are their marginal distributions on X, then $H(P'|Q') \leq H(P|Q)$. To see this, we use the deterministic kernel $(x, y) \mapsto \delta_x$ in Lemma 1.6, where $\Omega = \mathsf{X} \times \mathsf{Y}$ and $\Omega' = X$. Or equivalently, we take T in (1.8) to be the projection $(x, y) \mapsto x$.

1.2 Minimizing Entropy

The following compactness property of sets with bounded entropy is due to the superlinear growth of $h(x) = x \log x$.

Lemma 1.8. *Let $(Q_n)_{n\geq 1}$ satisfy $\sup_n H(Q_n|R) < \infty$. Then there are convex combinations $Q'_n \in \operatorname{conv}\{Q_n, Q_{n+1}, \dots\}$ that converge in variation.*

Proof. Let $Z_n := dQ_n/dR$. As $E[h(Z_n)]$ is bounded and h has superlinear growth, the la Vallée–Poussin theorem [2, Theorem 4.5.9, p. 272] shows that $(Z_n)_{n\geq 1}$ is uniformly integrable. By the Dunford–Pettis theorem [2, Theorem 4.7.18, p. 285] this is equivalent to weak precompactness in $L^1(R)$ and $(Z_n)_{n\geq 1}$ has a subsequence that converges weakly in $L^1(R)$—i.e., relative to the topology $\sigma(L^1(R), L^\infty(R))$.[3] Mazur's lemma [31, Theorem 3.13, p. 67] then yields convex combinations $Z'_n \in \operatorname{conv}\{Z_n, Z_{n+1}, \dots\}$ that converge (strongly) in $L^1(R)$, meaning that the corresponding convex combinations $Q'_n \in \operatorname{conv}\{Q_n, Q_{n+1}, \dots\}$ converge in variation. □

The compactness property in Lemma 1.8 shows that for (Q_n) such that $\lim_n H(Q_n|R)$ exists, suitable convex combinations Q'_n converge in total variation, but it is silent about the original sequence. We may note that if $Q_n \to Q$ in variation and $\lim H(Q_n|R) = H(Q|R) =: a$, then necessarily $\lim H(Q'_n|R) = a$ whenever $Q'_n \in \operatorname{conv}\{Q_n, Q_{n+1}, \dots\}$. Next, we show by a geometric argument that this equality of limits is sufficient for convergence of the original sequence.

[3] In fact, this is another noteworthy compactness property of $(Q_n)_{n\geq 1}$: there is a subsequence that converges set-wise.

Lemma 1.9. *Let $Q_n \in \mathcal{P}(\Omega)$. Suppose that $\lim_n H(Q_n|R) =: a \in \mathbb{R}$ exists and that $\limsup_{m,n\to\infty} H(Q_{m,n}|R) \geq a$ for $Q_{m,n} := (Q_m + Q_n)/2$.*[4] *Then (Q_n) converges in variation.*

Proof. We first note the "parallelogram identity" which follows directly from Definition 1.1:

$$H(Q_m|R) + H(Q_n|R) = 2H(Q_{m,n}|R) + H(Q_m|Q_{m,n}) + H(Q_n|Q_{m,n}).$$

By the assumption, taking $\limsup_{m,n\to\infty}$ on both sides shows that the last two terms converge to zero. Using

$$\|Q_m - Q_n\|_{TV} \leq \|Q_m - Q_{m,n}\|_{TV} + \|Q_n - Q_{m,n}\|_{TV}$$

as well as Pinsker's inequality (Lemma 1.2), we deduce the Cauchy property $\lim_{m,n} \|Q_m - Q_n\|_{TV} = 0$. □

We can now establish existence and uniqueness of a minimizer for $H(\,\cdot\,|R)$ within a suitable set $\mathcal{Q}$, sometimes called the *entropic projection* of R onto $\mathcal{Q}$.[5]

Theorem 1.10. *Let $\emptyset \neq \mathcal{Q} \subseteq \mathcal{P}(\Omega)$ be convex and closed in variation, and suppose that $\mathcal{Q}_{fin} := \{Q \in \mathcal{Q} :\, H(Q|R) < \infty\} \neq \emptyset$.*

(a) There exists a unique $Q_* \in \mathcal{Q}$ such that

$$H(Q_*|R) = \inf_{Q\in\mathcal{Q}} H(Q|R) \in [0, \infty).$$

Moreover, $Q_* \gg Q$ for any $Q \in \mathcal{Q}_{fin}$. In particular, if there exists $Q \in \mathcal{Q}_{fin}$ with $Q \sim R$, then $Q_* \sim R$.

(b) $Q_0 \in \mathcal{Q}$ is the minimizer Q_* if and only if $Z_0 := dQ_0/dR$ exists and[6]

$$E^Q[\log Z_0] \geq H(Q_0|R) \quad \textit{for all} \quad Q \in \mathcal{Q}_{fin}. \tag{1.9}$$

Proof. Let $Q_n \in \mathcal{Q}$ be such that $H(Q_n|R) \to a := \inf_{Q\in\mathcal{Q}} H(Q|R)$. By convexity we have $Q_{m,n} = (Q_m + Q_n)/2 \in \mathcal{Q}$ and hence $H(Q_{m,n}|R) \geq a$ for all m, n. Lemma 1.9 now shows that (Q_n) has a limit Q_* in variation, and Q_* is a minimizer by the lower semicontinuity of $H(\,\cdot\,|R)$; cf. Lemma 1.3. Uniqueness holds due to the strict convexity of $H(\,\cdot\,|R)$.

[4] As $\liminf_{m,n\to\infty} H(Q_{m,n}|R) \leq a$ by convexity, this condition is actually equivalent to $\lim_{m,n\to\infty} H(Q_{m,n}|R) = a$.

[5] Or I-projection, following [10] which used I instead of H to denote relative entropy.

[6] We have $E^Q[(\log Z_0)^+] < \infty$ for all $Q \in \mathcal{Q}_{fin}$ by Lemma 1.4 (a), so that the integral $E^Q[\log Z_0] \in [-\infty, \infty)$ is well-defined as soon as Z_0 exists.

Let $Q_0 \in \mathcal{Q}$ satisfy $Q_0 \ll R$, and $Q_1 \in \mathcal{Q}_{fin}$. For $\lambda \in [0,1]$, consider $Q_\lambda = \lambda Q_1 + (1-\lambda)Q_0$ and $Z_\lambda = dQ_\lambda/dR$. As $\lambda \mapsto h(Z_\lambda)$ is convex, its difference quotient decreases monotonically to $\partial_\lambda|_{\lambda=0+} H(Q_\lambda|R)$ as $\lambda \downarrow 0$,

$$h(Z_1) - h(Z_0) \geq \frac{h(Z_\lambda) - h(Z_0)}{\lambda} \downarrow (Z_1 - Z_0)h'(Z_0) = (Z_1 - Z_0)(1 + \log Z_0).$$

If $Q_0 \in \mathcal{Q}_{fin}$, the left-hand side is integrable and monotone convergence yields

$$\begin{aligned} \partial_\lambda|_{\lambda=0+} H(Q_\lambda|R) &= \partial_\lambda|_{\lambda=0+} E[h(Z_\lambda)] = E[(Z_1 - Z_0)h'(Z_0)] \\ &= E[(Z_1 - Z_0)(1 + \log Z_0)] \\ &= E^R[Z_1 \log Z_0] - H(Q_0|R) \in [-\infty, \infty). \end{aligned} \tag{1.10}$$

This identity remains valid for $Q_0 \in \mathcal{Q} \setminus \mathcal{Q}_{fin}$ if the derivative on the left-hand side is interpreted as $-\infty$.

Suppose that Q_0 is the minimizer Q_*. Then $\partial_\lambda|_{\lambda=0+} H(Q_\lambda|R) \geq 0$ and we conclude that

$$Z_1 \log Z_0 \in L^1(R) \quad \text{and} \quad E^R[Z_1 \log Z_0] \geq H(Q_0|R), \tag{1.11}$$

which is (1.9). Note also that $Q_1 \not\ll Q_0$ would imply $R\{Z_0 = 0, Z_1 > 0\} = R\{Z_1 \log Z_0 = -\infty\} > 0$, contradicting (1.11).

Conversely, if $Q_0 \in \mathcal{Q}$ satisfies (1.9) and $Q \in \mathcal{Q}_{fin}$ is arbitrary, then Lemma 1.4 (b) shows the equality in

$$H(Q|R) \geq H(Q|R) - H(Q|Q_0) = E^Q[\log \tfrac{dQ_0}{dR}] \geq H(Q_0|R). \tag{1.12}$$

In particular, $H(Q_0|R)$ is minimal. □

The last assertion in Theorem 1.10 (a) confirms the intuition that Q_*, being the "most diffuse" (relative to R) measure in $\mathcal{Q}$, should have the largest support among all $Q \in \mathcal{Q}$ with $Q \ll R$. Analytically, the reason is that $h'(0) = 1 + \log 0 = -\infty$; i.e., increasing the value of a vanishing density by a small amount leads to a large reduction in entropy.

As seen in the proof, Theorem 1.10 (b) is a variational first-order condition, stating that the directional derivative of the cost functional at the minimizer should be nonnegative in all admissible directions. In regular cases, we may expect that the minimizer is an interior solution and the derivative is zero in all admissible directions. (Indeed, it holds for the case of Schrödinger bridges; cf. the discussion following Proposition 2.17.) That corresponds to (1.9) holding with equality, or equivalently to $Q \mapsto E^Q[\log \frac{dQ_0}{dR}]$ being constant on $\mathcal{Q}_{fin}$. The latter will be used as a condition in several statements below.

Remark 1.11. The "if" part of Theorem 1.10 (b) holds as soon as $\mathcal{Q}_{fin} \neq \emptyset$, even for not necessarily convex or closed $\mathcal{Q}$—those conditions were not used in the proof. The "only if" part used convexity, though not closedness.

The following Pythagorean-type relationship reflects the strict convexity of the entropy minimization problem: if $H(Q|R)$ is close to $\inf_{\mathcal{Q}} H(\,\cdot\,|R)$, then Q is close to the minimizer Q_*.

Corollary 1.12. *Let $\mathcal{Q} \subset \mathcal{P}(\Omega)$ be convex. If Q_* minimizes $H(\,\cdot\,|R)$ over $\mathcal{Q}$, then*

$$H(Q|Q_*) \leq H(Q|R) - H(Q_*|R) \quad \textit{for all} \quad Q \in \mathcal{Q}_{fin}.$$

If $Q \mapsto E^Q[\log \frac{dQ_*}{dR}]$ is constant on $\mathcal{Q}_{fin}$, the above holds with equality.

Proof. Both assertions follow from (1.12) with $Q_0 = Q_*$ and Remark 1.11. The second assertion is also immediate from Lemma 1.4 (b) alone. □

The optimal log-density has the following integrability property.

Corollary 1.13. *Let $\mathcal{Q} \subset \mathcal{P}(\Omega)$ be convex. If Q_* minimizes $H(\,\cdot\,|R)$ over $\mathcal{Q}$, then*

$$\log \frac{dQ_*}{dR} \in L^1(Q) \quad \textit{for all} \quad Q \in \mathcal{Q}_{fin}.$$

Proof. Let $Z_* = dQ_*/dR$ and $Q \in \mathcal{Q}_{fin}$. As already stated in the footnote of Theorem 1.10, Lemma 1.4 (a) ensures that $E^Q[(\log Z_*)^+] < \infty$. On the other hand, (1.9) with $Q_0 = Q_*$ clearly implies $E^Q[\log Z_*] > -\infty$. □

The next corollary is stated merely for emphasis; it is obtained by specializing the inequality in (1.9) to an equality and recalling Remark 1.11.

Corollary 1.14. *Let $Q_0 \in \mathcal{Q} \subset \mathcal{P}(\Omega)$ and $\mathcal{Q}_{fin} \neq \emptyset$. If*

$$Q \mapsto E^Q\big[\log \tfrac{dQ_0}{dR}\big] \quad \textit{is constant on } \mathcal{Q}_{fin} \cup \{Q_0\},$$

then $Q_0 \in \mathcal{Q}_{fin}$ and $Q_0 \in \arg\min_{\mathcal{Q}} H(\,\cdot\,|R)$.

As mentioned above, the constancy in Corollary 1.14 corresponds to the directional derivative of the entropy at Q_0 being zero, for all directions within $\mathcal{Q}_{fin}$. The following sufficient condition for optimality is more general, and easier to verify in practice, as it only asks for an approximating sequence with a constancy as in Corollary 1.14.

Proposition 1.15. *Let $Q_0 \in \mathcal{Q} \subset \mathcal{P}(\Omega)$ satisfy $Q_0 \ll R$ and consider the log-density $\zeta := \log \frac{dQ_0}{dR}$. Suppose there exist $\zeta_n \in L^1(Q_0)$ such that* *(i)* $E^{Q}[\zeta_n] = E^{Q_0}[\zeta_n]$ *for all* $Q \in \mathcal{Q}_{fin}$*,*

(ii) $\limsup_n E^{Q_0}[\zeta_n] \geq E^{Q_0}[\zeta]$,

(iii) $\limsup_n E^{R}[e^{\zeta_n}] \leq 1$.

Then $H(Q_0|R) = \inf_{Q\in\mathcal{Q}} H(Q|R) \in [0,\infty]$*. In particular, if* $\mathcal{Q}_{fin} \neq \emptyset$*, then* Q_0 *is a minimizer.*

Proof. The claim is trivial if $\mathcal{Q}_{fin} = \emptyset$. Fix an arbitrary $Q \in \mathcal{Q}_{fin}$, set $Z := dQ/dR$ and $Z_0 := dQ_0/dR = \exp(\zeta)$. The convex function $\hbar(x) := h(x) - x = x\log x - x$ has Fenchel conjugate $\hbar^*(y) = \sup_x[xy - \hbar(x)] = e^y$, which yields Fenchel's inequality

$$\hbar(Z) \geq \zeta_n Z - \hbar^*(\zeta_n). \tag{1.13}$$

As $E^R[\zeta_n Z] = E^Q[\zeta_n] = E^{Q_0}[\zeta_n]$ by (i), taking expectations in (1.13) yields

$$H(Q|R) - 1 = E^R[\hbar(Z)] \geq E^R[\zeta_n Z] - E^R[\hbar^*(\zeta_n)] = E^{Q_0}[\zeta_n] - E^R[e^{\zeta_n}]$$

and then (ii) and (iii) allow us to conclude that

$$H(Q|R)-1 \geq \limsup E^{Q_0}[\zeta_n] - \limsup E^R[e^{\zeta_n}] \geq E^{Q_0}[\zeta] - 1 = H(Q_0|R)-1.$$

□

Remark 1.16. (a) Often the ζ_n in Proposition 1.15 are chosen as log-densities of some probabilities Q_n, and then (iii) is trivial as $E^R[e^{\zeta_n}] = 1$. In fact, the ζ_n can always be normalized in this fashion: an equivalent way to state the proposition would be to assume $E^R[e^{\zeta_n}] = 1$ and omit (iii).

(b) Proposition 1.17 below will imply that the sufficient condition of Proposition 1.15 is also necessary in many cases. In Proposition 1.17, $Q_0 = Q_*$ is the minimizer and we construct Q_n with log-density $\zeta_n = \log \frac{dQ_n}{dR}$ satisfying $\zeta_n \to \zeta$ in $L^1(Q_0)$, which of course implies (ii) and (iii). In many cases, we can construct the approximation such that ζ_n also satisfy (i).

The next result serves two purposes. First, we want to describe the minimizer Q_* over $\mathcal{Q}$ through an approximation with measures Q_n that can be constructed more explicitly. Indeed, most convex sets $\mathcal{Q}$ of interest can be characterized through countably many linear constraints, and then a natural polyhedral approximation $\mathcal{Q}_n$ can be found by enforcing only the first n constraints. In this context, one may be able to determine $Q_n := \arg\min_{\mathcal{Q}_n} H(\,\cdot\,|R)$ by elementary means, and then the next result yields that $Q_n \to Q_*$ and also that the log-densities converge. We will apply this idea to study the structure of Schrödinger bridges in Section 2.3.1. A second purpose is to relate the condition $\mathcal{Q}_{fin} \neq \emptyset$ to the convergence of the approximation. This will not be used until Section 5.2; for now the reader can focus on a situation where $\mathcal{Q}_{fin} \neq \emptyset$ is given.

Proposition 1.17. *Consider a decreasing sequence of sets $\mathcal{Q}_n \subset \mathcal{P}(\Omega)$ that are convex and closed in variation, and let $\mathcal{Q} := \cap_n \mathcal{Q}_n$. Suppose that $\mathcal{Q}_{n,fin} \neq \emptyset$ and let $Q_n = \arg\min_{\mathcal{Q}_n} H(\,\cdot\,|R)$ be the minimizer over $\mathcal{Q}_n$.*[7] *The following are equivalent:*

(i) $\mathcal{Q}_{fin} \neq \emptyset$,

(ii) Q_n *converge in variation and* $H(\lim_n Q_n|R) < \infty$,

(iii) $\lim_n H(Q_n|R) < \infty$.

If these conditions are satisfied, then

$$Q_n \to Q_* \quad \textit{in variation} \qquad \textit{and} \qquad H(Q_n|R) \to H(Q_*|R), \tag{1.14}$$

where $Q_* = \arg\min_{\mathcal{Q}} H(\,\cdot\,|R)$. Moreover, $Q_n \gg Q_{n+1} \gg Q_*$ as well as

$$H(Q_*|Q_n) \to 0 \qquad \textit{and} \qquad \log\frac{dQ_n}{dR} \to \log\frac{dQ_*}{dR} \quad \textit{in} \quad L^1(Q_*). \tag{1.15}$$

Proof. The inclusion $\mathcal{Q}_n \supset \mathcal{Q}_{n+1} \supset \mathcal{Q}$ implies that $a_n := H(Q_n|R)$ is increasing and $a_n \leq a_* := \inf_{Q\in\mathcal{Q}} H(Q|R)$, so that $a := \lim a_n \leq a_*$ and in particular (i)⇒(iii). For $m \geq n$ we have $Q_{m,n} := (Q_m + Q_n)/2 \in \mathcal{Q}_n$ and thus $H(Q_{m,n}|R) \geq a_n$, hence $\limsup_{m,n\to\infty} H(Q_{m,n}|R) \geq a$. If $a < \infty$, then Lemma 1.9 yields that Q_n converge in variation to some limit Q. As $Q \in \cap_n \mathcal{Q}_n = \mathcal{Q}$ and

$$H(Q|R) \leq \lim a_n = a \leq a_*$$

due to Lemma 1.3, we see that $Q \in \mathcal{Q}_{fin}$ and $Q \in \arg\min_{\mathcal{Q}} H(\,\cdot\,|R)$, so that $Q = Q_*$ by the uniqueness in Theorem 1.10 (note that $\mathcal{Q}$ is closed and convex, as an intersection of such sets). In particular, we have shown both (iii)⇒(i) and (iii)⇒(ii). On the other hand, $Q \in \cap_n \mathcal{Q}_n$ implies $H(Q|R) \geq a_n$ for all n and hence $H(Q|R) = \lim a_n$. This shows (ii)⇒(iii) as well as (1.14). Finally, Theorem 1.10 and $\mathcal{Q}_n \supset \mathcal{Q}_{n+1} \supset \mathcal{Q}$ also yield the stated absolute continuity.

It remains to prove (1.15). We use Corollary 1.12 for the problem over $\mathcal{Q}_n$ (where Q_n is the minimizer and $Q_* \in \mathcal{Q}_{n,fin}$ is a suboptimal measure) to find

$$H(Q_*|Q_n) \leq H(Q_*|R) - H(Q_n|R).$$

[7] A unique minimizer exists by Theorem 1.10. One can note that $\mathcal{Q}_{n,fin} \neq \emptyset$ is implied by $\mathcal{Q}_{fin} \neq \emptyset$.

As $H(Q_n|R) \to H(Q_*|R)$ was already shown, we deduce $H(Q_*|Q_n) \to 0$. The second part of (1.15) then follows by Lemma 1.5. □

In the application of Proposition 1.17 to Schrödinger bridges in Section 2.3.1, each of the sets $\mathcal{Q}_n$ is of the form considered in the next example, given by n linear equality constraints.

Example 1.18. Given bounded measurable functions $\phi_1, \dots, \phi_n : \Omega \to \mathbb{R}$, let

$$\mathcal{Q} = \{Q \in \mathcal{P}(\Omega) : E^Q[\phi_i] = 0,\ 1 \leq i \leq n\}.$$

Assume that $\mathcal{Q}_{fin} \neq \emptyset$. As $\mathcal{Q}$ is convex and closed, Theorem 1.10 then yields a unique minimizer $Q_* \in \mathcal{Q}$. We claim that Q_* is uniquely characterized (within $\mathcal{Q}$) by having a density of the form

$$\frac{dQ_*}{dR} = a\exp(b_1\phi_1 + \cdots + b_n\phi_n) \qquad \text{for some } b_i \in \mathbb{R} \text{ and } a > 0. \tag{1.16}$$

Sufficiency of (1.16) is immediate from Corollary 1.14; we need to show that there exists $Q_0 \in \mathcal{Q}$ of the form (1.16). To this end, we construct the minimizer using Lagrange multipliers. We confine ourselves to a sketch and refer to [19, Section 3, esp. Corollary 3.25] for a detailed treatment. Indeed, $\mathcal{Q}_{fin} \neq \emptyset$ implies that for all $b = (b_1, \dots, b_n) \in \mathbb{R}^n$,

$$R\{b \cdot \Phi > 0\} > 0 \implies R\{b \cdot \Phi < 0\} > 0, \tag{1.17}$$

where $\Phi = (\phi_1, \dots, \phi_n)$ and $\cdot$ is the Euclidean inner product. This can be used to show that the finite-dimensional concave optimization problem

$$\max_{b \in \mathbb{R}^n} E^R[-\exp(b \cdot \Phi)]$$

has a solution b. (This is known as the exponential utility maximization problem in financial economics, where (1.17) is interpreted as absence of arbitrage opportunities.) Moreover, the maximizer b satisfies the first-order condition $E^R[-\phi_i \exp(b \cdot \Phi)] = 0$ for $i = 1, \dots, n$. We can now define Q_0 via $\frac{dQ_0}{dR} = a\exp(b \cdot \Phi)$, where $a > 0$ is the normalizing constant. Then Q_0 is of the form (1.16) and the first-order condition states that $Q \in \mathcal{Q}$ as desired. While Corollary 1.14 directly implies that $Q_0 \in \mathcal{Q}$ has minimal entropy, we mention that this also follows through convex duality along the lines of Lemma 1.3. See [19, Section 3] for a detailed discussion which also shows that (1.17) implies $\mathcal{Q}_{fin} \neq \emptyset$ as $|\Phi|$ is bounded, or more generally whenever $E^R[e^{r|\Phi|}] < \infty$ for all $r > 0$. A different, more abstract proof of (1.16) is given in [10, Theorem 3.1].

2 Static Schrödinger Bridges

Let $(\mathsf{X},\mathcal{F}_{\mathsf{X}},\mu)$ and $(\mathsf{Y},\mathcal{F}_{\mathsf{Y}},\nu)$ be separable[8] probability spaces, $\mathsf{X}\times\mathsf{Y}$ their product (endowed with the product σ-field $\mathcal{F}_{\mathsf{X}}\otimes\mathcal{F}_{\mathsf{Y}}$) and $R\in\mathcal{P}(\mathsf{X}\times\mathsf{Y})$ a given reference probability measure. We denote by $\Pi(\mu,\nu)\subset\mathcal{P}(\mathsf{X}\times\mathsf{Y})$ the set of couplings; that is, the set of all $Q\in\mathcal{P}(\mathsf{X}\times\mathsf{Y})$ satisfying

$$\int_{\mathsf{X}\times\mathsf{Y}} f(x)\,Q(dx,dy)=\int_{\mathsf{X}} f(x)\,\mu(dx),\quad \int_{\mathsf{X}\times\mathsf{Y}} g(y)\,Q(dx,dy)=\int_{\mathsf{Y}} g(y)\,\nu(dy) \tag{2.1}$$

for all bounded measurable $f:\mathsf{X}\to\mathbb{R}$ and $g:\mathsf{Y}\to\mathbb{R}$. We observe that $\Pi(\mu,\nu)$ is convex and closed in variation, putting us in the framework of Theorem 1.10 with $\mathcal{Q}:=\Pi(\mu,\nu)$. Its requirement that $\Pi_{fin}(\mu,\nu):=\mathcal{Q}_{fin}$ is nonempty—i.e., that there exists $\pi\in\Pi(\mu,\nu)$ with $H(\pi|R)<\infty$—will of course depend on the choice of R. A simple sufficient condition is

$$R\gg\mu\otimes\nu \quad\text{and}\quad \log\frac{d(\mu\otimes\nu)}{dR}\in L^1(\mu\otimes\nu), \tag{2.2}$$

as this is equivalent to $\mu\otimes\nu\in\Pi_{fin}(\mu,\nu)$. We write $(\varphi\oplus\psi)(x,y):=\varphi(x)+\psi(y)$ for functions $\varphi:\mathsf{X}\to[-\infty,\infty)$ and $\psi:\mathsf{Y}\to[-\infty,\infty)$.

Theorem 2.1. *Let $\Pi_{fin}(\mu,\nu)\neq\emptyset$. Then there is a unique coupling*

$$\pi_*=\operatorname*{arg\,min}_{\Pi(\mu,\nu)} H(\,\cdot\,|R),$$

called the (static) Schrödinger bridge *from μ to ν.*

(a) Let $R\sim\mu\otimes\nu$. Then there are measurable functions $\varphi:\mathsf{X}\to\mathbb{R}$ and $\psi:\mathsf{Y}\to\mathbb{R}$, called Schrödinger potentials, *such that*

$$\frac{d\pi_*}{dR}=e^{\varphi\oplus\psi}\quad R\text{-a.s.}$$

The potentials are a.s. unique up to an additive constant.[9]

[8] A probability space $(\mathsf{X},\mathcal{F}_{\mathsf{X}},\mu)$ is called separable if there is a countable family $(A_n)\subset\mathcal{F}_{\mathsf{X}}$ such that for every $A\in\mathcal{F}_{\mathsf{X}}$ and $\varepsilon>0$, there exists n with $\mu(A\triangle A_n)<\varepsilon$. This property holds if and only if $L^1(\mathsf{X},\mathcal{F}_{\mathsf{X}},\mu)$ is separable (consider simple functions based on (A_n) or see [2, Exercise 4.7.64, p. 307]). All probability spaces of interest to us are separable; some very general sufficient conditions are detailed in [2, Section 7.14(iv), p. 132].

[9] I.e., if φ',ψ' are potentials, then $\varphi'=\varphi+a$ μ-a.s. and $\psi'=\psi-a$ ν-a.s. for some $a\in\mathbb{R}$.

(b) *Conversely, let $\pi_0 \in \Pi(\mu,\nu)$ admit a density of the form*

$$\frac{d\pi_0}{dR} = e^{\varphi\oplus\psi} \quad R\text{-a.s.}$$

for some measurable functions $\varphi : \mathsf{X} \to [-\infty,\infty)$, $\psi : \mathsf{Y} \to [-\infty,\infty)$. Then π_0 is the Schrödinger bridge.[10]

If (2.2) holds, then $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$.

Remark 2.2. In Theorem 2.1 (b), it is important that $\log \frac{d\pi_0}{dR} = \varphi\oplus\psi$ holds R-a.s. rather than merely π_0-a.s. Indeed, let $\mathsf{X} = \mathsf{Y}$ be a finite set with uniform measure $\mu = \nu$ and $R = \mu\otimes\nu$. Clearly the Schrödinger bridge is given by $\pi_* = R$. If π_0 is the identical coupling (i.e., the uniform distribution on the diagonal $\{(x,x) : x \in \mathsf{X}\}$), then $d\pi_0/dR = \exp(\varphi\oplus\psi)$ holds π_0-a.s. for $\varphi = 0$ and $\psi = \log(1/|\mathsf{X}|)$, but $\pi_0 \neq \pi_*$.

Occasionally we want to apply Theorem 2.1 (b) in a setting where it is not known a priori that $\Pi_{fin}(\mu,\nu) \neq \emptyset$. The following variant includes a sufficient condition for that.

Corollary 2.3. *Let $\pi_0 \in \Pi(\mu,\nu)$ admit a density of the form $\frac{d\pi_0}{dR} = e^{\varphi\oplus\psi}$ R-a.s. for some measurable functions $\varphi : \mathsf{X} \to [-\infty,\infty)$, $\psi : \mathsf{Y} \to [-\infty,\infty)$ satisfying $(\varphi\oplus\psi)^+ \in L^1(\mu\otimes\nu)$. Then $\pi_0 \in \Pi_{fin}(\mu,\nu)$ and π_0 is the Schrödinger bridge. Moreover, $(\varphi,\psi) \in L^1(\mu)\times L^1(\nu)$ and $H(\pi_0|R) = \mu(\varphi)+\nu(\psi)$.*

Proof. Lemma 2.23 will show that $(\varphi,\psi) \in L^1(\mu)\times L^1(\nu)$ and $H(\pi_0|R) = \mu(\varphi)+\nu(\psi)$. Thus $\pi_0 \in \Pi_{fin}(\mu,\nu) \neq \emptyset$ and now Theorem 2.1 (b) applies. □

Before proving Theorem 2.1, we detail three more corollaries. First, we emphasize a direct consequence of the fact that $\frac{d\pi_*}{dR} > 0$ in Theorem 2.1 (a); compare also Theorem 1.10.

Corollary 2.4. *If $\Pi_{fin}(\mu,\nu) \neq \emptyset$ and $R \sim \mu\otimes\nu$, the Schrödinger bridge satisfies $\pi_* \sim \mu\otimes\nu$. In particular, if there exists any coupling with finite entropy, there also exists a coupling with finite entropy that is equivalent to $\mu\otimes\nu$.*

[10] While $\pi_0 \in \Pi_{fin}(\mu,\nu)$ is not assumed a priori, it is part of the conclusion. Similarly, φ,ψ are a priori allowed to take the value $-\infty$, but Lemma 2.14 below shows that φ,ψ are necessarily finite a.s.

2.1 Schrödinger Equations

Next, we characterize the Schrödinger potentials as the solution to a system of two equations, the Schrödinger system. These equations will be used in Section 6 to define the iterates in Sinkhorn's algorithm.

Let $R \ll \mu \otimes \nu$, then $\frac{dR}{d(\mu\otimes\nu)}$ exists and we may define a measurable function $c : \mathsf{X} \times \mathsf{Y} \to (-\infty, \infty]$ via

$$e^{-c(x,y)} = \frac{dR}{d(\mu \otimes \nu)}.$$

(There is no particular necessity to write the density in exponential form; our notation is merely chosen to resemble the setting of entropic optimal transport in Section 4 below.) In most cases of interest to us, we have $R \sim \mu \otimes \nu$ and then c is $\mathbb{R}$-valued. For measurable functions $\varphi : \mathsf{X} \to [-\infty, \infty)$ and $\psi : \mathsf{Y} \to [-\infty, \infty)$, we study the so-called Schrödinger equations

$$\varphi(x) = -\log \int_{\mathsf{Y}} e^{\psi(y)-c(x,y)}\, \nu(dy) \quad \mu\text{-a.s.,} \tag{SE1}$$

$$\psi(y) = -\log \int_{\mathsf{X}} e^{\varphi(x)-c(x,y)}\, \mu(dx) \quad \nu\text{-a.s.} \tag{SE2}$$

Consider the measure $\pi(\varphi, \psi)$ defined by

$$d\pi(\varphi, \psi) := e^{\varphi \oplus \psi}\, dR = e^{\varphi \oplus \psi - c}\, d(\mu \otimes \nu).$$

Recalling that the marginal density is obtained by integrating the joint density over the other marginal, we see that

$$(\mathsf{SE1}) \quad \Longleftrightarrow \quad \text{the first marginal of } \pi(\varphi, \psi) \text{ is } \mu, \tag{2.3}$$

$$(\mathsf{SE2}) \quad \Longleftrightarrow \quad \text{the second marginal of } \pi(\varphi, \psi) \text{ is } \nu. \tag{2.4}$$

If $\pi(\varphi, \psi) \in \Pi(\mu, \nu)$, it follows that (φ, ψ) is a solution of (SE1)–(SE2). That is assertion (a) below, whereas (b) is a consequence of Theorem 2.1 (b).

Corollary 2.5 (Schrödinger Equations). *Let $e^{-c} = dR/d(\mu \otimes \nu)$.*

(a) If (φ, ψ) are Schrödinger potentials, then (φ, ψ) solve the Schrödinger equations (SE1)–(SE2)*.*

(b) Let $\varphi : \mathsf{X} \to [-\infty, \infty)$ and $\psi : \mathsf{Y} \to [-\infty, \infty)$ be measurable functions. If (φ, ψ) solve (SE1)*, then*

$$d\pi(\varphi, \psi) := e^{\varphi \oplus \psi}\, dR$$

defines a probability measure $\pi(\varphi, \psi)$ whose first marginal is μ. Denote by ν' its second marginal. If $\Pi_{fin}(\mu, \nu') \neq \emptyset$, then $\pi(\varphi, \psi)$ is the Schrödinger bridge from μ to ν'. The analogue holds for (SE2).

In particular, if (φ, ψ) solve (SE1)–(SE2) *and $\Pi_{fin}(\mu, \nu) \neq \emptyset$, then $\pi(\varphi, \psi)$ is the Schrödinger bridge from μ to ν. If $R \sim \mu \otimes \nu$, it follows that the solution of* (SE1)–(SE2) *is a.s. unique up to an additive constant.*

In Remark 3.4, we will further portray the Schrödinger system as the Euler–Lagrange equations (i.e., variational first-order conditions) describing the optimality of the potentials in a maximization problem.

2.2 Cyclical Invariance

Theorem 2.1 shows the relation between optimality of a coupling and the decomposition of its density as a product $e^{\varphi(x)}e^{\varphi(y)}$. In this section we introduce a reformulation for the existence of a decomposition that will be useful in the context of passing to limits (see Sections 5 and 6).

Definition 2.6. A probability measure $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is called cyclically invariant (with respect to R) if $\pi \sim R$ and its density $\frac{d\pi}{dR}$ admits a version $Z : \mathsf{X} \times \mathsf{Y} \to (0, \infty)$ such that for all $k \geq 2$,

$$\prod_{i=1}^{k} Z(x_i, y_i) = \prod_{i=1}^{k} Z(x_i, y_{i+1}) \quad \text{for all} \quad (x_i, y_i)_{i=1}^{k} \in (\mathsf{X} \times \mathsf{Y})^k, \tag{2.5}$$

with the cyclical convention $y_{k+1} := y_1$.

Similarly to the decomposition into potentials, one can see (2.5) as a first-order condition of optimality. Indeed, for a discrete problem where $\mathsf{X} = \{x_i, 1 \leq i \leq k\}$ and $\mathsf{Y} = \{y_i, 1 \leq i \leq k\}$, one can find by elementary perturbation arguments that the optimal density has to verify (2.5).

Lemma 2.7. *Let $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ satisfy $\pi \sim R$. Then π is cyclically invariant if and only if*

$$\frac{d\pi}{dR} = e^{\varphi \oplus \psi} \quad R\text{-a.s.} \tag{2.6}$$

for some measurable functions $\varphi : \mathsf{X} \to \mathbb{R}$, $\psi : \mathsf{Y} \to \mathbb{R}$.

Proof. Let (2.6) hold and $Z := e^{\varphi \oplus \psi}$. Then (2.5) boils down to

$$\exp\left(\sum_{i=1}^{k} \varphi(x_i) + \psi(y_i)\right) = \exp\left(\sum_{i=1}^{k} \varphi(x_i) + \psi(y_{i+1})\right)$$

which holds by simply rearranging the sum.

Conversely, let $Z : \mathsf{X}\times\mathsf{Y} \to (0,\infty)$ satisfy (2.5) and fix an arbitrary point $(x_*, y_*) \in \mathsf{X}\times\mathsf{Y}$. Define the measurable functions $\varphi : \mathsf{X} \to \mathbb{R}$, $\psi : \mathsf{Y} \to \mathbb{R}$ via $e^{\psi(y)} = Z(x_*, y)$ and $e^{\varphi(x)} = Z(x, y_*)e^{-\psi(y_*)}$, then $e^{\varphi(x_*)} = 0$ and

$$e^{\varphi(x)+\psi(y)} = \frac{Z(x,y_*)Z(x_*,y)}{e^{\psi(y_*)}} = \frac{Z(x_*,y)Z(x,y_*)}{Z(x_*,y_*)} = Z(x,y)$$

where the last equality holds due to (2.5) with $k = 2$ and $(x_1, y_1) = (x, y)$ and $(x_2, y_2) = (x_*, y_*)$. □

Remark 2.8. More generally, the Borwein–Lewis theorem [3, Theorem 3.3] states that a function $Z : S \to (0,\infty)$ on an arbitrary subset $S \subset \mathsf{X}\times\mathsf{Y}$ can be decomposed as $Z = e^{\varphi\oplus\psi}$ if and only if it satisfies a relation similar to (2.5). The proof is particularly simple when $S = \mathsf{X}\times\mathsf{Y}$ as in Lemma 2.7.

In view of Lemma 2.7, the following is a special case of Theorem 2.1.

Corollary 2.9. *Let $\Pi_{fin}(\mu,\nu) \neq \emptyset$ and $R \sim \mu\otimes\nu$. Then $\pi \in \Pi(\mu,\nu)$ is the Schrödinger bridge if and only if it is cyclically invariant.*

2.3 Proof of Theorem 2.1

In Theorem 2.1, existence and uniqueness of π_* follow immediately from Theorem 1.10. We start with the proof of part (a) in the next section, then continue with the verification part (b) and end with the integrability of the potentials.

2.3.1 Existence and Uniqueness of the Decomposition

Suppose that X, Y are separable. Then instead of using all bounded measurable functions f, g to define $\Pi(\mu,\nu)$ in (2.1), it suffices to test against suitable countable dense families. Indeed, we can find $(f_i)_{i\geq 1} \subset L^\infty(\mu)$ and $(g_i)_{i\geq 1} \subset L^\infty(\nu)$ such that $Q \in \Pi(\mu,\nu)$ if and only if

$$\int_{\mathsf{X}\times\mathsf{Y}} f_i(x)\, Q(dx,dy) = 0, \qquad \int_{\mathsf{X}\times\mathsf{Y}} g_i(y)\, Q(dx,dy) = 0, \quad i \geq 1. \tag{2.7}$$

This enables a natural approximation of $\Pi(\mu,\nu)$ defined by finitely many linear constraints: let $\mathcal{Q}_n$ be the set of all $Q \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$ satisfying (2.7) for $1 \leq i \leq n$ (instead of all $i \geq 1$). Then $\mathcal{Q}_n$ is convex, closed in variation and $\cap_n \mathcal{Q}_n = \Pi(\mu,\nu) = \mathcal{Q}$ as in Proposition 1.17, which shows that

$$\arg\min_{\mathcal{Q}_n} H(\,\cdot\,|R) =: \pi_n \to \pi_* \quad \text{in variation,}$$

or equivalently, $d\pi_n/dR \to d\pi_*/dR$ in $L^1(R)$. On the other hand, we obtain from Example 1.18 that

$$\frac{d\pi_n}{dR} = \exp(\varphi_n \oplus \psi_n)$$

for some bounded measurable functions $\varphi_n : \mathsf{X} \to \mathbb{R}$ and $\psi_n : \mathsf{Y} \to \mathbb{R}$; namely, φ_n and ψ_n are linear combinations of $f_i, i \leq n$ and $g_i, i \leq n$, respectively (and a constant function). After passing to a subsequence if necessary, we conclude that

$$\frac{d\pi_*}{dR} = \lim_{n\to\infty} \exp(\varphi_n \oplus \psi_n) \quad R\text{-a.s.} \tag{2.8}$$

We would like to pass from the existence of the R-a.s. limit (2.8) to the separate existence of limits $\varphi = \lim \varphi_n$ and $\psi = \lim \psi_n$ in $[-\infty, \infty)$. That is of course not possible at this stage, as there is a degree of freedom in choosing (φ_n, ψ_n): clearly $(\varphi_n - a_n, \psi_n + a_n)$ is another possible choice, for arbitrary $a_n \in \mathbb{R}$. Below, we show that under our assumption $R \sim \mu \otimes \nu$, this is indeed the only degree of freedom and the separate limits exist after a normalization of the form $\varphi_n(x_*) = 0$ for all n. To that end, we first study more closely the structure of the set S where the limit (2.8) exists.

To see where we are headed, note that constructing the separate limits would be straightforward if $\lim \varphi_n \oplus \psi_n$ existed (in $\mathbb{R}$, say) on a product set $A \times B$ of full measure: Choose a normalization $\varphi_n(x_*) = 0$ at an arbitrary $x_* \in A$. Given $(x, y) \in A \times B$, we also have $(x_*, y) \in A \times B$ by the product structure, and writing

$$\varphi_n(x) = \varphi_n(x) - \varphi_n(x_*) = (\varphi_n \oplus \psi_n)(x, y) - (\varphi_n \oplus \psi_n)(x_*, y)$$

shows that $\lim \varphi_n(x)$ exists. In general, a set of full product measure may fail to contain any measurable rectangles $A \times B$ of positive measure. The next lemma provides a slightly weaker property that serves as a proxy for our purposes. (The reader who is less technically inclined may take Corollary 2.12 below for granted and skip directly to Section 2.3.3 without loss of continuity.)

2.3.2 On Sets of Full Product Measure

Lemma 2.10. *Let $(\mathsf{X}, \mathcal{F}_\mathsf{X}, \mu)$ and $(\mathsf{Y}, \mathcal{F}_\mathsf{Y}, \mu)$ be probability spaces. If a set $S \in \mathcal{F}_\mathsf{X} \otimes \mathcal{F}_\mathsf{Y}$ has product measure $(\mu \otimes \nu)(S) = 1$, then $(\mu \otimes \nu)$-almost all $(x_*, y_*) \in S$ have the following property: there are $\mathsf{X}_0 \subset \mathsf{X}$ and $\mathsf{Y}_0 \subset \mathsf{Y}$ with $\mu(\mathsf{X}_0) = \nu(\mathsf{Y}_0) = 1$ such that $S_0 := S \cap (\mathsf{X}_0 \times \mathsf{Y}_0)$ satisfies $(x_*, y_*) \in S_0$ and*

$$(x, y) \in S_0 \quad \Longrightarrow \quad (x_*, y) \in S_0, \;\; (x, y_*) \in S_0.$$

Proof. Let $S_x = \{y : (x,y) \in S\}$ denote the section at $x \in \mathsf{X}$, and analogously for $y \in \mathsf{Y}$. Set $\mathsf{X}_1 = \{x \in \mathsf{X} : \nu(S_x) = 1\}$. In view of Fubini's theorem, $(\mu \otimes \nu)(S) = 1$ implies $\mu(\mathsf{X}_1) = 1$. Let $x_* \in \mathsf{X}_1$ and set $\mathsf{Y}_0 = \{y \in \mathsf{Y} : \mu(S_y) = 1\} \cap S_{x_*}$. Then $\nu(\mathsf{Y}_0) = 1$ following the same argument. Pick any $y_* \in \mathsf{Y}_0$ and note that $\mathsf{X}_0 := \mathsf{X}_1 \cap S_{y_*}$ satisfies $\mu(\mathsf{X}_0) = 1$.

Consider an arbitrary point $(x,y) \in \mathsf{X}_0 \times \mathsf{Y}_0$, then $(x_*, y) \in S_0$ and $(x, y_*) \in S_0$ by the construction of (x_*, y_*). In particular, this applies to any $(x,y) \in S_0$. □

We observe the following consequences for decompositions of functions.

Lemma 2.11. *Let X, Y be sets and $S_0 \subset \mathsf{X} \times \mathsf{Y}$. Suppose there exists a point $(x_*, y_*) \in S_0$ such that $(x_*, y) \in S_0$ and $(x, y_*) \in S_0$ for all $(x,y) \in S_0$. Write $\mathsf{X}_0 := \operatorname{proj}_{\mathsf{X}} S_0$ and $\mathsf{Y}_0 := \operatorname{proj}_{\mathsf{Y}} S_0$, and consider functions $\varphi, \varphi', \varphi_n : \mathsf{X}_0 \to [-\infty, \infty)$ and $\psi, \psi', \psi_n : \mathsf{Y}_0 \to [-\infty, \infty)$ that are finite at x_* and y_*, respectively.*

(i) If $\varphi \oplus \psi = \varphi' \oplus \psi'$ on S_0, then $\varphi = \varphi' + a$ on X_0 and $\psi = \psi' - a$ on Y_0, where $a := \varphi(x_*) - \varphi'(x_*)$.

(ii) Let $F := \lim (\varphi_n \oplus \psi_n) \in [-\infty, \infty)$ exist on S_0 with $F(x_*, y_*) \in \mathbb{R}$, where φ_n are normalized to $\varphi_n(x_*) = 0$. Then the pointwise limits $\varphi := \lim \varphi_n$ and $\psi := \lim \psi_n$ exist in $[-\infty, \infty)$ on X_0 and Y_0, respectively. Indeed, they are given by $\varphi(x) = F(x, y_*) - F(x_*, y_*)$ and $\psi(y) = F(x_*, y)$.

Proof. (i) Without loss of generality, $\varphi(x_*) = \varphi'(x_*)$. Set $F := \varphi \oplus \psi$. Given $(x,y) \in S_0$, we know that $(x_*, y) \in S_0$ and hence

$$\psi(y) = F(x_*, y) - \varphi(x_*) = F(x_*, y) - \varphi'(x_*) = \psi'(y),$$

and then also $\varphi(x) = F(x, y_*) - \psi(y_*) = F(x, y_*) - \psi'(y_*) = \varphi'(x)$.

(ii) Define $\varphi(x) := F(x, y_*) - F(x_*, y_*)$ and $\psi(y) := F(x_*, y)$. Writing $F_n := \varphi_n \oplus \psi_n$, we have $F_n \to F$ on S_0. Therefore,

$$\varphi_n(x) = F_n(x, y_*) - F_n(x_*, y_*) \to F(x, y_*) - F(x_*, y_*) = \varphi(x)$$

and

$$\psi_n(y) = F_n(x_*, y) - \varphi_n(x_*) = F_n(x_*, y) \to F(x_*, y) = \psi(y). \quad \square$$

For ease of reference, we record the combined result of the two lemmas.

Corollary 2.12. *Let $(\mathsf{X},\mathcal{F}_\mathsf{X},\mu)$ and $(\mathsf{Y},\mathcal{F}_\mathsf{Y},\mu)$ be probability spaces. Consider measurable functions $\varphi,\varphi',\varphi_n:\mathsf{X}\to[-\infty,\infty)$ and $\psi,\psi',\psi_n:\mathsf{Y}\to[-\infty,\infty)$.*

(i) If $(\mu\otimes\nu)\{\varphi\oplus\psi>-\infty\}>0$ and $\varphi\oplus\psi=\varphi'\oplus\psi'$ $(\mu\otimes\nu)$-a.s., then $\varphi=\varphi'+a$ μ-a.s. and $\psi=\psi'-a$ ν-a.s. for some $a\in\mathbb{R}$.

(ii) Let $F:=\lim(\varphi_n\oplus\psi_n)\in[-\infty,\infty)$ exist $(\mu\otimes\nu)$-a.s. and suppose that $(\mu\otimes\nu)\{F>-\infty\}>0$. Then $F=\varphi\oplus\psi$ for some measurable functions $\varphi:\mathsf{X}\to[-\infty,\infty)$ and $\psi:\mathsf{Y}\to[-\infty,\infty)$. Moreover, there are $a_n\in\mathbb{R}$ such that $\varphi=\lim(\varphi_n-a_n)$ μ-a.s. and $\psi=\lim(\psi_n+a_n)$ ν-a.s.

Proof. (i) We apply Lemma 2.10 to $S:=\{\varphi\oplus\psi=\varphi'\oplus\psi'\}$, then the assumption $(\mu\otimes\nu)\{\varphi\oplus\psi>-\infty\}>0$ allows us to chose $(x_*,y_*)\in\{\varphi\oplus\psi>-\infty\}$ and the claim follows from Lemma 2.11 (i).

(ii) Consider the measurable set $S:=\{\lim(\varphi_n\oplus\psi_n)$ exists in $[-\infty,\infty)\}$. We apply Lemma 2.10 to S, then $(\mu\otimes\nu)\{F>-\infty\}>0$ allows us to chose $(x_*,y_*)\in\{F>-\infty\}$. In the assertion of Lemma 2.10 we may assume that $\mathsf{X}_0=\mathrm{proj}_\mathsf{X}S_0$ and $\mathsf{Y}_0=\mathrm{proj}_\mathsf{Y}S_0$ and that these sets are measurable (otherwise remove an appropriate nullset). We normalize $\varphi_n(x_*)=0$ for all n (by subtracting a constant from φ_n and adding the same to ψ_n), then Lemma 2.11 (ii) shows that $\varphi=\lim\varphi_n$ and $\psi=\lim\psi_n$ exist a.s. in $[-\infty,\infty)$. They are measurable as limits of measurable functions. □

Remark 2.13. In this framework (and in contrast to Remark 2.15 below), the separate measurability of φ,ψ is generally not an issue. Specifically, let $\varphi:\mathsf{X}\to[-\infty,\infty)$ and $\psi:\mathsf{Y}\to[-\infty,\infty)$. If $F=\varphi\oplus\psi$ $(\mu\otimes\nu)$-a.s. where F is measurable and $(\mu\otimes\nu)\{F>-\infty\}>0$, then φ,ψ are a.s. measurable. Indeed, our proof shows that $\psi(\cdot)=F(x_*,\cdot)-\varphi(x_*)$ μ-a.s. and $\varphi(\cdot)=F(\cdot,y_*)-\psi(y_*)$ ν-a.s. for certain (x_*,y_*), and the right-hand sides are clearly measurable.

2.3.3 Completing the Proof of Existence and Uniqueness

After this excursion, let us return to (2.8) and complete the proof of the existence and uniqueness of the decomposition $\frac{d\pi_*}{dR}=e^{\varphi\oplus\psi}$ R-a.s.

The measurable set $S=\{F:=\lim(\varphi_n\oplus\psi_n)$ exists in $[-\infty,\infty)\}$ satisfies $R(S)=1$ by (2.8) and hence $(\mu\otimes\nu)(S)=1$. The set $S'=\{F>-\infty\}$ satisfies $\pi_*(S')=1$ and hence $R(S')>0$ and then $(\mu\otimes\nu)(S')>0$. Corollary 2.12 thus yields the existence and uniqueness of φ,ψ.

It remains to show that φ,ψ are a.s. finite. The following completes the proof of Theorem 2.1 (a).

Lemma 2.14. *Let $\pi_0 \in \Pi(\mu,\nu)$ admit a density of the form*

$$\frac{d\pi_0}{dR} = e^{\varphi\oplus\psi} \quad R\text{-a.s.}$$

for some measurable functions $\varphi : \mathsf{X} \to [-\infty,\infty)$, $\psi : \mathsf{Y} \to [-\infty,\infty)$. Then φ,ψ are a.s. finite and $\pi_0 \sim R$.

Proof. Let $A := \{\varphi = -\infty\}$, then $\frac{d\pi_0}{dR} = 0$ on $A \times \mathsf{Y}$ and $\pi_0 \in \Pi(\mu,\nu)$ yields

$$\mu(A) = \pi_0(A \times \mathsf{Y}) = \int_{A\times\mathsf{Y}} \frac{d\pi_0}{dR}\, dR = 0.$$

The proof that $\psi > -\infty$ ν-a.s. is analogous. □

Remark 2.15. The assumption $R \sim \mu\otimes\nu$ is important in Theorem 2.1 (a). If merely $R \ll \mu\otimes\nu$, a similar result can be shown, but the identity $\frac{d\pi_*}{dR} = e^{\varphi\oplus\psi}$ only holds π_*-a.s., and $\pi_* \sim R$ may fail. Moreover, the uniqueness of potentials is replaced by a countable number of normalizations (each on a different subset) instead of just one. The subsequent Example 2.16 illustrates these points in a simple case. For general $R \not\ll \mu\otimes\nu$, one can still decompose the density of π_*, but now even the measurability of φ,ψ can fail. Roughly speaking, an uncountable number of normalizations may need to be chosen. We refer to [3, 18, 34] for these more general situations, whereas our proof is closer to arguments going back to [15, 17].

Our assumption that X,Y are separable was made to obtain a constructive approximation as detailed below (2.7)—this is the one and only instance where separability will be used. Assuming separability does not seem to exclude any examples of interest. A less constructive approach based on Hahn–Banach separation, appropriate for general measurable spaces, is taken in [10, Theorem 3.1], which shows that the density of π_* always satisfies

$$\log\frac{d\pi_*}{dR} = \lim_n(\varphi_n \oplus \psi_n) \quad \text{in} \quad L^1(\pi_*) \tag{2.9}$$

for some $\varphi_n \in L^1(\mu)$ and $\psi_n \in L^1(\nu)$.[11] (One can observe that our constructive approximation shares the property (2.9), due to the last assertion of Proposition 1.17.) If $\pi_* \sim R \sim \mu\otimes\nu$, we can proceed exactly as above to deduce the conclusion of Theorem 2.1 (a). If $R \ll \mu\otimes\nu$, one can use the

[11] To avoid confusion, we remark that the corollary stated below Theorem 3.1 in [10] has a glitch. It asserts that a decomposition with integrable potentials is essentially always possible, but the proof overlooks the issue that passing to separate limits in (2.9) is not possible in general. See [18, 33] for more detailed comments.

arguments of [3, 18, 34] to obtain the decomposition at least π_*-a.s. To obtain it also R-a.s., [18] assumes a priori that there exists $\pi \in \Pi_{fin}(\mu,\nu)$ with $\pi \sim R$; this implies $\pi_* \sim R$ by Theorem 1.10. By contrast, we established that $\pi_* \sim R$ necessarily holds when $R \sim \mu \otimes \nu$; cf. Corollary 2.4.

Example 2.16. Consider $\mathsf{X} = \mathsf{Y} = \{0,1\}$ with uniform marginals μ,ν while R is the uniform distribution on $\{(0,0),(0,1),(1,1)\}$. Here $\Pi_{fin}(\mu,\nu)$ has the unique element $\pi_* = (\delta_{(0,0)} + \delta_{(1,1)})/2$, as this is the only coupling absolutely continuous wrt. R. We observe that $\pi_* \not\sim R$. The potentials only need to satisfy

$$\varphi(0) + \psi(0) = \log \tfrac{3}{2} \qquad \text{and} \qquad \varphi(1) + \psi(1) = \log \tfrac{3}{2}, \tag{2.10}$$

so that *two* normalizations are needed to enforce uniqueness. In particular, the uniqueness of potentials up to an additive constant no longer holds. Moreover, the formula $\frac{d\pi_*}{dR} = e^{\varphi \oplus \psi}$ cannot hold at $(1,0)$, as $\pi_*\{(1,0)\} = 0$ would imply $\varphi(1) = -\infty$ or $\psi(0) = -\infty$, contradicting (2.10). That is, the decomposition $\frac{d\pi_*}{dR} = e^{\varphi\oplus\psi}$ holds π_*-a.s., but not R-a.s.

2.3.4 Decomposition Implies Optimality

Next, we prove the "Verification" Theorem 2.1 (b). Suppose that $\pi_0 \in \Pi(\mu,\nu)$ has a density of the form $\frac{d\pi_0}{dR} = \exp(\varphi \oplus \psi)$. If $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$, then $E^\pi[\varphi \oplus \psi] = \mu(\varphi) + \nu(\psi)$ is independent of $\pi \in \Pi(\mu,\nu)$ and Corollary 1.14 directly implies that $\pi_0 = \arg\min_{\pi\in\Pi(\mu,\nu)} H(\,\cdot\,|R)$. (For brevity, we sometimes denote $\mu(\varphi) := \int \varphi \, d\mu$.) In this section, we show by an approximation argument that the conclusion remains valid even without assuming the integrability. In fact, the following result is slightly more precise.

Proposition 2.17. *Let $\Pi_{fin}(\mu,\nu) \neq \emptyset$ and let $\pi_0 \in \Pi(\mu,\nu)$ admit a density*

$$\log \frac{d\pi_0}{dR} = \varphi \oplus \psi \quad R\text{-a.s.}$$

for some measurable functions $\varphi : \mathsf{X} \to [-\infty,\infty)$ and $\psi : \mathsf{Y} \to [-\infty,\infty)$. Then

$$\pi \mapsto E^\pi[\log \tfrac{d\pi_0}{dR}] \quad \text{is constant over} \quad \Pi_{fin}(\mu,\nu) \cup \{\pi_0\} \tag{2.11}$$

and $\pi_0 = \pi_* \in \Pi_{fin}(\mu,\nu)$ is the Schrödinger bridge.

Of course the constant value in (2.11) is $E^{\pi_0}[\log \frac{d\pi_0}{dR}] = H(\pi_0|R)$. While we have not assumed a priori that $H(\pi_0|R) < \infty$, this is part of the conclusion, obtained on the strength of the assumption that $\Pi_{fin}(\mu,\nu) \neq \emptyset$.

Comparing with Theorem 1.10 and the discussion thereafter, Proposition 2.17 states that the density of π_0 satisfies (1.9) with equality, meaning that all directional derivatives vanish at π_0. In particular, this holds for the density of the Schrödinger bridge π_* under the conditions of Theorem 2.1 (a).

The main step for the proof of Proposition 2.17 is the following.

Lemma 2.18. *Let $\varphi : \mathsf{X} \to [-\infty, \infty]$ and $\psi : \mathsf{Y} \to [-\infty, \infty]$ be measurable. Then*

$$\pi \mapsto E^{\pi}[\varphi \oplus \psi] \quad \textit{is constant}$$

over $\{\pi \in \Pi(\mu, \nu) : E^{\pi}[(\varphi \oplus \psi)^+] < \infty$ or $E^{\pi}[(\varphi \oplus \psi)^-] < \infty\}$.

Proof. Consider the bounded functions

$$\varphi_n = (-n) \vee \varphi \wedge n \quad \text{and} \quad \psi_n = (-n) \vee \psi \wedge n. \tag{2.12}$$

Writing $\zeta_n = \varphi_n \oplus \psi_n$ and $\zeta = \varphi \oplus \psi$ and $A := \{\zeta > 0\}$, we have the properties

$$\{\zeta_n > 0\} \subseteq A \subseteq \{\zeta \geq 0\} \subseteq \{\zeta_n \geq 0\}, \tag{2.13}$$

$$0 \leq \zeta_n \uparrow \zeta \quad \text{on} \quad A. \tag{2.14}$$

Let $\pi \in \Pi(\mu, \nu)$. Clearly (2.13) implies that $E^{\pi}[\zeta^+] = E^{\pi}[\zeta \mathbf{1}_A]$ and $E^{\pi}[\zeta_n^+] = E^{\pi}[\zeta_n \mathbf{1}_A]$. Using monotone convergence, (2.14) then implies

$$E^{\pi}[\zeta^+] = E^{\pi}[\zeta \mathbf{1}_A] = \lim E^{\pi}[\zeta_n \mathbf{1}_A] = \lim E^{\pi}[\zeta_n^+].$$

Analogous assertions hold for the negative part with $B := \{\zeta < 0\}$ instead of A. If $E^{\pi}[\zeta^+] < \infty$ or $E^{\pi}[\zeta^-] < \infty$, we can combine the two limits and conclude that

$$E^{\pi}[\zeta] = E^{\pi}[\zeta^+] - E^{\pi}[\zeta^-] = \lim E^{\pi}[\zeta_n^+] - \lim E^{\pi}[\zeta_n^-] = \lim E^{\pi}[\zeta_n].$$

But $E^{\pi}[\zeta_n]$ is constant over $\pi \in \Pi(\mu, \nu)$ as ζ_n is a sum of bounded marginal functions, and the claim follows. □

The following also completes the proof of Theorem 2.1.

Proof of Proposition 2.17. Let $\pi \in \Pi_{fin}(\mu, \nu)$, then $E^{\pi}[(\varphi \oplus \psi)^+] < \infty$ by Lemma 1.4 (a). On the other hand, $E^{\pi_0}[(\varphi \oplus \psi)^-] < \infty$ as $\varphi \oplus \psi$ is the log-density of π_0. Thus Lemma 2.18 implies (2.11) and the last claim follows via Corollary 1.14. □

Remark 2.19. Another way to argue the optimality of π_0 in Proposition 2.17, is to define $\zeta_n = \varphi_n \oplus \psi_n$ as in (2.12) and check the conditions (i)–(iii) of Proposition 1.15. Indeed, (i) is clear. For (ii), we use $|\zeta_n| \leq |\zeta|$ and monotone convergence to see $\lim_n E^{\pi_0}[\zeta_n] = E^{\pi_0}[\zeta]$. And for (iii), $e^{\zeta_n} \leq 1 + e^{\zeta} \in L^1(R)$ implies $\lim_n E^R[e^{\zeta_n}] = E^R[e^{\zeta}] = 1$.

The following corollary shows in particular that there can be at most one coupling with density of the form $\frac{d\pi}{dR} = e^{\varphi \oplus \psi}$ R-a.s., and as a consequence, also at most one cyclically invariant coupling. Notably, this holds even if $\Pi_{fin}(\mu, \nu) = \emptyset$, which was exploited in [1, 23] to give a meaning to the Schrödinger bridge π_* in that situation. When $\Pi_{fin}(\mu, \nu) = \emptyset$, all couplings have infinite relative entropy and thus the Schrödinger bridge problem is not immediately meaningful as an optimization problem.

Corollary 2.20. *Let $\pi, \pi' \in \Pi(\mu, \nu)$ satisfy $\pi' \ll \pi \ll R$ and assume that*

$$\frac{d\pi'}{dR} = \frac{d\pi}{dR} e^{\tilde{\varphi} \oplus \tilde{\psi}} \quad R\text{-a.s.}$$

for measurable functions $\tilde{\varphi} : \mathsf{X} \to [-\infty, \infty)$, $\tilde{\psi} : \mathsf{Y} \to [-\infty, \infty)$. Then $\pi = \pi'$.

Proof. The assumption implies that $\frac{d\pi'}{d\pi} = e^{\tilde{\varphi} \oplus \tilde{\psi}}$ π-a.s. Thus, Proposition 2.17 (with π playing the role of the reference measure) yields that π' minimizes $H(\,\cdot\,|\pi)$ over $\Pi(\mu, \nu)$. But as $\pi \in \Pi(\mu, \nu)$, it is clear that π is the unique minimizer of $H(\,\cdot\,|\pi)$, thus $\pi = \pi'$. □

2.3.5 Integrability of Potentials

Finally, we show the claimed integrability of the potentials. As π_* has finite entropy, it is clear that its log-density $\varphi \oplus \psi$ is π_*-integrable. The subtlety is that in general, this does not imply that φ, ψ are separately integrable for the marginals μ, ν of π_* (cf. Remark 2.22 below). A simple counterexample can be found in [33, Example 1]; see also [18].

Lemma 2.21. *Suppose that (2.2) holds, or equivalently $\mu \otimes \nu \in \Pi_{fin}(\mu, \nu)$. If $\pi_* \in \Pi(\mu, \nu)$ is the Schrödinger bridge and*

$$\log \frac{d\pi_*}{dR} = \varphi \oplus \psi \quad R\text{-a.s.}$$

for some measurable functions $\varphi : \mathsf{X} \to [-\infty, \infty)$ and $\psi : \mathsf{Y} \to [-\infty, \infty)$, then $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$.

Proof. Recall from Corollary 1.13 that

$$\log \frac{d\pi_*}{dR} \in L^1(\pi_0) \quad \text{for all} \quad \pi_0 \in \Pi_{fin}(\mu,\nu).$$

In particular, $\varphi \oplus \psi \in L^1(\mu \otimes \nu)$, which by Fubini implies $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$; cf. Remark 2.22 below. □

Remark 2.22. A note of caution regarding the last step in the preceding proof: as mentioned above, for a coupling $\pi \neq \mu \otimes \nu$, in general, $\varphi \oplus \psi \in L^1(\pi)$ does not imply $\varphi \in L^1(\mu)$ or $\psi \in L^1(\nu)$.

Let $\pi \in \Pi(\mu,\nu)$ have disintegration $\pi = \mu(dx) \otimes K_x(dy)$. If $\varphi \oplus \psi \in L^1(\pi)$, Fubini's theorem for kernels (i.e., tower property of conditional expectation) states that for a.e. x we have $\varphi(x) + \psi(\cdot) \in L^1(K_x)$ and $\int [\varphi(x) + \psi(y)]\, K_x(dy) \in L^1(\mu)$, and moreover that $\iint [\varphi(x) + \psi(y)]\, K_x(dy)\, \mu(dx) = \int \varphi \oplus \psi\, \pi(dx, dy)$.

Clearly $\int [\varphi(x) + \psi(y)]\, K_x(dy) = \varphi(x) + \int \psi(y)\, K_x(dy) =: \varphi(x) + \Psi(x)$. On the other hand, the fact that this sum is in $L^1(\mu)$ does not imply that φ (or Ψ) is μ-integrable on its own.

The situation is different for the particular coupling $\pi = \mu \otimes \nu$ used in the proof of Lemma 2.21. As the kernel $K_x \equiv \nu$ does not depend on x, the above function $\Psi(x) := \int \psi(y)\, \nu(dy)$ cannot depend on x. The constant $a = \Psi(x)$ must be finite, because $\varphi(x) + \Psi(x)$ must be finite μ-a.s. Now, the fact that $\varphi + a \in L^1(\mu)$ indeed tells us that $\varphi \in L^1(\mu)$.

The above proof of Lemma 2.21 through Corollary 1.13 is short yet somewhat indirect. Next, we offer an alternate argument.

Lemma 2.23. *Let $\pi \in \Pi(\mu,\nu)$ and $\log \frac{d\pi}{dR} = \varphi \oplus \psi$ R-a.s. for some measurable functions $\varphi : \mathsf{X} \to [-\infty, \infty)$ and $\psi : \mathsf{Y} \to [-\infty, \infty)$. If*

$$(\varphi \oplus \psi)^+ \in L^1(\mu \otimes \nu),$$

then $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. Moreover, $\mu(\varphi) + \nu(\psi) = H(\pi|R)$.

Proof. We have $E^\pi[(\varphi \oplus \psi)^-] < \infty$ as $\varphi \oplus \psi$ is the log-density of π; in fact, $E^\pi[\varphi \oplus \psi] = H(\pi|R) \geq 0$. In view of our assumption, Lemma 2.18 then guarantees that $E^{\mu\otimes\nu}[\varphi \oplus \psi] = E^\pi[\varphi \oplus \psi] = H(\pi|R) \geq 0$. In particular, we also have $(\varphi \oplus \psi)^- \in L^1(\mu \otimes \nu)$. Thus $\varphi \oplus \psi \in L^1(\mu \otimes \nu)$ which by Fubini implies $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$; cf. Remark 2.22. □

Another Proof of Lemma 2.21. Let $\mu \otimes \nu \in \Pi_{fin}(\mu,\nu)$ and let $\pi \in \Pi(\mu,\nu)$ satisfy $\log \frac{d\pi}{dR} = \varphi \oplus \psi$ R-a.s. Then $E^{\mu\otimes\nu}[(\varphi \oplus \psi)^+] < \infty$ by Lemma 1.4 (a), so that Lemma 2.23 yields $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. □

3 Duality for Static Schrödinger Bridges

In this section, we characterize the Schrödinger potentials φ_*, ψ_* as the solution to a "dual" optimization problem. Let (X, μ) and (Y, ν) be probability spaces. We fix a reference measure $R \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ and assume that there exists $\pi_* \in \Pi(\mu, \nu)$ with a density of the form

$$\frac{d\pi_*}{dR} = e^{\varphi_* \oplus \psi_*} \quad R\text{-a.s.} \quad \text{where} \quad \varphi_* \in L^1(\mu), \quad \psi_* \in L^1(\nu). \tag{3.1}$$

By Corollary 2.3, this implies that $H(\pi_*|R) = \mu(\varphi_*) + \nu(\psi_*) < \infty$ and that $\pi_* = \arg\min_{\Pi(\mu,\nu)} H(\,\cdot\,|R)$ is the unique Schrödinger bridge. We have seen in Theorem 2.1 (a) how (3.1) is necessarily satisfied when $R \sim \mu \otimes \nu$ and $\mu \otimes \nu \in \Pi_{fin}(\mu, \nu)$, but those conditions will not be needed directly.

The following fact, sometimes called weak duality, is the first half of the duality relation.

Lemma 3.1. *Let $\pi \in \Pi(\mu, \nu)$ and $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$. Then*

$$H(\pi|R) \geq \mu(\varphi) + \nu(\psi) - \int e^{\varphi \oplus \psi} \, dR + 1.$$

Proof. Let $\pi \in \Pi(\mu, \nu)$ and $\Phi \in L^1(\pi)$. As in (1.13), Fenchel's inequality yields $\alpha \log \alpha - \alpha \geq \beta\alpha - e^\beta$, or equivalently

$$\alpha \log \alpha - \beta\alpha \geq \alpha - e^\beta \quad \text{for all} \quad \alpha \geq 0, \quad \beta \in \mathbb{R}. \tag{3.2}$$

Write $Z = d\pi/dR$. Using (3.2) with $\alpha = Z(x, y)$ and $\beta = \Phi(x, y)$ yields the inequality in

$$\begin{aligned} H(\pi|R) = \int Z \log Z \, dR &= \int \Phi Z \, dR + \int (Z \log Z - \Phi Z) \, dR \\ &\geq \int \Phi Z \, dR + \int (Z - e^\Phi) \, dR = \int \Phi \, d\pi - \int e^\Phi \, dR + 1. \end{aligned}$$

If $\Phi = \varphi \oplus \psi$ for $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$, then $\int \Phi \, d\pi = \mu(\varphi) + \nu(\psi)$ as π is a coupling. The claim follows. □

The next result shows that the Schrödinger potentials φ_*, ψ_* are the maximizers for the concave *dual problem*

$$\sup_{\varphi \in L^1(\mu),\, \psi \in L^1(\nu)} G(\varphi, \psi), \quad G(\varphi, \psi) := \mu(\varphi) + \nu(\psi) - \int e^{\varphi \oplus \psi} \, dR + 1 \tag{3.3}$$

and that there is no "duality gap" between the primal (entropy minimization) problem and the dual problem.

Theorem 3.2 (Duality)**.** *Let* (3.1) *hold.*[12] *We have*

$$\inf_{\pi\in\Pi(\mu,\nu)} H(\pi|R) = \sup_{\varphi\in L^1(\mu),\, \psi\in L^1(\nu)} \mu(\varphi)+\nu(\psi) - \int e^{\varphi\oplus\psi}\, dR + 1, \qquad (3.4)$$

the supremum is attained by the Schrödinger potentials (φ_*, ψ_*)*, and*

$$H(\pi_*|R) = \inf_{\pi\in\Pi(\mu,\nu)} H(\pi|R) = \mu(\varphi_*)+\nu(\psi_*).$$

The maximizers are unique in the sense that if (φ,ψ) *are other maximizers, then* $\varphi\oplus\psi = \varphi_*\oplus\psi_*$ *R-a.s.*[13]

Proof. The inequality "$\geq$" in (3.4) follows from Lemma 3.1. On the other hand, (3.1) yields

$$H(\pi_*|R) = \int (\varphi_*\oplus\psi_*)\, d\pi_* = \mu(\varphi_*)+\nu(\psi_*)$$

and

$$\int e^{\varphi_*\oplus\psi_*}\, dR = \int 1\, d\pi_* = 1,$$

so that equality is attained in (3.4) for π_*, φ_*, ψ_*. This shows (3.4), that (φ_*, ψ_*) are maximizers, and also that $H(\pi_*|R) = \mu(\varphi_*)+\nu(\psi_*)$. Uniqueness follows from the strict concavity of the dual problem. □

Remark 3.3. An alternate way to write a dual problem is

$$\sup_{\varphi\in L^1(\mu),\, \psi\in L^1(\nu)} \mu(\varphi)+\nu(\psi) - \log\int e^{\varphi\oplus\psi}\, dR. \qquad (3.5)$$

Lemma 3.1 and Theorem 3.2 apply to this dual problem without changes. One way to obtain the weak duality for (3.5) is to recall (1.1) which, with $\phi := \varphi\oplus\psi$, already yields $H(\pi|R) \geq \mu(\varphi)+\nu(\psi) - \log\int e^{\varphi\oplus\psi}\, dR$ for all bounded φ, ψ. This extends to integrable φ,ψ by a dominated convergence argument, and now the analogue of Theorem 3.2 follows as before.

Remark 3.4 (Euler–Lagrange)**.** The Schrödinger equations (SE1)–(SE2) in Section 2.1 can be interpreted as the Euler–Lagrange equations of the concave maximization problem (3.3); i.e., as the variational first-order condition for optimality. To see this, fix $\varphi\in L^1(\mu)$, $\psi\in L^1(\nu)$ and let μ' be the first

[12] The integrability condition can be weakened to $(\varphi_*\oplus\psi_*)^+ \in L^1(\mu\otimes\nu)$ by Lemma 2.23.

[13] If $R\sim\mu\otimes\nu$, this implies that φ_*, ψ_* are a.s. unique up to an additive constant; cf. Corollary 2.12.

marginal of the measure $d\pi(\varphi,\psi) := e^{\varphi\oplus\psi}\,dR$. Consider a bounded measurable function $\varphi_0 : \mathsf{X} \to \mathbb{R}$ and $\varepsilon \in \mathbb{R}$, then

$$\begin{aligned}G(\varphi,\psi) - G(\varphi+\varepsilon\varphi_0,\psi) &= \int (e^{\varepsilon\varphi_0} - 1)\,d\pi(\varphi,\psi) - \varepsilon\mu(\varphi_0)\\ &= \int (e^{\varepsilon\varphi_0} - 1)\,d\mu' - \varepsilon\mu(\varphi_0)\\ &= \varepsilon[\mu'(\varphi_0) - \mu(\varphi_0)] + o(\varepsilon).\end{aligned}$$

If $\varphi = \arg\max G(\cdot,\psi)$, we must have $G(\varphi,\psi) - G(\varphi+\varepsilon\varphi_0,\psi) \geq 0$, hence $\mu'(\varphi_0) = \mu(\varphi_0)$ for all bounded φ_0. It follows that $\mu' = \mu$, or equivalently that φ solves (SE1). Similarly, (SE2) is the first-order condition for $\psi = \arg\max G(\varphi,\cdot)$.

4 Entropic Optimal Transport

Let (X,μ) and (Y,ν) be probability spaces and $\Pi(\mu,\nu)$ the set of couplings on the product $\mathsf{X}\times\mathsf{Y}$. We also fix a probability measure $P \in \mathcal{P}(\mathsf{X}\times\mathsf{Y})$; it can be arbitrary for now but will soon be chosen as the **p**roduct $P = \mu\otimes\nu$. Given a measurable "cost" function $c : \mathsf{X}\times\mathsf{Y} \to (-\infty,\infty]$ with $P\{c < \infty\} > 0$, the entropic optimal transport (or EOT) problem with regularization parameter $\varepsilon > 0$ is

$$\mathcal{C}_\varepsilon = \mathcal{C}_\varepsilon(\mu,\nu,c) = \inf_{\pi\in\Pi(\mu,\nu)} \int c\,d\pi + \varepsilon H(\pi|P). \qquad (\varepsilon\mathsf{EOT})$$

For simplicity, we assume for the moment that c is uniformly bounded from below, so that the terms on the right-hand side are always well-defined in $(-\infty,\infty]$. Dividing (εEOT) by ε transforms it into a similar problem with $\varepsilon = 1$ and cost function c/ε,

$$\mathcal{C}_\varepsilon(\mu,\nu,c) = \varepsilon\mathcal{C}_1(\mu,\nu,c/\varepsilon),$$

and moreover these problems have the same minimizers.[14] For our basic considerations, it thus suffices to consider $\varepsilon = 1$, simplifying the notation:

$$\mathcal{C}_1 = \mathcal{C}_1(\mu,\nu,c) = \inf_{\pi\in\Pi(\mu,\nu)} \int c\,d\pi + H(\pi|P). \qquad (\mathsf{EOT})$$

As c is bounded from below, e^{-c} is bounded and in particular

$$e^{-c} \in L^1(P), \qquad (4.1)$$

[14] However, in contrast to the unregularized transport problem, there is no simple relation between $\mathcal{C}_1(\mu,\nu,c/\varepsilon)$ and $\mathcal{C}_1(\mu,\nu,c)$ or between the corresponding minimizers.

so we can introduce the auxiliary **r**eference measure $R \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ via

$$dR := \alpha^{-1} e^{-c}\, dP = e^{-(c+\log\alpha)}\, dP, \qquad \alpha := \int e^{-c}\, dP > 0. \tag{4.2}$$

We note that $R \ll P$ and moreover $R \sim P$ if and only if $c < \infty$ P-a.s.

For $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ with $\pi \ll R$, we have

$$\begin{aligned} H(\pi|R) &= \int \log\left(\frac{d\pi}{dR}\right) d\pi = \int \log\left(\frac{d\pi}{dP}\frac{dP}{dR}\right) d\pi \\ &= H(\pi|P) + \int \log\left(\alpha e^{c}\right) d\pi = \int c\, d\pi + H(\pi|P) + \log\alpha. \end{aligned}$$

This identity extends to arbitrary $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$; in brief,

$$F(\pi) := \int c\, d\pi + H(\pi|P) = H(\pi|R) - \log\alpha, \tag{4.3}$$

showing that the entropic transport problem (EOT) is equivalent to the static Schrödinger bridge problem for R:

$$\mathcal{C}_1 = \inf_{\pi \in \Pi(\mu,\nu)} H(\pi|R) - \log\alpha \tag{4.4}$$

and both problems have the same minimizers. If $\int e^{-c}\, dP = 1$, the problems are identical.

Using the notation $\Pi_{fin}(\mu,\nu)$ for couplings with finite entropy $H(\,\cdot\,|R)$ as in Section 2, we also see that

$$\pi \in \Pi_{fin}(\mu,\nu) \quad \iff \quad F(\pi) < \infty \quad \iff \quad c \in L^1(\pi) \text{ and } H(\pi|P) < \infty. \tag{4.5}$$

We have reduced (EOT) to the Schrödinger bridge problem. Conversely, starting with any probability measure $R \ll P$, we can introduce a measurable function $c : \mathsf{X} \times \mathsf{Y} \to (-\infty, \infty]$ with $P\{c < \infty\} > 0$ via

$$c(x,y) := -\log\frac{dR}{dP}.$$

Then $\alpha = \int e^{-c}\, dP = 1$ and we see that the Schrödinger bridge problem is also a particular case of (EOT), as least when c is bounded from below. To remove the latter restriction, we extend (EOT) as follows.[15]

[15] In most cases of interest, the cost function c is bounded from below and finite-valued, like the quadratic cost $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$. We consider more general cases mainly have the full equivalence with the Schrödinger bridge formulation.

Remark 4.1. We extend (EOT) to costs c which are not necessarily bounded from below, but instead merely satisfy (4.1). There is a slight delicacy regarding the right-hand side of (EOT), as it might read $-\infty + \infty$. We define

$$F(\pi) := \int c \, d\pi + \varepsilon H(\pi|P) := \int \left(c + \log \frac{d\pi}{dP} \right) d\pi.$$

The latter integral is always well-defined in $(-\infty, \infty]$; indeed, it is equal to $H(\pi|R) - \log \alpha$. Thus we avoid the expression $-\infty + \infty$ and more importantly, (4.3) remains valid for all $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$. More generally, all the considerations above (and below) remain valid for this extension, except possibly the last part of (4.5)

While we have considered a general P so far, let us now specialize to $P = \mu \otimes \nu$. Then (4.5) for the product coupling $\pi = \mu \otimes \nu$ becomes

$$\mu \otimes \nu \in \Pi_{fin}(\mu, \nu) \iff c \in L^1(\mu \otimes \nu). \tag{4.6}$$

In particular, the condition $\mu \otimes \nu \in \Pi_{fin}(\mu, \nu)$, or equivalently (2.2), boils down to a standard integrability condition on the cost function.

The next result translates Theorem 2.1 to the present setting; for part (a) our proof thus assumes that (X, μ) and (Y, ν) are separable.

Theorem 4.2 (Structure Theorem for EOT)**.** *Let $\mathcal{C}_1 < \infty$. Then there is a unique minimizer π_* for the entropic optimal transport problem* (EOT)*.*

(a) Let $c < \infty$ $\mu \otimes \nu$-a.s. Then $\pi_* \sim \mu \otimes \nu$ and there are measurable functions $\varphi : \mathsf{X} \to \mathbb{R}$, $\psi : \mathsf{Y} \to \mathbb{R}$, called EOT potentials, *such that*

$$\frac{d\pi_*}{d(\mu \otimes \nu)} = e^{\varphi \oplus \psi - c} \quad \mu \otimes \nu\text{-a.s.}$$

The EOT potentials are a.s. unique up to an additive constant.[16] *If $c \in L^1(\mu \otimes \nu)$, then $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$.*

(b) Conversely, let $\pi_0 \in \Pi(\mu, \nu)$ admit a density of the form

$$\frac{d\pi_0}{d(\mu \otimes \nu)} = e^{\varphi \oplus \psi - c} \quad \mu \otimes \nu\text{-a.s.}$$

for some functions $\varphi : \mathsf{X} \to [-\infty, \infty)$, $\psi : \mathsf{Y} \to [-\infty, \infty)$. Then π_0 is the minimizer π_* and φ, ψ are the EOT potentials.

[16] I.e., if φ', ψ' are potentials, then $\varphi' = \varphi + a$ μ-a.s. and $\psi' = \psi - a$ ν-a.s. for some $a \in \mathbb{R}$.

Proof. Taking into account (4.2), (4.3) and (4.6), this is a direct application of Theorem 2.1 and Corollary 2.4. In (b), we did not require explicitly that φ, ψ are measurable. Indeed, $\varphi \oplus \psi = c + \log \frac{d\pi_0}{d(\mu\otimes\nu)}$ is necessarily measurable, and hence φ and ψ are (a.s.) measurable due to Remark 2.13. □

As a consequence of Theorem 4.2, π_* can be also be characterized by a cyclical invariance property as in Section 2.2. We do not detail this here, but instead state the definition in Section 5.3 where it is also applied.

Remark 4.3. The EOT potentials in Theorem 4.2 differ from the Schrödinger potentials in Theorem 2.1 by a constant. In the literature, both are often called Schrödinger potentials, a slight inconsistency. Indeed, if φ, ψ are the Schrödinger potentials as in Theorem 2.1,

$$\frac{d\pi_*}{d(\mu \otimes \nu)} = \frac{d\pi_*}{dR}\frac{dR}{d(\mu \otimes \nu)} = e^{(\varphi \oplus \psi - \log \alpha) - c},$$

so that the sum of EOT potentials in the sense of Theorem 4.2 is shifted by $\log \alpha$. On the other hand, this is convenient as it corresponds exactly to the shift in (4.4). For instance, the duality in Theorem 4.7 below takes the same form as Theorem 3.2.

In contrast to a scaling of the cost as discussed below ($\varepsilon\mathsf{EOT}$), the next two remarks treat transformations that are separable and hence do not affect the optimizers. A simple transformation is to shift the cost c by a function of x or y alone; in particular, this allows us to relax the integrability condition (4.1).

Remark 4.4 (Shift of Cost)**.** Let $c_1 \in L^1(\mu)$ and $c_2 \in L^1(\nu)$. Then

$$\mathcal{C}_1(\mu, \nu, c + c_1 \oplus c_2) = \mathcal{C}_1(\mu, \nu, c) + \mu(c_1) + \nu(c_2)$$

and, if finite, both problems have *the same minimizer* $\pi_* \in \Pi(\mu, \nu)$.

Sometimes the entropic optimal transport problem is considered with entropy relative to a measure different from the product of the marginals. As long as that measure is also product, it still suffices to study (EOT).

Remark 4.5 (Different Reference)**.** Let $\mu' \in \mathcal{P}(\mathsf{X}), \nu' \in \mathcal{P}(\mathsf{Y})$ and consider

$$\mathcal{C}_1' := \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi + H(\pi | \mu' \otimes \nu'). \tag{4.7}$$

Then

$$\mathcal{C}_1' = \mathcal{C}_1 + H(\mu|\mu') + H(\nu|\nu')$$

and if $\mathcal{C}_1'$ is finite, both problems have *the same minimizer* $\pi_* \in \Pi(\mu,\nu)$. More generally, we have

$$H(\pi|\mu' \otimes \nu') = H(\pi|\mu \otimes \nu) + H(\mu|\mu') + H(\nu|\nu') \quad \text{for all} \quad \pi \in \Pi(\mu,\nu).$$

Indeed, for $\pi \in \Pi(\mu,\nu)$ with $\pi \ll \mu' \otimes \nu'$, we necessarily have $\mu \ll \mu'$ and $\nu \ll \nu'$. This implies $\mu \otimes \nu \ll \mu' \otimes \nu'$ and then

$$\begin{aligned} H(\pi|\mu' \otimes \nu') &= \int \log\left(\frac{d\pi}{d(\mu\otimes\nu)}\frac{d(\mu\otimes\nu)}{d(\mu'\otimes\nu')}\right) d\pi \\ &= H(\pi|\mu\otimes\nu) + \int \left[\log\left(\frac{d\mu}{d\mu'}(x)\right) + \log\left(\frac{d\nu}{d\nu'}(y)\right)\right] \pi(dx,dy) \\ &= H(\pi|\mu\otimes\nu) + H(\mu|\mu') + H(\nu|\nu'). \end{aligned}$$

Whereas if $\pi \not\ll \mu' \otimes \nu'$, then either $\pi \not\ll \mu \otimes \nu$ or $\mu \not\ll \mu'$ or $\nu \not\ll \nu'$, so that both sides are infinite.

We can note that Remark 4.4 and Remark 4.5 are two sides of the same medal: if $H(\mu|\mu') < \infty$ and $H(\nu|\nu') < \infty$, then $c_1 := \log \frac{d\mu}{d\mu'} \in L^1(\mu)$ and $c_2 := \log \frac{d\nu}{d\nu'} \in L^1(\nu)$. For $\pi \in \Pi(\mu,\nu)$ we thus have $\int c_1 \oplus c_2 \, d\pi = H(\mu|\mu') + H(\nu|\nu')$, so that (4.7) can be seen as the problem (EOT) with our usual entropy relative to $\mu \otimes \nu$ but changed cost $\tilde{c} := c + c_1 \oplus c_2$.

In the context of entropic optimal transport, the general inequality of Corollary 1.12 is strengthened to an equality: the suboptimality of a coupling π (in terms of cost) is exactly given by its entropy relative to π_*. If we think of the entropy as a notion of distance, this reflects the strict convexity of entropic optimal transport in a quantitative way.

Proposition 4.6. *Let $\mathcal{C}_1 < \infty$ and $c < \infty$ $\mu \otimes \nu$-a.s. Denote $F(\pi) := \int c \, d\pi + H(\pi|\mu \otimes \nu)$ as in* (4.3). *The minimizer π_* of* (EOT) *satisfies*

$$F(\pi) - F(\pi_*) = H(\pi|\pi_*) \quad \textit{for all} \quad \pi \in \Pi(\mu,\nu) \textit{ with } F(\pi) < \infty.$$

Proof. In view of Theorem 4.2 and Proposition 2.17, Corollary 1.12 yields $H(\pi|\pi_*) = H(\pi|R) - H(\pi_*|R)$ and we conclude via (4.3). □

Finally, we translate our duality results to the current setting. For simplicity, we use the slightly less general condition $c \in L^1(\mu \otimes \nu)$.

Theorem 4.7 (EOT Duality)**.** *Let $c \in L^1(\mu \otimes \nu)$. Then*

$$\mathcal{C}_1 = \sup_{\varphi \in L^1(\mu),\, \psi \in L^1(\nu)} \mu(\varphi) + \nu(\psi) - \int e^{\varphi \oplus \psi - c} \, d(\mu \otimes \nu) + 1, \tag{4.8}$$

the supremum is attained by the EOT potentials $\varphi_* \in L^1(\mu), \psi_* \in L^1(\nu)$, and

$$\mathcal{C}_1 = \mu(\varphi_*) + \nu(\psi_*). \tag{4.9}$$

The maximizers are a.s. unique up to an additive constant.

Proof. In view of Theorem 4.2 and $\mu \otimes \nu \sim R$, $c \in L^1(\mu \otimes \nu)$ implies the assumption of Theorem 3.2. The claims then follow from Theorem 3.2 and Theorem 2.1. □

Remark 4.8. (a) In Theorem 4.7, the assumption $c \in L^1(\mu \otimes \nu)$ guarantees that the minimizer $\pi_* \in \Pi(\mu, \nu)$ exists and has a density of the form

$$\frac{d\pi_*}{d(\mu \otimes \nu)} = e^{\varphi_* \oplus \psi_* - c} \quad \mu \otimes \nu\text{-a.s.} \quad \text{with} \quad \varphi_* \in L^1(\mu),\ \psi_* \in L^1(\nu). \tag{4.10}$$

As in Theorem 3.2, we can omit the assumption $c \in L^1(\mu \otimes \nu)$ and directly assume (4.10) instead; moreover, we can weaken the integrability requirement to $(\varphi_* \oplus \psi_*)^+ \in L^1(\mu \otimes \nu)$.

(b) As in Remark 3.3, Theorem 4.7 also holds for the dual problem

$$\sup_{\varphi \in L^1(\mu),\, \psi \in L^1(\nu)} \mu(\varphi) + \nu(\psi) - \log \int e^{\varphi \oplus \psi - c}\, d(\mu \otimes \nu).$$

4.1 Regularity of Potentials

Next, we exemplify how the Schrödinger equations can be used to establish regularity properties of the potentials. As in Theorem 4.2 (a), let φ, ψ be the EOT potentials such that

$$\frac{d\pi_*}{d(\mu \otimes \nu)} = e^{\varphi \oplus \psi - c} \quad \mu \otimes \nu\text{-a.s.}$$

As in Section 2.1, the fact that π_* is a probability measure with marginals μ and ν implies the Schrödinger equations

$$\begin{aligned} \varphi(x) &= -\log \int e^{\psi(y) - c(x,y)}\, \nu(dy) \quad \mu\text{-a.s.}, \\ \psi(y) &= -\log \int e^{\varphi(x) - c(x,y)}\, \mu(dx) \quad \nu\text{-a.s.} \end{aligned} \tag{4.11}$$

A priori, φ, ψ are only defined μ-a.s. and ν-a.s., respectively. To discuss pointwise properties such as continuity, we would like to define the potentials at every point, not only almost-surely. Indeed, we can choose natural versions of φ, ψ defined everywhere on X, Y by using the right-hand sides as of (4.11) as pointwise definitions for the left-hand sides. With that choice, (4.11) holds everywhere on $\mathsf{X} \times \mathsf{Y}$, even without the a.s. qualifiers.

As the Schrödinger equations express one (exponentiated) potential as a convolution of the other potential with the kernel $e^{-c(x,y)}$, they can be used to study how the potentials inherit regularity properties from c (e.g., [5, 13, 27, 28]). Below, we give two basic examples of such a priori estimates; one could similarly estimate derivatives, etc. We mention that (4.11) can be seen as an analogue of the c-conjugacy relation between the Kantorovich potentials in unregularized optimal transport (e.g., [35]) which has been used used extensively to derive regularity properties in that context.

Lemma 4.9. *Let $c \in L^1(\mu \otimes \nu)$ and let φ, ψ be versions of the EOT potentials satisfying* (4.11) *everywhere as well as $\mu(\varphi) \geq 0$ and $\nu(\psi) \geq 0$. Then*

$$\inf_{y \in \mathsf{Y}} \big[c(x,y) - \psi(y)\big] \leq \varphi(x) \leq \int c(x,y)\, \nu(dy),$$
$$\inf_{x \in \mathsf{X}} \big[c(x,y) - \varphi(x)\big] \leq \psi(y) \leq \int c(x,y)\, \mu(dx)$$

for all $x \in \mathsf{X}$ and $y \in \mathsf{Y}$. If c is bounded, then

$$a \leq \varphi \leq b, \qquad a \leq \psi \leq b$$

for $a = -\|c^-\|_\infty - \|c^+\|_\infty$ and $b = \|c^+\|_\infty$. Whereas if c is merely bounded from below and $e^c \in L^1(\mu \otimes \nu)$, then

$$\tilde{a} \leq \varphi, \qquad \tilde{a} \leq \psi$$

for $\tilde{a} = -\|c^-\|_\infty - \log \|e^c\|_{L^1(\mu \otimes \nu)}$.

Proof. Using (4.11), Jensen's inequality and $\nu(\psi) \geq 0$,

$$\begin{aligned}\varphi(x) &= -\log \int e^{\psi(y) - c(x,y)}\, \nu(dy) \\ &\leq \int [-\psi(y) + c(x,y)]\, \nu(dy) \leq \int c(x,y)\, \nu(dy).\end{aligned}$$

For the lower bound, (4.11) yields

$$\begin{aligned}\varphi(x) &\geq -\log \int e^{\sup_{y \in \mathsf{Y}} [\psi(y) - c(x,y)]}\, \nu(dy) \\ &= -\sup_{y \in \mathsf{Y}} \big[\psi(y) - c(x,y)\big] = \inf_{y \in \mathsf{Y}} \big[c(x,y) - \psi(y)\big].\end{aligned}$$

The proof of the first claim for ψ is symmetric.

If c is bounded, the above upper bounds imply $\varphi, \psi \leq \|c^+\|_\infty$. Using this in the lower bound then also yields the lower bound a. If c is merely bounded from below, we have

$$e^{-\varphi(x)} = \int e^{\psi(y)-c(x,y)}\,\nu(dy) \leq e^{\|c^-\|_\infty} \int e^{\psi(y)}\,\nu(dy),$$

and for the latter integral, the upper bound and Jensen's inequality imply

$$\int e^{\psi(y)}\,\nu(dy) \leq \int e^{\int c(x,y)\,\mu(dx)} \nu(dy) \leq \int e^c d(\mu \otimes \nu).$$

The claimed lower bound $\tilde{a}$ follows by taking logarithm. $\square$

Remark 4.10. In Lemma 4.9, the condition that $\mu(\varphi) \geq 0$ and $\nu(\psi) \geq 0$ depends on the chosen normalization. If $\int e^{-c} d(\mu \otimes \nu) \leq 1$, then (4.9) implies that the condition holds for two popular choices, the normalization $\mu(\varphi) = 0$ and the symmetric normalization $\mu(\varphi) = \nu(\psi)$.

Let $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ be a modulus of continuity; i.e., continuous at 0 with $\omega(0) = 0$. In the following lemma, we assume that some metrics $d_{\mathsf{X}}, d_{\mathsf{Y}}$ are given on X, Y.

Lemma 4.11. *Let φ, ψ be versions of the EOT potentials satisfying* (4.11). *If c is uniformly continuous with modulus ω in both variables, then φ, ψ are uniformly continuous with the same modulus ω.*

Proof. Let $x_1, x_2 \in \mathsf{X}$ satisfy $\varphi(x_1) \geq \varphi(x_2)$. Then

$$\begin{aligned}
&|\varphi(x_1) - \varphi(x_2)| \\
&\quad = \log \int e^{\psi(y)-c(x_2,y)}\,\nu(dy) - \log \int e^{\psi(y)-c(x_1,y)}\,\nu(dy) \\
&\quad = \log \int e^{c(x_1,y)-c(x_2,y)+\psi(y)-c(x_1,y)}\,\nu(dy) - \log \int e^{\psi(y)-c(x_1,y)}\,\nu(dy) \\
&\quad \leq \log \left[e^{\sup_{y \in \mathsf{Y}} |c(x_1,y)-c(x_2,y)|} \int e^{\psi(y)-c(x_1,y)}\,\nu(dy) \right] - \log \int e^{\psi(y)-c(x_1,y)}\,\nu(dy) \\
&\quad = \sup_{y \in \mathsf{Y}} |c(x_1,y) - c(x_2,y)| \leq \omega(d_{\mathsf{X}}(x_1,x_2)).
\end{aligned}$$

The case $\varphi(x_1) \leq \varphi(x_2)$ follows by symmetry and the proof for ψ is analogous. $\square$

Lemma 4.11 shows in particular that the potentials can inherit a Lipschitz constant from c. On the other hand, uniform continuity is a strong assumption if the spaces X, Y are unbounded. We remark that the proof of Lemma 4.11 can be modified to alleviate this, for instance if a decay condition on the tails of μ, ν is given. As an example, $c(x,y) = -x \cdot y$ on $\mathbb{R}^d \times \mathbb{R}^d$ is not uniformly continuous (this is the quadratic cost $\|x-y\|^2/2$ after applying Remark 4.4), but assuming the that marginals are subgaussian, the conjugacy relations still imply regularity of the potentials; cf. [27].

For ease of reference, we conclude this section by explicitly stating some formulas for the problem (εEOT) with regularization parameter $\varepsilon \neq 1$.

Remark 4.12 (Rescaled EOT Potentials)**.** Let $\mathcal{C}_\varepsilon < \infty$, then Theorem 4.2 applied to $\tilde{c} := c/\varepsilon$ immediately yields that the solution π_ε of (εEOT) satisfies $\frac{d\pi_\varepsilon}{d(\mu\otimes\nu)} = e^{\varphi\oplus\psi - c/\varepsilon}$ $\mu \otimes \nu$-a.s. It is sometimes convenient to consider the *rescaled EOT potentials*[17] $\varphi_\varepsilon := \varepsilon\varphi$ and $\psi_\varepsilon := \varepsilon\psi$, for which the optimal density takes the form

$$\frac{d\pi_\varepsilon}{d(\mu\otimes\nu)} = e^{\frac{\varphi_\varepsilon\oplus\psi_\varepsilon - c}{\varepsilon}} \quad \mu\otimes\nu\text{-a.s.} \tag{4.12}$$

There is no benefit regarding the Schrödinger equations, which now read

$$\begin{aligned} \varphi_\varepsilon(x) &= -\varepsilon \log \int e^{\frac{\psi_\varepsilon(y) - c(x,y)}{\varepsilon}} \, \nu(dy) \quad \mu\text{-a.s.}, \\ \psi_\varepsilon(y) &= -\varepsilon \log \int e^{\frac{\varphi_\varepsilon(x) - c(x,y)}{\varepsilon}} \, \mu(dx) \quad \nu\text{-a.s.} \end{aligned} \tag{4.13}$$

On the other hand, the regularity results take the same form as for $\varepsilon = 1$: after replacing (φ, ψ) by $(\varphi_\varepsilon, \psi_\varepsilon)$, Lemma 4.9 and Lemma 4.11 hold verbatim. This may be a first hint that $(\varphi_\varepsilon, \psi_\varepsilon)$ are at a natural scale. If $c \in L^1(\mu\otimes\nu)$, the duality can be stated as

$$\mathcal{C}_\varepsilon = \sup_{\varphi\in L^1(\mu),\, \psi\in L^1(\nu)} \mu(\varphi) + \nu(\psi) - \varepsilon \int e^{\frac{\varphi\oplus\psi - c}{\varepsilon}} \, d(\mu\otimes\nu) + \varepsilon \tag{4.14}$$

with the supremum now attained by the rescaled EOT potentials $(\varphi_\varepsilon, \psi_\varepsilon)$. Thus

$$\mathcal{C}_\varepsilon = \mu(\varphi_\varepsilon) + \nu(\psi_\varepsilon), \tag{4.15}$$

again taking the same form as for $\varepsilon = 1$.

Another motivation to use the rescaled potentials will be detailed in Section 5.4: in the limit $\varepsilon \to 0$, the potentials without rescaling would

[17] In the literature, all these functions are generally called Schrödinger potentials or merely potentials. We are adding some terminology to differentiate them more clearly.

blow up, whereas $(\varphi_\varepsilon, \psi_\varepsilon)$ converge to their analogues in the unregularized transport problem (the Kantorovich potentials). This again suggests that $(\varphi_\varepsilon, \psi_\varepsilon)$ are at a natural scale.

5 Convergence to Optimal Transport

As in Section 4, let (X, μ) and (Y, ν) be probability spaces, $\Pi(\mu, \nu)$ the set of couplings and $P \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$. Moreover, $c : \mathsf{X} \times \mathsf{Y} \to (-\infty, \infty]$ is measurable with $P\{c < \infty\} > 0$. In this section we study the limit $\varepsilon \to 0$ of

$$\mathcal{C}_\varepsilon = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi + \varepsilon H(\pi|P). \tag{εEOT}$$

Under suitable conditions, we expect convergence to the (unregularized) optimal transport problem that corresponds to $\varepsilon = 0$,

$$\mathcal{C}_0 = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi. \tag{OT}$$

Specifically, we are interested in the convergence of three objects: the value functions $\mathcal{C}_\varepsilon$, the optimal couplings π_ε, and the rescaled EOT potentials $(\varphi_\varepsilon, \psi_\varepsilon)$.

We first recall some notions from optimal transport. The *dual optimal transport problem* is[18]

$$\sup_{\varphi \in L^1(\mu),\, \psi \in L^1(\nu),\, \varphi \oplus \psi \leq c} \mu(\varphi) + \nu(\psi), \tag{5.1}$$

and we call any solution (φ_0, ψ_0) a pair of *Kantorovich potentials.* The optimal transport problem also has an analogue of the cyclical invariance property, but here the definition refers to the support of the coupling. A set $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ is called *c-cyclically monotone* if for all $k \geq 1$,

$$\sum_{i=1}^k c(x_i, y_i) \leq \sum_{i=1}^k c(x_i, y_{i+1}) \quad \text{for} \quad (x_i, y_i) \in \Gamma, \quad 1 \leq i \leq k, \tag{5.2}$$

with the cyclical convention $y_{k+1} := y_1$. A measure π is called c-cyclically monotone if it is concentrated on a c-cyclically monotone set Γ. We will only use this notion when the cost function c is continuous, and then it

[18] More precisely, the supremum is taken over $\varphi \in L^1(\mu), \psi \in L^1(\nu)$ admitting versions with $\varphi \oplus \psi \leq c$ everywhere on $\mathsf{X} \times \mathsf{Y}$. If X, Y are Polish and c is upper semicontinuous, that holds as soon as $\varphi \oplus \psi \leq c$ $\mu \otimes \nu$-a.s.

is equivalent to require that the support $\operatorname{spt} \pi$ be c-cyclically monotone, where $\operatorname{spt} \pi$ is the smallest closed set $A \subset \mathsf{X} \times \mathsf{Y}$ with $\pi(A) = 1$.

The optimal transport problem lacks the general smoothness of (εEOT), hence the regularity properties of c will be more important in this section. If X, Y are Polish spaces and $c \in L^1(\mu \otimes \nu)$ is lower semicontinuous and bounded from below, the following results are standard (see [35, Theorem 5.10, Remark 5.14]): (OT) admits a minimizer ("optimal transport") $\pi_0 \in \Pi(\mu, \nu)$, the dual problem (5.1) admits Kantorovich potentials (φ_0, ψ_0), and we have the optimal transport duality

$$\mathcal{C}_0 = \mu(\varphi_0) + \nu(\psi_0). \tag{5.3}$$

Moreover, $\pi \in \Pi(\mu, \nu)$ is an optimal transport if and only if it is c-cyclically monotone. The problems (OT) and (5.1) are linear, hence uniqueness does not hold in general. Nevertheless, uniqueness is known for many important examples (see [35]).

For our goal of connecting (εEOT) with (OT), the following example illustrates that lower semicontinuity of c alone is not sufficient.

Example 5.1. Consider the lower semicontinuous cost function

$$c(x, y) = \begin{cases} 1, & x \neq y, \\ 0, & x = y \end{cases}$$

with marginals $\mu = \nu = \operatorname{Unif}[0, 1]$ and $P = \mu \otimes \nu$. Then any $\pi \in \Pi(\mu, \nu)$ with $\pi \ll P$ has transport cost $\int c \, d\pi = 1$. On the other hand, the identical coupling $\pi_0 = \mu \otimes_x \delta_x$ (which corresponds to the transport map $T(x) = x$) has vanishing transport cost, and it is the unique optimal transport. Moreover, any Kantorovich potentials φ_0, ψ_0 of (OT) must satisfy $\varphi_0(x) + \psi_0(x) = 0$ μ-a.s. We observe that

(i) $\mathcal{C}_\varepsilon \equiv 1$ does not converge to $\mathcal{C}_0 = 0$,

(ii) $\pi_\varepsilon \equiv \mu \otimes \nu$ does not converge to π_0,

(iii) the rescaled EOT potentials $\varphi_\varepsilon \oplus \psi_\varepsilon \equiv 1$ do not converge to Kantorovich potentials.

The disconnect between the problems is apparent: the optimal transport π_0 exploits the smaller values of c on the diagonal, whereas the regularized problem does not "see" the diagonal (or any $\mu \otimes \nu$-nullsets), hence for this problem the cost c is equivalent to a constant cost $\tilde{c} \equiv 1$.

We are mainly interested in continuous costs c and entropy relative to $P = \mu \otimes \nu$ in (εEOT). Some basic results can be stated in greater generality without any additional effort. One restriction we impose throughout is that

$$c \text{ is bounded from below.} \tag{5.4}$$

This ensures that $\int c\, d\pi$ is well defined and the transport cost $\pi \mapsto \int c\, d\pi$ is weakly lower semicontinuous on $\Pi(\mu, \nu)$ when c is lower semicontinuous. The lower bound can often be relaxed to $c \geq c_1 \oplus c_2 \in L^1(\mu) \oplus L^1(\nu)$ by applying our results to $\tilde{c} = c - c_1 \oplus c_2 \geq 0$ and using Remark 4.4 as well as its analogue for optimal transport.

5.1 Convergence of Value Functions

Clearly the value function $\mathcal{C}_\varepsilon$ of (εEOT) is monotone decreasing as $\varepsilon \geq 0$ decreases, hence $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon \geq \mathcal{C}_0$. Our aim is to show that $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$; that is, we retrieve the value function of (OT) as the regularization parameter tends to zero. In this section, X, Y are Polish and $P \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is arbitrary.

To show that $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$, we can work with the primal or the dual problem. On the primal side, it suffices to show that there exists an almost-optimal transport with finite entropy as follows.

Lemma 5.2. *Suppose that given $\eta > 0$, there exists $\pi^\eta \in \Pi(\mu, \nu)$ with $\int c\, d\pi^\eta \leq \mathcal{C}_0 + \eta$ and $H(\pi^\eta|P) < \infty$. Then $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.*

Proof. Given $\eta > 0$, we have

$$\mathcal{C}_\varepsilon \leq \int c\, d\pi^\eta + H(\pi^\eta|P) \leq \mathcal{C}_0 + \eta + \varepsilon H(\pi^\eta|P).$$

Thus $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon \leq \mathcal{C}_0 + \eta$, and $\eta > 0$ was arbitrary. □

There are different ways to produce this almost-optimal transport π^η. One is to take the entropic optimizer π_ε (which of course has finite entropy) and check that it is almost-optimal for (OT). This idea is investigated in Section 5.3. Another idea is to take an optimal transport for (OT) and "smear it out" such as to ensure finite entropy without affecting the transport cost too much, which is the approach we present next.

Lemma 5.3 (Block Approximation)**.** *Let c be continuous and bounded. Given $\eta > 0$ and $\pi \in \Pi(\mu, \nu)$, there exist $\tilde{\pi} \in \Pi(\mu, \nu)$ such that*

$$\left| \int c\, d\tilde{\pi} - \int c\, d\pi \right| \leq \eta \qquad \text{and} \qquad \frac{d\tilde{\pi}}{d(\mu \otimes \nu)} \quad \text{is bounded.}$$

Proof. Step 1. We first suppose that π is concentrated on a compact set $K = K_{\mathsf{X}} \times K_{\mathsf{Y}}$. Given $\delta > 0$, compactness of K_{X} yields a measurable finite partition $A_1, \dots, A_N$ of K_{X} with $\operatorname{diam} A_i \leq \delta$, and similarly $B_1, \dots, B_{N'}$ for K_{Y}. Consider the "block approximation"

$$\tilde{\pi} := \sum_{i,j} \pi(A_i \times B_j)\, \mu_i \otimes \nu_j, \quad \mu_i := \mu(A_i)^{-1} \mu|_{A_i}, \quad \nu_j := \nu(B_j)^{-1} \nu|_{B_j}.$$

Note that $\tilde{\pi} \in \Pi(\mu, \nu)$ and $d\tilde{\pi}/d(\mu \otimes \nu)$ is bounded (for fixed δ). Moreover, uniform continuity of c on the compact K implies that $\int c\, d\tilde{\pi} \to \int c\, d\pi$ as $\delta \to 0$; note that both integrals are comparable to $\sum_{i,j} c(x_i, y_j) \pi(A_i \times B_j)$ for arbitrary $x_i \in A_i$ and $y_j \in B_j$.

Step 2. To treat the general case, let $\delta > 0$. As π is tight, we can find a compact $K = K_{\mathsf{X}} \times K_{\mathsf{Y}}$ such that $\pi(K) > 1 - \delta$. Let

$$\pi' := \pi|_K, \qquad \pi'' := \pi - \pi'.$$

By Step 1, there is a measure $\tilde{\pi}'$ having the same marginals as π' such that $d\tilde{\pi}'/d(\mu \otimes \nu)$ is bounded and $|\int c\, d\tilde{\pi}' - \int c\, d\pi'| < \eta$. If $\pi'' \neq 0$, let (μ'', ν'') be the marginals of π'' and let $\tilde{\pi}''$ be their product coupling

$$\tilde{\pi}'' = \mu''(\mathsf{X})^{-1}\, \mu'' \otimes \nu''.$$

As c is bounded and $\pi''(\mathsf{X} \times \mathsf{Y}) < \delta$, we have $\int c\, d\tilde{\pi}'' \to 0$ and $\int c\, d\pi'' \to 0$ for $\delta \to 0$. Moreover, $d\tilde{\pi}''/d(\mu \otimes \nu) \leq \mu''(\mathsf{X})^{-1}$. In summary, the coupling $\tilde{\pi} := \tilde{\pi}' + \tilde{\pi}'' \in \Pi(\mu, \nu)$ satisfies the assertion. □

Corollary 5.4. *Let c be continuous and bounded. If $H(\mu \otimes \nu | P) < \infty$, then* $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.

Proof. Let $\eta > 0$ and $\pi \in \Pi(\mu, \nu)$ an optimal transport for (OT). Applying Lemma 5.3 to π yields $\tilde{\pi} \in \Pi(\mu, \nu)$ with $\int c\, d\tilde{\pi} \leq \mathcal{C}_0 + \eta$ and $d\tilde{\pi}/d(\mu \otimes \nu)$ bounded. The latter implies $H(\tilde{\pi}|P) < \infty$ due to $H(\mu \otimes \nu|P) < \infty$, so Lemma 5.2 applies. □

While not needed here, we mention that the block construction can be used to further quantify the distance between π and $\tilde{\pi}$. In the proof of Lemma 5.3, boundedness of c is used for the convergence $\int c\, d\tilde{\pi}'' \to 0$. At least in some cases, a block approximation can also be implemented for unbounded costs (see [6, 22]). For the present purpose it is not important to obtain a bounded density; finite entropy is sufficient. But even so, the simple approximation from Step 2 in the proof is rather crude. Instead, one may typically have to divide the whole domain into small blocks and carefully control the resulting entropy. In Sections 5.3 and 5.4 below, we explore a different route to obtain $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$ for unbounded cost: the convergence of primal and dual optimizers, respectively.

5.2 Convergence of Optimal Couplings for Finite Entropy

Since π_ε minimizes the entropy among all couplings having the same (or smaller) transport cost $\int c\,d\pi$, it is intuitive that the limit as $\varepsilon \to 0$ should have the same property: π_ε should converge to the optimal transport with minimal entropy—provided that such a transport exists. The next result makes this precise. It turns out that the driving ingredient is the convergence $\mathcal{C}_\varepsilon \to \mathcal{C}_0$ of the value functions (as provided by Corollary 5.4 above or Theorem 5.10 and Corollary 5.17 below); if that is taken as a primitive, the analysis falls into the general framework of Section 1 and the conclusions follow easily in a very general setting. Thus, in this section, X, Y are general measurable spaces and $P \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is arbitrary. We suppose that $\mathcal{C}_\varepsilon < \infty$ for some (and then all) $\varepsilon > 0$ and write $\Pi_{opt}(\mu, \nu)$ for the set of all optimal transports; i.e., all $\pi \in \Pi(\mu, \nu)$ with $\int c\,d\pi = \mathcal{C}_0$.

Theorem 5.5. *Suppose that* $\lim_{\varepsilon\to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. *Then the following are equivalent:*

(i) $\lim_{\varepsilon\to 0} H(\pi_\varepsilon|P) < \infty$,

(ii) (π_ε) *converges in variation and* $H(\lim_{\varepsilon\to 0} \pi_\varepsilon|P) < \infty$,

(iii) there exists $\pi \in \Pi_{opt}(\mu, \nu)$ *with* $H(\pi|P) < \infty$.

Under those conditions, the limit $\pi_* = \lim_\varepsilon \pi_\varepsilon$ *is the (unique) optimal transport with minimal entropy:*

$$\pi_* = \operatorname*{arg\,min}_{\Pi_{opt}(\mu,\nu)} H(\,\cdot\,|P).$$

Moreover, $H(\pi_\varepsilon|P) \to H(\pi_*|P)$ *as well as*

$$H(\pi_*|\pi_\varepsilon) \to 0 \qquad \textit{and} \qquad \log \frac{d\pi_\varepsilon}{dP} \to \log \frac{d\pi_*}{dP} \quad \textit{in} \quad L^1(\pi_*).$$

Proof. Let π_ε be the optimizer of ($\varepsilon\mathsf{EOT}$). The additive form of ($\varepsilon\mathsf{EOT}$) and the optimality of the couplings imply that

$$H(\pi_\varepsilon|P) \le H(\pi_{\varepsilon'}|P) \quad \text{and} \quad \int c\,d\pi_\varepsilon \ge \int c\,d\pi_{\varepsilon'} \quad \text{for} \quad \varepsilon \ge \varepsilon' > 0. \tag{5.5}$$

Denote $\mathcal{Q} := \Pi_{opt}(\mu, \nu)$ and

$$\mathcal{Q}_\varepsilon := \left\{\pi \in \Pi(\mu, \nu) : \int c\,d\pi \le \int c\,d\pi_\varepsilon\right\}.$$

Then $\mathcal{Q}_\varepsilon$ is a closed convex set and $\pi_\varepsilon = \arg\min_{\mathcal{Q}_\varepsilon} H(\,\cdot\,|P)$. The second part of (5.5) implies that $\mathcal{Q}_\varepsilon \supset \mathcal{Q}_{\varepsilon'}$ for $\varepsilon \geq \varepsilon'$, and moreover $\cap_\varepsilon \mathcal{Q}_\varepsilon = \mathcal{Q}$ due to $\int c\,d\pi \leq \int c\,d\pi_\varepsilon \leq \mathcal{C}_\varepsilon \to \mathcal{C}_0$ for $\pi \in \cap_\varepsilon \mathcal{Q}_\varepsilon$. This puts us in the setting of Proposition 1.17 which provides all the claims. □

The typical applications for Theorem 5.5 concern discrete and semi-discrete optimal transport (semi-discrete meaning that one marginal is discrete and the other continuous). When both marginals are continuous, typically all optimal transports are singular with respect to P, and then Theorem 5.5 only tells us that that (π_ε) cannot converge in variation. In any event, we can deduce the following dichotomy about the speed of convergence $\mathcal{C}_\varepsilon \to \mathcal{C}_0$.

Corollary 5.6. *We have $\mathcal{C}_\varepsilon - \mathcal{C}_0 = O(\varepsilon)$ if and only if* (OT) *admits an optimal transport with finite entropy. More precisely,*

$$\mathcal{C}_\varepsilon = \mathcal{C}_0 + \varepsilon H(\pi_*|P) + o(\varepsilon) \quad \textit{if} \quad \pi_* = \operatorname*{arg\,min}_{\Pi_{opt}(\mu,\nu)} H(\,\cdot\,|P)$$

whereas $\lim \varepsilon^{-1}(\mathcal{C}_\varepsilon - \mathcal{C}_0) = \infty$ *if* (OT) *admits no optimal transport with finite entropy.*

Proof. Clearly $\int c\,d\pi_\varepsilon \geq \mathcal{C}_0$, hence we have the lower bound

$$\frac{\mathcal{C}_\varepsilon - \mathcal{C}_0}{\varepsilon} = \frac{\int c\,d\pi_\varepsilon + \varepsilon H(\pi_\varepsilon|P) - \mathcal{C}_0}{\varepsilon} \geq H(\pi_\varepsilon|P). \tag{5.6}$$

Suppose there is no optimal transport with finite entropy. If $\mathcal{C}_\varepsilon \to \mathcal{C}_0$, then Theorem 5.5 applies and yields $\lim H(\pi_\varepsilon|P) = \infty$, so that (5.6) implies $\lim \varepsilon^{-1}(\mathcal{C}_\varepsilon - \mathcal{C}_0) = \infty$. If $\mathcal{C}_\varepsilon \not\to \mathcal{C}_0$, then $\lim \varepsilon^{-1}(\mathcal{C}_\varepsilon - \mathcal{C}_0) = \infty$ is trivial.

Otherwise, let $\pi_* := \arg\min_{\Pi_{opt}(\mu,\nu)} H(\,\cdot\,|P)$ be the optimal transport with minimal entropy. We have

$$\mathcal{C}_\varepsilon \leq \int c\,d\pi_* + \varepsilon H(\pi_*|P) = \mathcal{C}_0 + \varepsilon H(\pi_*|P)$$

and hence the upper bound

$$\frac{\mathcal{C}_\varepsilon - \mathcal{C}_0}{\varepsilon} \leq H(\pi_*|P). \tag{5.7}$$

This shows in particular that $\lim_\varepsilon \mathcal{C}_\varepsilon = \mathcal{C}_0$. Hence Theorem 5.5 yields $\lim_\varepsilon H(\pi_\varepsilon|P) = H(\pi_*|P)$ and now the claim follows from (5.6) and (5.7). □

5.3 Weak Convergence of Optimal Couplings

In this section, X, Y are Polish and we study the convergence of (π_ε) in the sense of weak convergence. In many continuous examples, (OT) admits a unique optimal transport π_* given by a transport map (see [35]) and hence $\pi_* \not\ll \mu \otimes \nu$. On the other hand, $\pi_\varepsilon \ll \mu \otimes \nu$ (at least when $P = \mu \otimes \nu$), so that π_ε cannot converge to π_* in variation. By contrast, we shall see that weak convergence holds under general conditions. As before, $P \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is arbitrary.

As a first step, we recall that $\Pi(\mu, \nu)$ is weakly compact. We state a slightly more general result for families of marginals.

Lemma 5.7. *If $\mathcal{M} \subset \mathcal{P}(\mathsf{X})$ and $\mathcal{N} \subset \mathcal{P}(\mathsf{Y})$ are weakly compact, then the set $\{\pi \in \Pi(\mu, \nu) : \mu \in \mathcal{M},\, \nu \in \mathcal{N}\} \subset \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ is weakly compact.*

Proof. Let $\delta > 0$. By Prokhorov's theorem there is a compact set $K_\mathsf{X} \subset \mathsf{X}$ with $\mu(K_X) > 1 - \delta$ for all $\mu \in \mathcal{M}$, and similarly for Y. If $\pi \in \Pi(\mu, \nu)$ for some $(\mu, \nu) \in \mathcal{M} \times \mathcal{N}$, then $\pi(K_\mathsf{X} \times K_\mathsf{Y}) \geq 1 - 2\delta$. By the reverse direction of Prokhorov's theorem, this shows that $\{\pi \in \Pi(\mu, \nu) : \mu \in \mathcal{M},\, \nu \in \mathcal{N}\}$ is relatively compact. To see that it is weakly closed, consider a sequence $\pi_n \in \Pi(\mu_n, \nu_n)$ such that $\pi_n \to \pi$ weakly for some $\pi \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$. As the coordinate projection $\mathsf{X} \times \mathsf{Y} \to \mathsf{X}$ is continuous, it follows that μ_n converges weakly to the first marginal μ of π, and $\mu \in \mathcal{M}$ by the closedness of $\mathcal{M}$. Similarly for the second marginal. □

For the fixed marginals (μ, ν), Lemma 5.7 shows in particular that $\Pi(\mu, \nu)$ is weakly compact, and hence the following.

Lemma 5.8. *If $\varepsilon_n \to 0$, then (π_{ε_n}) has a weakly convergent subsequence and the limit is in $\Pi(\mu, \nu)$.*

We want to show that this limit is an optimal transport. Two approaches will be exemplified. The first is straightforward: we take $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$ for granted (e.g., from Corollary 5.4) and deduce optimality from the lower semicontinuity of the transport cost.

Proposition 5.9. *Let c be lower semicontinuous and $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0 < \infty$. If $\varepsilon_n \to 0$ and $\lim_n \pi_{\varepsilon_n} = \pi$ weakly, then $\pi \in \Pi(\mu, \nu)$ is an optimal transport. If (OT) admits a unique optimal transport π_*, it follows that $\lim_{\varepsilon \to 0} \pi_\varepsilon = \pi_*$ weakly.*

Proof. As c is lower semicontinuous and bounded from below, Portmanteau's theorem yields

$$\int c \, d\pi \leq \liminf \int c \, d\pi_{\varepsilon_n} \leq \liminf \mathcal{C}_{\varepsilon_n} = \mathcal{C}_0.$$

On the strength of Lemma 5.8, the second assertion is a consequence. □

The second approach is to show optimality of the limit intrinsically: we show that any weak limit must be c-cyclically monotone, hence an optimal transport. Under a mild integrability condition, this also implies that $\lim_{\varepsilon\to 0}\mathcal{C}_\varepsilon = \mathcal{C}_0$. In the remainder of this section, we specialize to $P = \mu\otimes\nu$. The main result reads as follows.

Theorem 5.10. *Let $P = \mu\otimes\nu$ and let c be continuous with $c \le c_1 \oplus c_2$ for some $c_1 \in L^1(\mu)$ and $c_2 \in L^1(\nu)$.*[19]

(a) We have $\lim_{\varepsilon\to 0}\mathcal{C}_\varepsilon = \mathcal{C}_0$.

(b) If $\varepsilon_n \to 0$ and $\lim_n \pi_{\varepsilon_n} = \pi$ weakly, then $\pi \in \Pi(\mu,\nu)$ is an optimal transport.

(c) If (OT) *admits a unique optimal transport π_*, then $\lim_{\varepsilon\to 0}\pi_\varepsilon = \pi_*$ weakly.*

To prove the key part (b), we use the intrinsic characterization of π_ε by cyclical invariance. Recalling that (εEOT) corresponds to a reference measure $dR \propto e^{-c/\varepsilon}dP$ in the setting of Schrödinger bridges, Definition 2.6 translates as follows.

Definition 5.11. A coupling $\pi \in \Pi(\mu,\nu)$ is called (c,ε)-*cyclically invariant* if $\pi \sim P$ and its density admits a version $\frac{d\pi}{dP} : \mathsf{X}\times\mathsf{Y} \to (0,\infty)$ such that

$$\prod_{i=1}^{k}\frac{d\pi}{dP}(x_i,y_i) = \exp\left(-\frac{1}{\varepsilon}\left[\sum_{i=1}^{k}c(x_i,y_i) - \sum_{i=1}^{k}c(x_i,y_{i+1})\right]\right)\prod_{i=1}^{k}\frac{d\pi}{dP}(x_i,y_{i+1}) \tag{5.8}$$

for all $k \in \mathbb{N}$ and $(x_i,y_i)_{i=1}^k \subset \mathsf{X}\times\mathsf{Y}$, where $y_{k+1} := y_1$.

We observe that the exponent contains the sums from the definition (5.2) of c-cyclical monotonicity. If their difference is positive, the right-hand side of (5.8) decays exponentially fast as $\varepsilon \to 0$. In [1], that observation is exploited to derive a large deviations principle for (π_ε) as $\varepsilon \to 0$. Here, we only use the fact that the right-hand side tends to zero.

Proposition 5.12. *Let c be continuous and let π_ε be (c,ε)-cyclically invariant for $\varepsilon > 0$. If $\varepsilon_n \to 0$ and $\lim \pi_{\varepsilon_n} = \pi$ weakly, then $\pi \in \Pi(\mu,\nu)$ is c-cyclically monotone. In particular, π is an optimal transport as soon as $\mathcal{C}_0 < \infty$.*

[19] The theorem remains valid if $c \le c_1 \oplus c_2$ is replaced by $c \in L^1(\mu\otimes\nu)$; cf. Corollary 5.17

The proof is based on the following lemma capturing the aforementioned exponential decay.

Lemma 5.13. *Let $k \geq 2$ and $\delta \geq 0$. Define*

$$A_k(\delta) := \left\{ (x_i, y_i)_{i=1}^k \in (\mathsf{X} \times \mathsf{Y})^k : \sum_{i=1}^k c(x_i, y_i) - \sum_{i=1}^k c(x_i, y_{i+1}) \geq \delta \right\}.$$

Then $\pi_\varepsilon^k := \prod_{i=1}^k \pi_\varepsilon(dx_i, dy_i) \in \mathcal{P}((\mathsf{X} \times \mathsf{Y})^k)$ satisfies

$$\pi_\varepsilon^k(A_k(\delta)) \leq e^{-\delta/\varepsilon} \quad \textit{for all} \quad \varepsilon > 0. \tag{5.9}$$

Proof. Set $Z = d\pi_\varepsilon/dP$. Using (5.8), we have for P^k-a.e. $(x_i, y_i)_{i=1}^k \in A$ that

$$\begin{aligned} \prod Z(x_i, y_i) &= \exp\left\{-\varepsilon^{-1}\left[\textstyle\sum c(x_i, y_i) - \sum c(x_i, y_{i+1})\right]\right\} \prod Z(x_i, y_{i+1}) \\ &\leq e^{-\delta/\varepsilon} \prod Z(x_i, y_{i+1}). \end{aligned}$$

Let $\bar{A} := \left\{(x_i, y_{i+1})_{i=1}^k : (x_i, y_i)_{i=1}^k \in A\right\}$. Integrating over A with respect to $P^k = \prod P(dx_i, dy_i) = \prod P(dx_i, dy_{i+1})$ yields

$$\pi_\varepsilon^k(A) \leq e^{-\delta/\varepsilon} \pi_\varepsilon^k(\bar{A}) \leq e^{-\delta/\varepsilon},$$

where we have used that π_ε^k is a probability measure. □

Proof of Proposition 5.12. We show that $\operatorname{spt} \pi$ is a c-cyclically invariant set. Suppose for contradiction that there are $(x_i, y_i) \in \operatorname{spt} \pi$, $1 \leq i \leq k$ with $\sum_i c(x_i, y_i) > \sum_i c(x_i, y_{i+1})$. By continuity of c there exist $\delta > 0$ and open neighborhoods $U_i \ni (x_i, y_i)$ such that $\sum_i c(\tilde{x}_i, \tilde{y}_i) \geq \delta + \sum_i c(\tilde{x}_i, \tilde{y}_{i+1})$ for all $(\tilde{x}_i, \tilde{y}_i) \in U_i$. Moreover, $\pi(U_i) > 0$ and hence $\liminf_n \pi_{\varepsilon_n}(U_i) > 0$. On the other hand, $U_1 \times \cdots \times U_k \subset A_k(\delta)$ implies $\pi_{\varepsilon_n}^k(U_1 \times \cdots \times U_k) \to 0$ by (5.9), a contradiction. □

The following fact will be used to derive $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$ from the convergence of optimizers; i.e., to derive (a) from (b) in Theorem 5.10.

Lemma 5.14. *Let $|c| \leq c_1 \oplus c_2$ for some $c_1 \in L^1(\mu)$ and $c_2 \in L^1(\nu)$. Then c is $\Pi(\mu, \nu)$-uniformly integrable.*

Proof. We may assume that $c, c_1, c_2 \geq 0$. By the la Vallée–Poussin theorem [2, Theorem 4.5.9, p. 272], $2(c_1 \oplus c_2) \in L^1(\mu \otimes \nu)$ implies that there exists a convex, increasing, superlinearly growing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\phi(2(c_1 \oplus c_2)) \in L^1(\mu \otimes \nu)$. Fubini's theorem then shows that also $\phi(2c_1) \in L^1(\mu)$ and $\phi(2c_2) \in L^1(\nu)$. Let $\pi \in \Pi(\mu, \nu)$, then by convexity,

$$\begin{aligned}\int \phi(c)\, d\pi \leq \int \phi\left(2\frac{c_1 \oplus c_2}{2}\right) d\pi &\leq \frac{1}{2}\int \phi(2c_1) \oplus \phi(2c_2)\, d\pi \\ &= \frac{1}{2}\int \phi(2c_1)\, d\mu + \int \phi(2c_2)\, d\nu < \infty.\end{aligned}$$

Thus $\sup_{\pi\in\Pi(\mu,\nu)} \int \phi(c)\, d\pi < \infty$ and now the converse direction of the la Vallée–Poussin theorem yields the claim. □

Proof of Theorem 5.10. The assumption clearly implies $\mathcal{C}_\varepsilon < \infty$. As π_ε is (c, ε)-cyclically invariant by Corollary 2.9, (b) follows from Proposition 5.12. In the light of Lemma 5.8 and Lemma 5.14, (b) implies (a) and (c). □

Remark 5.15. Uniqueness of the optimal transport is known in many important examples, for instance for quadratic cost $c(x, y) = \|x - y\|^2$ on $\mathbb{R}^d \times \mathbb{R}^d$ when μ (or ν) is absolutely continuous (see [35]). It seems plausible that $\lim_{\varepsilon\to 0} \pi_\varepsilon$ exists even without this uniqueness; i.e., the entropic regularization would select a particular optimal transport in the limit. We have seen this in Theorem 5.5 where the minimal entropy optimal transport $\pi_* = \arg\min_{\Pi_{opt}(\mu,\nu)} H(\,\cdot\,|P)$ is selected, but in general it is open how to formalize an analogue if all optimal transports have infinite entropy. One example where selection is known, is the 1-dimensional Monge problem [14].

5.4 Convergence of Potentials

Throughout this section, X, Y are Polish, $P = \mu \otimes \nu$ and $c \in L^1(P)$ is continuous. By Section 4, the rescaled EOT potentials $(\varphi_\varepsilon, \psi_\varepsilon)$ exist, are integrable, and solve the dual EOT problem. Our aim is the show the convergence of $\varphi_\varepsilon \oplus \psi_\varepsilon$ to a dual solution $\varphi_0 \oplus \psi_0$ of the optimal transport problem. To state a separate convergence of φ_ε and ψ_ε, it is clearly necessary to choose a normalization. We use the symmetric convention $\mu(\varphi_\varepsilon) = \nu(\psi_\varepsilon)$, and then the same will hold for the limit. Results for other normalizations, for instance $\mu(\varphi_\varepsilon) = 0$, are an immediate consequence.

Theorem 5.16. *Let $P = \mu \otimes \nu$ and let $c \in L^1(P)$ be continuous. Let $(\varphi_\varepsilon, \psi_\varepsilon)$ be the rescaled EOT potentials for $\varepsilon > 0$.*

(a) Given $\varepsilon_n \to 0$, there is a subsequence (ε_k) such that φ_{ε_k} converges in $L^1(\mu)$ and ψ_{ε_k} converges in $L^1(\nu)$.

(b) If $\lim_n \varphi_{\varepsilon_n} = \varphi$ μ-a.s. and $\lim_n \psi_{\varepsilon_n} = \psi$ ν-a.s. for $\varepsilon_n \to 0$, then (φ, ψ) are Kantorovich potentials and the convergence also holds in L^1.

In particular, if the Kantorovich potentials (φ_0, ψ_0) are a.s. unique, then $\lim_\varepsilon \varphi_\varepsilon = \varphi_0$ *in* $L^1(\mu)$ *and* $\lim_\varepsilon \psi_\varepsilon = \psi_0$ *in* $L^1(\nu)$*. Moreover, if* X, Y *are compact,* L^1*-convergence can be strengthened to uniform convergence in all assertions.*

In many important examples, it is known that the Kantorovich potentials are a.s. unique (e.g., [1, Appendix B]), and then Theorem 5.16 yields a clear-cut result on the convergence of the potentials in L^1. In the case of non-uniqueness, the situation is similar as in the primal problem (Remark 5.15): we have a compactness result and we may conjecture that a particular limit is selected, but it seems that this is known only for discrete marginals (see [9]).

The main difficulty in Theorem 5.16 is to establish compactness in a suitable sense. We only detail the proof in the special case where X, Y are compact, so that we can use the standard Arcelà–Ascoli theorem (see also [24]). For the general result, see [28].

Proof (Compact Case). When X, Y are compact, the continuous function c is bounded and uniformly continuous. Lemma 4.9 and Lemma 4.11 (and Remark 4.12) then show that $\varphi_\varepsilon, \psi_\varepsilon$ are bounded and uniformly continuous, uniformly in ε (after choosing suitable versions). By the Arcelà–Ascoli theorem, $(\varphi_\varepsilon)_{\varepsilon>0}$ is relatively compact in the topology of uniform convergence, and similarly for $(\psi_\varepsilon)_{\varepsilon>0}$. This shows (a). Regarding (b), it also shows that if $\lim_n \varphi_{\varepsilon_n} = \varphi$ μ-a.s. and $\lim_n \psi_{\varepsilon_n} = \psi$ ν-a.s., this convergence necessarily uniform and the limits φ, ψ are continuous (again, after choosing suitable versions). Let $\delta > 0$ and $A := \{\varphi \oplus \psi \geq c + 2\delta\}$. The uniform convergence implies that for $\varepsilon > 0$ small enough,

$$\frac{d\pi_\varepsilon}{dP} = e^{\frac{\varphi_\varepsilon \oplus \psi_\varepsilon - c}{\varepsilon}} \geq e^{\delta/\varepsilon} \quad \text{on} \quad A,$$

hence $P(A) > 0$ would imply $\pi_\varepsilon(A) > 1$ for $\varepsilon > 0$ small. We conclude that $\varphi \oplus \psi \leq c$ P-a.s. But since φ, ψ, c are all continuous, this already implies that $\varphi \oplus \psi \leq c$ on the support $\operatorname{spt} P = \operatorname{spt} \mu \times \operatorname{spt} \nu$. After possibly changing φ, ψ on nullsets, we have that $\varphi \oplus \psi \leq c$ on $\mathsf{X} \times \mathsf{Y}$.

It remains to show that φ, ψ are dual optimizers. Combining the uniform convergence with the EOT duality (4.9) and Corollary 5.4, we have

$$\mu(\varphi) + \nu(\psi) = \lim \left[\mu(\varphi_\varepsilon) + \nu(\psi_\varepsilon)\right] = \lim \mathcal{C}_\varepsilon = \mathcal{C}_0.$$

In view of the optimal transport duality (5.3), this shows that φ, ψ are dual optimizers and the proof is complete. □

Corollary 5.17. *Let* $P = \mu \otimes \nu$ *and let* $c \in L^1(P)$ *be continuous. Then* $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.

This follows from the L^1-convergence in Theorem 5.16 via duality:

$$\mathcal{C}_0 = \mu(\varphi) + \nu(\psi) = \lim \left[\mu(\varphi_{\varepsilon_k}) + \nu(\psi_{\varepsilon_k})\right] = \lim \mathcal{C}_{\varepsilon_k}.$$

We have cheated in this argument, though, as the proof of the L^1-convergence in [28] partially uses $\lim_{\varepsilon\to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. Specifically, [28] first shows the convergence of the potentials in probability. Using the uniformly integrable upper bound resulting from Lemma 4.9, this implies $\limsup_{\varepsilon\to 0} \mathcal{C}_\varepsilon \leq \mathcal{C}_0$, which in view of $\mathcal{C}_\varepsilon \geq \mathcal{C}_0$ is enough to conclude $\limsup_{\varepsilon\to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. In [28], L^1-convergence is then deduced by a Scheffé argument.

6 Sinkhorn's Algorithm

Let (X, μ) and (Y, ν) be probability spaces. We fix $R \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ satisfying $R \ll \mu \otimes \nu$ and write, as in Section 2.1,

$$e^{-c(x,y)} = \frac{dR}{d(\mu \otimes \nu)}(x, y).$$

Thus $c : \mathsf{X} \times \mathsf{Y} \to (-\infty, \infty]$ satisfies $c < \infty$ $\mu \otimes \nu$-a.s. if and only if $R \sim \mu \otimes \nu$, and that is the case of principal interest to us. On the other hand, some results hold in the more general setting without additional effort. *Throughout this section, we assume that the Schrödinger problem is finite;* i.e., $\Pi_{fin}(\mu, \nu) \neq \emptyset$. In particular, the (unique) Schrödinger bridge $\pi_* \in \Pi(\mu, \nu)$ exists with $H(\pi_*|R) < \infty$.

In the dual perspective, our aim is to solve the Schrödinger equations

$$\varphi(x) = -\log \int_{\mathsf{Y}} e^{\psi(y) - c(x,y)}\, \nu(dy) \quad \mu\text{-a.s.}, \tag{SE1}$$

$$\psi(y) = -\log \int_{\mathsf{X}} e^{\varphi(x) - c(x,y)}\, \mu(dx) \quad \nu\text{-a.s.} \tag{SE2}$$

We may start with some function φ_0, say $\varphi_0 := 0$, and alternatingly solve the two equations: for $t \geq 0$,

(1) define ψ_t as the solution of (SE2) with $\varphi := \varphi_t$,

(2) define φ_{t+1} as the solution (SE1) with $\psi := \psi_t$,

and iterate. This is the basic idea of *Sinkhorn's algorithm.* If the algorithm reaches a fixed point, $(\varphi_t, \psi_t) = (\varphi_{t+1}, \psi_{t+1})$, then (φ_t, ψ_t) is a solution of (SE1)–(SE2) and we have found the desired potentials. In most cases, it will not reach a fixed point for finite t, but we shall prove convergence to a fixed point under suitable conditions.

To add further motivation beyond the general idea of alternatingly solving the two equations, consider the dual problem from Section 3 and its objective function

$$G(\varphi,\psi) := \mu(\varphi) + \nu(\psi) - \int e^{\varphi\oplus\psi}\, dR + 1.$$

Recalling the interpretation of the Schrödinger equations as Euler–Lagrange equations for optimality (Remark 3.4), the algorithm can be seen as a coordinate ascent scheme for a concave maximization problem: for $t \geq 0$, iterate

(1) $\psi_t := \arg\max G(\varphi_t, \cdot)$,

(2) $\varphi_{t+1} := \arg\max G(\cdot, \psi_t)$.

This implies the monotonicity of the scheme,

$$G(\varphi_t,\psi_t) \leq G(\varphi_{t+1},\psi_t) \leq G(\varphi_{t+1},\psi_{t+1}),$$

and in particular $\lim_t G(\varphi_t,\psi_t)$ exists. By the strict concavity of G, each iteration will strictly increase the value $G(\varphi_t,\psi_t)$, unless a fixed point has been reached. Thus we may expect that $G(\varphi_t,\psi_t) \to G(\varphi_*,\psi_*)$ as well as $\varphi_t \to \varphi_*$ and $\psi_t \to \psi_*$, for some potentials (φ_*,ψ_*).

Algorithm 6.1 (Sinkhorn, Dual Formulation)**.** Set $\varphi_0 := 0$. For $t \geq 0$,

$$\psi_t(y) := -\log \int_{\mathsf{X}} e^{\varphi_t(x)-c(x,y)}\, \mu(dx),$$
$$\varphi_{t+1}(x) := -\log \int_{\mathsf{Y}} e^{\psi_t(y)-c(x,y)}\, \nu(dy).$$

We also define

$$d\pi(\varphi,\psi) := e^{\varphi\oplus\psi}\, dR = e^{\varphi\oplus\psi-c}\, d(\mu\otimes\nu),$$

$$\pi_{2t} := \pi(\varphi_t,\psi_t), \quad \pi_{2t-1} := \pi(\varphi_t,\psi_{t-1}), \quad t \geq 0,$$

were $\psi_{-1} := 0$ and thus $\pi_{-1} = R$.

The primal problem offers another perspective on Sinkhorn's algorithm. Starting with φ_t, Step (1) corresponds to choosing $\psi = \psi_t$ such that the measure

$$e^{\varphi_t\oplus\psi-c}\, d(\mu\otimes\nu)$$

has the required second marginal ν. Then, the choice (2) for $\varphi = \varphi_{t+1}$ corresponds to "fitting" the first marginal of $e^{\varphi\oplus\psi_t-c}\, d(\mu\otimes\nu)$ to be μ. Of course, the second marginal may now be off again, and the iteration continues. This explains why Sinkhorn's algorithm is also known as *iterative proportional fitting procedure,* or IPFP.

There are many choices fitting one marginal. Setting the above algorithm aside for the moment, one natural choice for a primal algorithm is to minimize entropy relative to the last iterate, starting with R. That is, we set $\pi_{-1} := R$ and iterate for $t \geq 0$,

$$\pi_{2t} := \operatorname*{arg\,min}_{\Pi(*,\nu)} H(\,\cdot\,|\pi_{2t-1}), \tag{6.1}$$

$$\pi_{2t+1} := \operatorname*{arg\,min}_{\Pi(\mu,*)} H(\,\cdot\,|\pi_{2t}), \tag{6.2}$$

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 analogous. Next, we rewrite this implicit algorithm in a more explicit form. Let us focus on the second step (6.2), where given $\pi' \ll \mu \otimes \nu$, the next iterate is

$$\pi := \operatorname*{arg\,min}_{\Pi(\mu,*)} H(\,\cdot\,|\pi'). \tag{6.3}$$

To determine the minimizer, we can disintegrate a generic $\pi \in \Pi(\mu,*)$ into $\pi = \mu \otimes K$ and compare with $\pi' = \mu' \otimes K'$, where μ' is the first marginal of π'. Assuming that $\pi \ll \pi'$, we have

$$H(\pi|\pi') = H(\mu|\mu') + \int H(K|K')\, d\mu.$$

The first term is independent of $\pi \in \Pi(\mu,*)$, hence the minimum is attained for $K := K'$ as that makes the second term vanish. For this choice of π,

$$H(\pi|\pi') = H(\mu|\mu') \tag{6.4}$$

as well as

$$\frac{d\pi}{d\pi'}(x,y) = \frac{d\mu}{d\mu'}\frac{dK}{dK'} = \frac{d\mu}{d\mu'}.$$

In conclusion, the iteration (6.1)–(6.2) can be stated explicitly as follows.

Algorithm 6.2 (Sinkhorn, Primal Formulation)**.** Set $\pi_{-1} := R$. For $t \geq 0$, define $\pi_t \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ via

$$\frac{d\pi_{2t}}{d\pi_{2t-1}}(x,y) := \frac{d\nu}{d\nu_{2t-1}}(y),$$
$$\frac{d\pi_{2t+1}}{d\pi_{2t}}(x,y) := \frac{d\mu}{d\mu_{2t}}(x),$$

where (μ_t, ν_t) denotes the marginal distributions of π_t.

Next, we observe that this coincides with the dual formulation given in Algorithm 6.1. Indeed, the conditional density $\frac{dK'}{d\nu}(x,y)$ of π' given x is the quotient of the joint density $\frac{d\pi'}{d(\mu\otimes\nu)}$ and marginal density $\int \frac{d\pi'}{d(\mu\otimes\nu)}\,d\nu$. With $\pi' := \pi_{2t}$ defined as in Algorithm 6.1, this yields

$$\frac{dK'}{d\nu}(x,y) = \frac{e^{\varphi_t(x)+\psi_t(y)-c(x,y)}}{\int e^{\varphi_t(x)+\psi_t(y)-c(x,y)}\nu(dy)} = \frac{e^{\psi_t(y)-c(x,y)}}{\int e^{\psi_t(y)-c(x,y)}\nu(dy)}.$$

In the primal formulation, π_{2t+1} was defined with the kernel $K = K'$ and its marginal density is $d\mu/d\mu = 1$, hence the joint density is

$$\frac{d\pi_{2t+1}}{d(\mu\otimes\nu)}(x,y) = \frac{d\mu}{d\mu}(x)\frac{dK}{d\nu}(x,y) = \frac{e^{\psi_t(y)-c(x,y)}}{\int e^{\psi_t(y)-c(x,y)}\nu(dy)},$$

or equivalently, the log-density is

$$-\log\int e^{\psi_t(y)-c(x,y)}\nu(dy) + \psi_t(y) - c(x,y).$$

The log-density implied by Algorithm 6.1 is $\varphi_{t+1}(x)+\psi_t(y)-c(x,y)$, hence the above matches the definition $\varphi_{t+1}(x) = -\int e^{\psi_t(y)-c(x,y)}\nu(dy)$ in Algorithm 6.1. Similar arguments hold for the update from $2t-1$ to $2t$, and as a result, the two formulation of Sinkhorn's algorithms are equivalent. Below, we mostly use the dual formulation.

Remark 6.3. Sinkhorn's iteration is also used in the context of the entropic optimal transport problem (EOT) where the cost c is not necessarily normalized; i.e., $\alpha := \int e^{-c}\,d(\mu\otimes\nu) \neq 1$. The following shows how our results can be extended to this situation with minimal changes. Indeed, we can introduce a normalized cost

$$\hat{c} := c + \log\alpha$$

which satisfies $\int e^{-\hat{c}}\,d(\mu\otimes\nu) = 1$; i.e., $dR = e^{-\hat{c}} = d(\mu\otimes\nu)$ is a probability.

Denote by (φ_t,ψ_t) the iterates of Algorithm 6.1 for cost c and by $(\hat{\varphi}_t,\hat{\psi}_t)$ the iterates for cost $\hat{c}$. Then a simple induction shows that

$$\varphi_t = \hat{\varphi}_t, \qquad \psi_t = \hat{\psi}_t - \log\alpha, \qquad t \geq 0.$$

(This is analogous to Remark 4.3.) As $e^{\varphi_t\oplus\psi_s-c}\,d(\mu\otimes\nu) = e^{\hat{\varphi}_t\oplus\hat{\psi}_s-\hat{c}}\,d(\mu\otimes\nu)$, the induced measures π_n are the same for both iterations. By applying our results to $\hat{c}$, we can easily deduce the corresponding results for c. In fact, most of the results below are stated in terms of differences such as $\psi_t - \psi_s$ and hence the formulas hold in the unnormalized case without changes.

6.1 Basic Properties and Marginal Convergence

Denoting by (μ_t, ν_t) the marginal distributions of π_t, we recall that

$$\mu_{2t+1} = \mu, \qquad \nu_{2t} = \nu, \qquad t \geq 0; \tag{6.5}$$

that is, every other marginal is correct. One aim of this section is to show that the "incorrect" marginals converge to the correct ones as $t \to \infty$. We first provide some basic properties.

Lemma 6.4. *For all $t \geq 0$ and $n \geq 0$,*

(i) $\varphi_t \in L^1(\mu)$ *and* $\psi_t \in L^1(\nu)$*,*

(ii) $H(\pi_{2t}|\pi_{2t-1}) = \nu(\psi_t - \psi_{t-1})$ *and* $H(\pi_{2t+1}|\pi_{2t}) = \mu(\varphi_{t+1} - \varphi_t)$*,*

(iii) $\mu(\varphi_n) = \sum_{t=0}^{n-1} H(\pi_{2t+1}|\pi_{2t})$ *and* $\nu(\psi_n) = \sum_{t=0}^{n} H(\pi_{2t}|\pi_{2t-1})$*; in particular,* $\mu(\varphi_n)$ *and* $\nu(\psi_n)$ *are nonnegative and increasing,*[20]

Proof. (ii) For $t \geq 0$,

$$H(\pi_{2t}|\pi_{2t-1}) = \int \log \frac{d\pi_{2t}}{d\pi_{2t-1}}\, d\pi_{2t} = \int (\psi_t - \psi_{t-1})\, d\pi_{2t} = \int (\psi_t - \psi_{t-1})\, d\nu \tag{6.6}$$

where we have used (6.5) and the integrals are necessarily well-defined in $[0, \infty]$; in particular, $(\psi_t - \psi_{t-1})^- \in L^1(\nu)$. Similarly,

$$H(\pi_{2t+1}|\pi_{2t}) = \int (\varphi_{t+1} - \varphi_t)\, d\mu.$$

(i) Clearly ψ_{-1} and φ_0 are integrable. Suppose that $\psi_{t-1} \in L^1(\nu)$ and $\varphi_t \in L^1(\mu)$, we show by induction that $\psi_t \in L^1(\nu)$ and $\varphi_{t+1} \in L^1(\mu)$. Indeed, (6.6) and $\psi_{t-1} \in L^1(\nu)$ imply $(\psi_t)^- \in L^1(\nu)$. On the other hand, using $H(\pi_*|R) < \infty$, Lemma 1.4 (a) yields $\log(d\pi_{2t}/dR)^+ = (\varphi_t + \psi_t)^+ \in L^1(\pi_*)$. In view of $\pi_* \in \Pi(\mu, \nu)$ and the hypothesis that $\varphi_t \in L^1(\mu)$, it follows that $(\psi_t)^+ \in L^1(\nu)$. We have shown $\psi_t \in L^1(\nu)$, and the proof that $\varphi_{t+1} \in L^1(\mu)$ is analogous.

(iii) Summing up (6.6) yields

$$\sum_{t=0}^{n} H(\pi_{2t}|\pi_{2t-1}) = \sum_{t=0}^{n} \int (\psi_t - \psi_{t-1})\, d\nu = \int (\psi_n - \psi_{-1})\, d\nu = \nu(\psi_n).$$

Similarly, $\sum_{t=0}^{n} H(\pi_{2t+1}|\pi_{2t}) = \mu(\varphi_{n+1})$. □

[20] In the case of an unnormalized cost c (i.e., $\alpha := \int e^{-c}\, d(\mu \otimes \nu) \neq 1$), the result for ψ_n changes to $\nu(\psi_n) = -\log \alpha + \sum_{t=0}^{n} H(\pi_{2t}|\pi_{2t-1})$ and hence $\nu(\psi_n) \geq -\log \alpha$ instead of $\nu(\psi_n) \geq 0$. See Remark 6.3.

We can now state a key property for the convergence analysis.

Proposition 6.5. *For all $n \geq -1$,*

$$H(\pi_*|\pi_n) = H(\pi_*|R) - \sum_{t=0}^{n} H(\pi_t|\pi_{t-1}). \tag{6.7}$$

In particular, $H(\pi_*|\pi_n)$ is decreasing in n.

Proof. Recalling $\pi_{-1} = R$, the case $n = -1$ is clear. Let $n \geq 0$. From Lemma 6.4 (ii) we obtain $\sum_{t=0}^{2n} H(\pi_t|\pi_{t-1}) = \mu(\varphi_n) + \nu(\psi_n)$. On the other hand, Lemma 1.4 (b) and $H(\pi_*|R) < \infty$ yield

$$\begin{aligned} H(\pi_*|R) - H(\pi_*|\pi_{2n}) = E^{\pi_*}[\log(d\pi_{2n}/dR)] &= E^{\pi_*}[\varphi_n + \psi_n] \\ &= \mu(\varphi_n) + \nu(\psi_n), \end{aligned}$$

where $\pi_* \in \Pi(\mu, \nu)$ and the integrability from Lemma 6.4 (i) were used for the last equality. It follows that

$$H(\pi_*|R) = H(\pi_*|\pi_{2n}) + \sum_{t=0}^{2n} H(\pi_t|\pi_{t-1}).$$

Similarly for $2n$ replaced by $2n + 1$, showing the claim. □

As a consequence, we obtain that the marginals (μ_t, ν_t) of π_t converge to the correct marginals (μ, ν).

Corollary 6.6. *For all $t \geq 1$,*

$$H(\mu_t|\mu) + H(\nu_t|\nu) \leq H(\pi_t|\pi_{t-1}),$$

and the right-hand side is summable with

$$\sum_{t\geq 1} H(\pi_t|\pi_{t-1}) \leq H(\pi_*|R) - H(\pi_0|R) \leq H(\pi_*|R).$$

In particular, $H(\mu_t|\mu) \to 0$ and $H(\nu_t|\nu) \to 0$, and then also $\mu_t \to \mu$ and $\nu_t \to \nu$ in variation.

Proof. If $t \geq 2$ is even, then $\mu = \mu_{t-1}$ and $\nu = \nu_t$ by (6.5), so that

$$H(\mu_t|\mu) + H(\nu_t|\nu) = H(\mu_t|\mu_{t-1}) \leq H(\pi_t|\pi_{t-1}),$$

where the inequality is due the data processing inequality (Example 1.7). Similarly for odd $t \geq 1$, and the first inequality follows. For the second inequality, note that (6.7) yields $\sum_{t=1}^{\infty} H(\pi_t|\pi_{t-1}) \leq H(\pi_*|R) - H(\pi_0|R)$ due to $\pi_{-1} = R$. The convergence in variation follows by Pinsker's inequality (Lemma 1.2). □

The next lemma lists further properties of the Sinkhorn marginals.

Lemma 6.7. *For all $t \geq 0$, we have*

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

as well as $\frac{d\mu_{2t+1}}{d\mu} = 1$ *and* $\frac{d\nu_{2t}}{d\nu} = 1$. *In particular,*

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

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

$$\begin{aligned} H(\mu_{2t+2}|\mu) - H(\mu_{2t+2}|\mu_{2t}) &= \mu_{2t+2}(\varphi_t - \varphi_{t+1}), \\ H(\nu_{2t+1}|\nu) - H(\nu_{2t+1}|\nu_{2t-1}) &= \nu_{2t+1}(\psi_{t-1} - \psi_t). \end{aligned}$$

Proof. We write the marginal density by integrating out the second marginal from the joint density,

$$\begin{aligned} \frac{d\mu_{2t}}{d\mu}(x) &= \int_{\mathsf{Y}} \frac{d\pi_{2t}}{d(\mu \otimes \nu)}(x,y)\,\nu(dy) = \int_{\mathsf{Y}} e^{\varphi_t(x) + \psi_t(y) - c(x,y)}\,\nu(dy) \\ &= e^{\varphi_t(x)} \int_{\mathsf{Y}} e^{\psi_t(y) - c(x,y)}\,\nu(dy) = e^{\varphi_t(x)} e^{-\varphi_{t+1}(x)} \end{aligned}$$

where the last step used the definition of φ_{t+1}. The proof for ν_{2t-1} is analogous. While $\frac{d\mu_{2t+1}}{d\mu} = 1$ and $\frac{d\nu_{2t}}{d\nu} = 1$ can be obtained in the same way, this is also a restatement of (6.5).

The formulas for $H(\mu_{2t}|\mu)$ and $H(\nu_{2t-1}|\nu)$ are now immediate. It also follows that

$$H(\mu|\mu_{2t}) = \int \log \frac{d\mu}{d\mu_{2t}}\,d\mu = -\int \log \frac{d\mu_{2t}}{d\mu}\,d\mu = -\mu(\varphi_t - \varphi_{t+1})$$

which is equal to $H(\pi_{2t+1}|\pi_{2t})$ by Lemma 6.4 (ii). Similarly for $H(\nu|\nu_{2t-1})$. For the last pair of formulas, we use Lemma 1.4 (b) to see that

$$H(\mu_{2t+2}|\mu) - H(\mu_{2t+2}|\mu_{2t}) = \int \log \frac{d\mu_{2t}}{d\mu}\,d\mu_{2t+2} = \mu_{2t+2}(\varphi_t - \varphi_{t+1}),$$

and similarly for ν_{2t+1}. □

Remark 6.8. The formula $H(\mu|\mu_{2t}) = H(\pi_{2t+1}|\pi_{2t})$ in Lemma 6.7 was already derived in (6.4). Taking it as a starting point gives a slightly different way to understand the minimization property (6.3) of Sinkhorn's projection: the data processing inequality (Example 1.7) shows $H(\mu|\mu_{2t}) \leq H(\pi|\pi_{2t})$ for any $\pi \in \Pi(\mu, *)$, and as π_{2t+1} attains this bound, we have

$$\pi_{2t+1} = \operatorname*{arg\,min}_{\pi \in \Pi(\mu,*)} H(\pi|\pi_{2t}).$$

Remark 6.9. We have seen that $H(\mu_t|\mu) \to 0$ and $H(\nu_t|\nu) \to 0$; cf. Corollary 6.6. The reverse entropies also converge: $H(\mu|\mu_t) \to 0$ and $H(\nu|\nu_t) \to 0$. Indeed, $H(\mu|\mu_{2t}) = H(\pi_{2t+1}|\pi_{2t}) \to 0$ by Lemma 6.7 and Corollary 6.6, and of course $H(\mu|\mu_{2t+1}) = 0$. Similarly for ν_t.

6.2 Rate for Marginal Convergence

While Corollary 6.6 shows that $H(\mu_{2t}|\mu) \to 0$, our next aim is to prove a rate for this convergence (and also for $H(\mu|\mu_{2t}) \to 0$). The rate is stated in Corollary 6.12 below; the key step is the monotonicity of $H(\mu_{2t}|\mu)$. These properties were obtained in [25], where the Sinkhorn marginals are framed as a Bregman gradient descent scheme and the properties are derived as structural consequences. Below, we give a proof through on an entropy calculation which is elementary but may lack the deeper explanation.

Proposition 6.10. *For $t \geq 0$, we have*

$$H(\mu_{2t}|\mu) \geq H(\nu|\nu_{2t+1}) \geq H(\mu_{2t+2}|\mu) \geq H(\nu|\nu_{2t+3}) \geq \dots;$$

more precisely,

$$H(\nu|\nu_{2t+1}) = H(\mu_{2t+2}|\mu) + H(\pi_{2t+2}|\pi_{2t}) - H(\mu_{2t+2}|\mu_{2t}) \geq H(\mu_{2t+2}|\mu),$$
$$H(\nu|\nu_{2t+1}) = H(\mu_{2t}|\mu) - H(\pi_{2t}|\pi_{2t+2}) \leq H(\mu_{2t}|\mu).$$

Similarly, $H(\mu|\mu_{2t}) \geq H(\nu_{2t+1}|\nu) \geq H(\mu|\mu_{2t+2}) \geq H(\nu_{2t+3}|\nu) \geq \dots$ for $t \geq 0$. In particular, the sequences

$$\{H(\mu_{2t}|\mu)\}_{t\geq 0}, \quad \{H(\mu|\mu_{2t})\}_{t\geq 0}, \quad \{H(\nu|\nu_{2t+1})\}_{t\geq 0}, \quad \{H(\nu_{2t+1}|\nu)\}_{t\geq 0}$$

are monotone decreasing.

Proof. The last two displays imply the other claims. Moreover, the two inequalities therein are clear: the data processing inequality (Example 1.7) shows $H(\pi_{2t+2}|\pi_{2t}) \geq H(\mu_{2t+2}|\mu_{2t})$, and of course $H(\pi_{2t}|\pi_{2t+2}) \geq 0$. Hence, it suffices to prove the equalities. On the one hand, Lemma 6.7 yields

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

which is the second equality. On the other hand, Lemma 6.7 yields

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

which is the first equality.

The proof for $H(\mu|\mu_{2t}) \geq H(\nu_{2t+1}|\nu) \geq H(\mu|\mu_{2t+2})$ is analogous, at least for $t \geq 1$. For $t = 0$, while $H(\mu|\mu_0) \geq H(\nu_1|\nu)$ formally follows from the data processing inequality and computing $H(\pi_1|\pi_{-1}) - H(\nu_1|\nu_{-1}) = H(\mu|\mu_0) - H(\nu_1|\nu)$, it does not seem obvious that $H(\nu_1|\nu_{-1}) < \infty$. To circumvent this, consider disintegrations $\pi_1(dx, dy) = \nu_1(dy) \otimes K_1(y, dx)$ and $\pi_{-1}(dx, dy) = \nu_{-1}(dy) \otimes K_{-1}(y, dx)$. Formally, $H(\pi_1|\pi_{-1}) - H(\nu_1|\nu_{-1}) = \int H(K_1|K_{-1})\, d\nu_1$, but the right-hand side is meaningful even if $H(\pi_1|\pi_{-1})$ and $H(\nu_1|\nu_{-1})$ are both infinite. We have $dK_1/d\mu = e^{\varphi_1 \oplus \psi_0 - c}/e^{\psi_0 - \psi_1}$ and $dK_{-1}/d\mu = e^{\varphi_0 \oplus \psi_{-1} - c}/e^{\psi_{-1} - \psi_0}$, yielding

$$\begin{aligned} 0 \leq \int H(K_1|K_{-1})\, d\nu_1 &= \int (\varphi_1 - \varphi_0) \oplus (\psi_1 - \psi_0)\, d\pi_1 \\ = \mu(\varphi_1 - \varphi_0) - \nu_1(\psi_0 - \psi_1) &= H(\mu|\mu_0) - H(\nu_1|\nu) \end{aligned}$$

as desired. □

We recall the following facts about monotone series.

Lemma 6.11. *Let $(a_n)_{n\geq 1} \subset [0, \infty)$ be decreasing and $A := \sum_{n\geq 1} a_n < \infty$. Then*

$$a_n = o(1/n) \qquad \text{and} \qquad a_n \leq A/n, \quad n \geq 1.$$

Proof. We have $na_n \leq \sum_{k=1}^{n} a_k \leq A$ and hence $a_n \leq A/n$. Moreover, given $\varepsilon > 0$, there exists m such that $\sum_{k=m}^{n-1} a_k \leq \varepsilon$ for all $n > m$, thus $(n-m)a_n \leq \varepsilon$ and then $\limsup_{n\to\infty} na_n \leq \varepsilon + \limsup_{n\to\infty} ma_n = \varepsilon$. □

Corollary 6.12 (Sublinear Rate for Marginals)**.** *We have*

$$H(\mu_{2t}|\mu) = o(t), \qquad H(\nu|\nu_{2t-1}) = o(t)$$

and, with $A := H(\pi_*|R) - H(\pi_0|R)$*,*

$$H(\mu_{2t}|\mu) \leq A/t, \qquad H(\nu|\nu_{2t-1}) \leq A/t, \qquad t \geq 1.$$

Analogous results hold for $H(\nu_{2t-1}|\nu)$ *and* $H(\mu|\mu_{2t-2})$*.*

Proof. For $t \geq 0$, recall that $H(\mu_{2t}|\mu) \leq H(\pi_{2t}|\pi_{2t-1})$ by Corollary 6.6 and $H(\nu|\nu_{2t-1}) = H(\pi_{2t}|\pi_{2t-1})$ by Lemma 6.7. As $\sum_{t\geq 1} H(\pi_t|\pi_{t-1}) \leq A$ by Corollary 6.6, we see that

$$\sum_{t\geq 1} H(\mu_{2t}|\mu) \leq A, \qquad \sum_{t\geq 1} H(\nu|\nu_{2t-1}) \leq A.$$

In view of the monotonicity stated in Proposition 6.10, Lemma 6.11 yields the claim. □

Remark 6.13. In the spirit of Proposition 6.10, one can also check that

$$\|\mu - \mu_{2t}\|_{TV}, \qquad \|\nu - \nu_{2t+1}\|_{TV}, \qquad \|\pi_{t+1} - \pi_t\|_{TV}, \qquad t \geq 0$$

are decreasing sequences. See [11, Lemmas 33, 34].

6.3 Strong Convergence

The convergence of Sinkhorn's algorithm can be shown under certain conditions on c. We will see in the proof of Theorem 6.15 below that convergence (of primal and dual iterates, as well as the values) follows once uniform integrability of certain sequences is guaranteed. The simplest case would be to assume that c is bounded, which implies uniform bounds on φ_t, ψ_t (see Lemma 6.14 below). In Theorem 6.15, we use a uniform lower bound on c and an exponential integrability condition; this gives uniform lower bounds on φ_t, ψ_t and a suitable integrability of the positive parts.

Lemma 6.14. *Let* $x \in \mathsf{X}$ *and* $y \in \mathsf{Y}$*.*[21]

(i) For $t \geq 0$*,*

$$\inf_{y\in\mathsf{Y}} \big[c(x,y) - \psi_t(y)\big] \leq \varphi_{t+1}(x) \leq \int c(x,y)\,\nu(dy),$$

$$\inf_{x\in\mathsf{X}} \big[c(x,y) - \varphi_t(x)\big] \leq \psi_t(y) \leq \int c(x,y)\,\mu(dx).$$

[21] For unnormalized c with $\alpha := \int e^{-c}\, d(\mu \otimes \nu) > 1$, the bounds in Lemma 6.14 need to be adjusted by a constant; cf. Remark 6.3.

(ii) If c is bounded, then for $t \geq 0$,

$$-2\|c\|_\infty \leq \varphi_t(x) \leq \|c\|_\infty, \qquad -2\|c\|_\infty \leq \psi_t(y) \leq \|c\|_\infty.$$

In particular, $\|\varphi_t\|_\infty \leq 2\|c\|_\infty$ and $\|\psi_t\|_\infty \leq 2\|c\|_\infty$.

(iii) If $e^{pc} \in L^1(\mu \otimes \nu)$ for some $p > 0$, then for $t \geq 0$,

$$\|e^{p\varphi_{t+1}}\|_{L^1(\mu)} \leq \|e^{pc}\|_{L^1(\mu\otimes\nu)}, \qquad \|e^{p\psi_t}\|_{L^1(\nu)} \leq \|e^{pc}\|_{L^1(\mu\otimes\nu)}.$$

(iv) If $c \geq c_0 \in (-\infty, 0)$, then for $t \geq 1$,

$$\begin{aligned}\varphi_t(x) &\geq c_0 - \log\|e^{\psi_{t-1}}\|_{L^1(\nu)} \geq c_0 - \log\|e^c\|_{L^1(\mu\otimes\nu)},\\ \psi_t(y) &\geq c_0 - \log\|e^{\varphi_t}\|_{L^1(\mu)} \geq c_0 - \log\|e^c\|_{L^1(\mu\otimes\nu)}.\end{aligned}$$

Proof. We only detail the proofs for ψ_t.

(i),(ii) Recall that $\mu(\varphi_t) \geq 0$ and $\nu(\psi_t) \geq 0$ by Lemma 6.4 (iii). Jensen's inequality and $\mu(\varphi_t) \geq 0$ yield

$$\begin{aligned}\psi_t(y) &= -\log \int e^{\varphi_t(x)-c(x,y)}\,\mu(dx)\\ &\leq \int [-\varphi_t(x) + c(x,y)]\,\mu(dx) \leq \int c(x,y)\,\mu(dx) \leq \|c\|_\infty.\end{aligned}$$

Similarly, $\varphi_t(x) \leq \int c(x,y)\,\nu(dy) \leq \|c\|_\infty$, completing the proof of the upper bounds. For the lower bound, we note that

$$\begin{aligned}\psi_t(y) &\geq -\log \int e^{\sup_{x\in\mathsf{X}}[\varphi_t(x)-c(x,y)]}\,\mu(dx)\\ &= -\sup_{x\in\mathsf{X}}\big[\varphi_t(x) - c(x,y)\big] \geq -\|c\|_\infty - \|\varphi_t^+\|_\infty \geq -2\|c\|_\infty,\end{aligned}$$

here the upper bound of φ_t (resp. $\varphi_0 = 0$ for $t = 0$) was used in the last step.

(iii) Using the upper bound in (i) and Jensen's inequality,

$$\int e^{p\psi_t}\,d\nu \leq \int e^{p\int c(x,y)\,\mu(dx)}\nu(dy) \leq \int e^{pc}d(\mu\otimes\nu).$$

(iv) Using the definition of ψ_t and (iii) with $p = 1$,

$$e^{-\psi_t(y)} = \int e^{\varphi_t(x)-c(x,y)}\,\mu(dx) \leq e^{-c_0}\|e^{\varphi_t}\|_{L^1(\mu)} \leq e^{-c_0}\|e^c\|_{L^1(\mu\otimes\nu)}. \qquad \square$$

We can now show the convergence of Sinkhorn's algorithm under an exponential integrability of c.

Theorem 6.15. *Let c be bounded from below and such that*

$$\int e^{rc}\, d(\mu \otimes \nu) < \infty \quad \text{for some} \quad r > 1. \tag{6.8}$$

Then, for some Schrödinger potentials φ_*, ψ_* and the Schrödinger bridge π_*,

(i) $\varphi_t(x) \to \varphi_*(x)$ *and* $\psi_t(y) \to \psi_*(y)$ *for all* $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

(ii) $\varphi_t \to \varphi_*$ *in* $L^p(\mu)$ *and* $\psi_t \to \psi_*$ *in* $L^p(\nu)$ *for all* $p \in [1, \infty)$,

(iii) $H(\pi_*|\pi_t) \to 0$ *and* $\pi_t \to \pi_*$ *in variation,*

(iv) $H(\pi_t|R) \to H(\pi_*|R)$.

Proof. Note that $\mu \otimes \nu \in \Pi_{fin}(\mu, \nu)$, so that π_* exists and $H(\pi_*|R) < \infty$. In view of Lemma 6.14 (iii),(iv) and the la Vallée–Poussin theorem [2, Theorem 4.5.9, p. 272],

$$(e^{\varphi_t})_{t\geq 1}, \quad (\varphi_t)_{t\geq 1}, \quad (e^{\varphi_t - \varphi_{t+1}}\varphi_t)_{t\geq 1} \quad \text{are uniformly integrable in } L^1(\mu)$$

and similarly for ψ_t. For later use, we also recall from Lemma 6.4 that $\mu(\varphi_t)_{t\geq 0}$ is increasing, hence the limit $m := \lim_t \mu(\varphi_t)$ exists.

By the Dunford–Pettis theorem [2, Theorem 4.7.18, p. 285], after passing to a subsequence, the uniformly integrable sequence e^{φ_t} converges weakly in $L^1(\mu)$ to some function Φ; i.e., relative to the topology $\sigma(L^1(\mu), L^\infty(\mu))$. We write $\Phi = e^{\varphi_*}$. As $e^{-c(\cdot,y)} \in L^\infty(\mu)$ due to the lower bound of c, it follows for fixed $y \in \mathsf{Y}$ that

$$\lim_{t\to\infty} \psi_t(y) = -\log \lim_{t\to\infty} \int e^{\varphi_t(x)} e^{-c(x,y)}\, \mu(dx) = -\log \int e^{\varphi_*(x) - c(x,y)}\, \mu(dx).$$

We write $\psi_*(y)$ for the right-hand side, so that $\psi_t \to \psi_*$ pointwise, and by uniform integrability also $e^{\psi_t} \to e^{\psi_*}$ in $L^1(\nu)$. Noting that $(e^{\psi_t(\cdot) - c(x,\cdot)})_{t\geq 0} \subset L^1(\nu)$ is uniformly integrable for every $x \in \mathsf{X}$, we also have

$$\lim_{t\to\infty} \varphi_{t+1}(x) = -\log \lim_{t\to\infty} \int e^{\psi_t(y) - c(x,y)}\, \nu(dy) = -\log \int e^{\psi_*(y) - c(x,y)}\, \nu(dy)$$

and the right-hand side must coincide with φ_* μ-a.s. Choosing a suitable version of φ_*, we have $\varphi_t(x) \to \varphi_*(x)$ for all $x \in \mathsf{X}$.

In brief, (φ_*, ψ_*) is a solution of the Schrödinger equations; cf. Corollary 2.5. In view of the exponential integrability and uniform lower bounds, it is clear that $\varphi_t \to \varphi_*$ in $L^p(\mu)$ for all $p \in [1, \infty)$, and similarly for ψ_t. To see that the original sequence (φ_t) converges to φ_*, we still need to argue that the potential φ_* does not depend on the subsequence chosen above. Indeed, the potential is unique up to a constant (Theorem 2.1) and hence completely determined by its mean. As noted in the beginning of the proof, the limit $m = \lim_t \mu(\varphi_t) = \mu(\varphi_*)$ exists along the original sequence, so that φ_* cannot depend on the subsequence.

As (φ_*, ψ_*) is a solution of the Schrödinger equations, $\pi(\varphi_*, \psi_*) = \pi_*$ is the Schrödinger bridge. Clearly $H(\pi_*|\pi_{2t}) = \mu(\varphi_* - \varphi_t) + \nu(\psi_* - \psi_t) \to 0$, and similarly for π_{2t+1}. By Pinsker's inequality (Lemma 1.2), this also yields $\pi_t \to \pi_*$ in variation. Finally, using Lemma 6.7,

$$\begin{aligned} H(\pi_{2t}|R) - H(\pi_*|R) &= \mu_{2t}(\varphi_t) - \mu(\varphi_*) + \nu(\psi_t) - \nu(\psi_*) \\ &= \mu(e^{\varphi_t - \varphi_{t+1}}\varphi_t) - \mu(\varphi_*) + \nu(\psi_t) - \nu(\psi_*) \to 0 \end{aligned}$$

due to $\varphi_t(x) - \varphi_{t+1}(x) \to 0$ and the uniform integrability of $e^{\varphi_t - \varphi_{t+1}}\varphi_t$. Similarly for π_{2t+1}. □

Remark 6.16. Condition (6.8) can be weakened to

$$\int e^{r\int c(x,y)\,\nu(dy)}\,\mu(dx) < \infty \quad \text{and} \quad \int e^{r\int c(x,y)\,\mu(dx)}\,\nu(dy) < \infty \tag{6.9}$$

for some $r > 1$ (which is weaker due to Jensen's inequality). The proof of Theorem 6.15 remains unchanged except that we now use Lemma 6.14 (i) instead of (iii) to obtain the uniform integrability of (e^{φ_t}) and (e^{ψ_t}).

Remark 6.17. [32] gives a different set of integrability conditions to derive uniform integrability and hence convergence of Sinkhorn's algorithm; see Conditions (A1), (B1), (B2) in [32]. Those conditions are not directly comparable to the ones above—neither includes the other. Instead of exponential integrability, [32] assumes among several other properties that $\int e^{-c(x,y)}\,\mu(dx)$ is uniformly bounded away from zero, which is quite general if c is bounded in the y variable but restrictive otherwise.

On a related note, we remark that (6.8) is still a fairly strong condition: in the context of (εEOT) where c is replaced by c/ε, it requires that c has an exponential moment of order larger than $1/\varepsilon$.

One can certainly think of other sets of conditions to prove a version of Theorem 6.15. However, it is worth noting that the convergence of Algorithm 6.1 is less general than the convergence of the marginals in Corollary 6.6, which did not even require finiteness of c.

Example 6.18 (Divergence of Sinkhorn Iterates)**.** As in Example 2.16, consider $\mathsf{X} = \mathsf{Y} = \{0, 1\}$ with uniform marginals μ, ν while R is the uniform distribution on $\{(0,0),(0,1),(1,1)\}$, thus corresponding to a cost function taking the value $c(1,0)=\infty$. As observed in Example 2.16, $\Pi(\mu,\nu)$ has a unique element absolutely continuous wrt. R, the Schrödinger bridge

$$\pi_* = \frac{1}{2}(\delta_{(0,0)} + \delta_{(1,1)}).$$

We know from Corollary 6.6 that the Sinkhorn marginals (μ_t,ν_t) converge to (μ,ν). From the finiteness of $\mathsf{X}\times\mathsf{Y}$, it is clear that (π_t) admits cluster points. Any cluster point π must have the limit marginals (μ,ν) and satisfy $\pi \ll R$, which already implies $\pi=\pi_*$ by the above. As a result, $\pi_t\to\pi_*$ in variation.

Next, consider the Sinkhorn iterates φ_t,ψ_t. By construction, we have $\varphi_t(x)+\psi_t(y)=\log(\frac{d\pi_{2t}}{dR}(x,y))$ R-a.s., and the fact that $\pi_{2t}\to\pi_*$ in variation means that $\frac{d\pi_{2t}}{dR}\to\frac{d\pi_*}{dR}$ R-a.s. More explicitly,

$$\varphi_t(0)+\psi_t(0)\to\log\frac{3}{2},\quad \varphi_t(0)+\psi_t(1)\to-\infty,\quad \varphi_t(1)+\psi_t(1)\to\log\frac{3}{2}.$$

This implies that both φ_t and ψ_t diverge.

Sharp conditions for the convergence of (π_n) in variation do not seem to be known at present. The next result at least shows that if (π_n) converges in variation, then the limit is indeed the Schrödinger bridge π_*.

Proposition 6.19. *Let $c<\infty$ $\mu\otimes\nu$-a.s., or equivalently $R\sim\mu\otimes\nu$. Suppose that a subsequence (π_{n_k}) of the Sinkhorn iterates of Algorithm 6.2 converges in total variation. Then the limit is the Schrödinger bridge π_*. Moreover, if $n_k=2t_k$, the iterates φ_{t_k},ψ_{t_k} of Algorithm 6.1 satisfy $\varphi_{t_k}\oplus\psi_{t_k}\to\varphi_*\oplus\psi_*$ in $\mu\otimes\nu$-probability, where $\varphi_*\oplus\psi_*$ is the (uniquely determined) sum of the Schrödinger potentials. The analogue holds for $n_k=2t_k-1$.*

Proof. For simplicity, we denote the subsequence by (π_n). Let $\pi_n\to\pi_0$ in variation. As $\pi_n\ll R$ for all n, it follows that $\pi_0\ll R$ and, for $n=2t$,

$$e^{\varphi_t\oplus\psi_t}=\frac{d\pi_n}{dR}\to\frac{d\pi_0}{dR}=:e^F \quad \text{in} \quad L^1(R).$$

Here F is a measurable function with values in $[-\infty,\infty)$. As π_0 is a probability, $\pi_0\{F>-\infty\}=1$ and hence $R\{F>-\infty\}>0$. After passing to a subsequence, we have $\varphi_t\oplus\psi_t\to F$ R-a.s. Recalling $R\sim\mu\otimes\nu$, Corollary 2.12 (b) then states that $F=\varphi\oplus\psi$ for some measurable functions $\varphi:\mathsf{X}\to[-\infty,\infty)$ and $\psi:\mathsf{Y}\to[-\infty,\infty)$, and now Theorem 2.1 (b) shows that $\pi_0 = \pi_*$. Thus, convergence must hold along the original sequence (π_n). More precisely, Corollary 2.12 (b) and Theorem 2.1 (b) imply that

$$e^{\varphi_t \oplus \psi_t} \to e^{\varphi_* \oplus \psi_*} \quad \text{in} \quad L^1(R),$$

where (φ_*, ψ_*) are the Schrödinger potentials (which are finite and unique up to additive constant). This implies $\varphi_t \oplus \psi_t \to \varphi_* \oplus \psi_*$ in $\mu \otimes \nu$-probability. Similarly for ψ_{t-1} instead of ψ_t. □

In the discrete case where X and Y are finite sets, it is clear that (π_n) converges in variation after passing to a subsequence, simply because the weights $(\pi_n(x, y))$ form a bounded set in a finite-dimensional space. For finite cost c, we can then apply Proposition 6.19 to deduce that the whole sequence converges to π_*. (As c is even bounded, this is also clear from Theorem 6.15.) But when X, Y are not discrete, compactness for the total variation topology can be hard to establish. The next section offers an alternate approach using weak convergence. The advantage of the weak topology is that (relative) compactness is immediate.

6.4 Weak Convergence and the Link to Stability

In this section we frame the convergence of Sinkhorn's algorithm as a more general question, the stability of entropic optimal transport problems.

Given the cost c and marginals μ, ν, we have seen that if there exists a coupling with finite entropy wrt. R, there exists a unique Schrödinger bridge $\pi_* = \pi(\mu, \nu) \in \Pi(\mu, \nu)$. Or equivalently, if we do not enforce the normalization for c, a unique solution $\pi_* \in \Pi(\mu, \nu)$ of the entropic optimal transport problem

$$\mathcal{C}_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int c \, d\pi + H(\pi | \mu \otimes \nu). \tag{6.10}$$

Consider sequences of marginals $\mu_n \to \mu$ and $\nu_n \to \nu$, where the convergence is in a sense to be chosen. Assuming suitable integrability, there are unique solutions $\pi_n \in \Pi(\mu_n, \nu_n)$ of[22]

$$\mathcal{C}_1(\mu_n, \nu_n) = \inf_{\pi \in \Pi(\mu_n, \nu_n)} \int c \, d\pi + H(\pi | \mu_n \otimes \nu_n).$$

[22] As $\int e^{-c} \, d(\mu_n \otimes \nu_n)$ typically depends on n, using the formulation of entropic optimal transport avoids introducing an additional constant depending on n, whence our preference over the Schrödinger bridge formulation.

It is then natural to ask whether $\pi_n \to \pi_*$ (in a suitable sense). In words, we expect the solution to be *stable* wrt. the marginals. Similar questions can be asked for the potentials and the values $\mathcal{C}_1$.

Several results on stability have been obtained in the recent literature [7, 12, 23, 16]; we detail and apply the one of [23]. We say that $\pi \in \Pi(\mu, \nu)$ is *c-cyclically invariant* Definition 5.11 holds with $\varepsilon = 1$, or equivalently if Definition 2.6 holds for $dR \propto e^{-c} d(\mu \otimes \nu)$. Under the condition $\mathcal{C}_1(\mu, \nu) < \infty$, we have seen that the solution π_* of (6.10) is c-cyclically invariant and that it is uniquely characterized by that property (Corollary 2.9). In [1, 23], the authors do not impose the condition $\mathcal{C}_1(\mu, \nu) < \infty$ but work directly with cyclical invariance. In that language, the question on stability of optimizers becomes: if $\pi_n \in \Pi(\mu_n, \nu_n)$ are c-cyclically invariant, does convergence of the marginals imply convergence of π_n to a c-cyclically invariant coupling of the limiting marginals?

The following is special case of [23, Theorem 1.4]. The general assumptions are that X, Y are Polish metric spaces and that $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ is continuous and bounded from below. To allow for the techniques used in [23], the spaces X and $\mathsf{X} \times \mathsf{Y}$ are assumed to satisfy the assertion of Lebesgue's theorem on differentiation of measures:

Assumption 6.20. Given $\rho, \lambda \in \mathcal{P}(\mathsf{X})$ satisfying $\rho \ll \lambda$, there exists $\mathsf{X}_0 \subset \mathsf{X}$ of full λ-measure such that

$$f(x) := \lim_{r \to 0} \frac{\rho(B_r(x))}{\lambda(B_r(x))}, \qquad x \in \mathsf{X}_0 \tag{6.11}$$

defines a version of the Radon–Nikodym density $d\rho/d\lambda$. The analogous property is assumed on the space $\mathsf{X} \times \mathsf{Y}$.

This holds in particular when X, Y are Euclidean spaces, the main example of interest. The conclusion then reads as follows.

Theorem 6.21 (Weak Stability)**.** *For $n \geq 1$, let $(\mu_n, \nu_n) \in \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{Y})$ and let $\pi_n \in \Pi(\mu_n, \nu_n)$ be c-cyclically invariant. Suppose that μ_n, ν_n converge weakly to some limits μ, ν. Then π_n converges weakly to the unique c-cyclically invariant coupling $\pi_* \in \Pi(\mu, \nu)$.*

Weak convergence of (π_n), at least along a subsequence, is clear by the compactness stated in Lemma 5.7. The main insight in Theorem 6.21 is that any limit π must be equivalent to $\mu \otimes \nu$ and that one can pass to the limit in the definition of c-cyclical invariance. We refer to [23] for the proof.

Let us now return to Sinkhorn's algorithm. Here we start from marginals μ, ν and the algorithm itself generates certain marginals μ_n, ν_n. These satisfy $\mu_n \sim \mu$ and $\nu_n \sim \nu$ (cf. Lemma 6.7); moreover, $\mu_n \to \mu$ and $\nu_n \to \nu$ in variation (Corollary 6.6), which of course implies weak convergence. Our approach is to see the iterates π_n as solutions to EOT problems with marginals μ_n, ν_n, so that the stability theorem can be turned into a convergence result for the algorithm.

Indeed, writing φ_t, ψ_t for the Sinkhorn iterates as before, we have by construction that

$$\frac{d\pi_{2t}}{d(\mu \otimes \nu)} = e^{\varphi_t \oplus \psi_t - c} \quad \mu \otimes \nu\text{-a.s.}$$

and similarly for π_{2t-1}. This is not quite our standard form with the EOT potentials, because μ and ν are not the marginals of π_{2t}, but it can be translated as follows.

Lemma 6.22. *Let π_t, φ_t, ψ_t be the Sinkhorn iterates of Algorithm 6.1. Then*

$$\frac{d\pi_n}{d(\mu_n \otimes \nu_n)} = e^{\tilde{\varphi}_n \oplus \tilde{\psi}_n - c} \quad \mu_n \otimes \nu_n\text{-}a.s.$$

for $n \geq 1$, where

$$\begin{cases} \tilde{\varphi}_n := \varphi_{t+1}, & \tilde{\psi}_n := \psi_t \qquad \text{if } \ n = 2t, \\ \tilde{\varphi}_n := \varphi_t, & \tilde{\psi}_n := \psi_t \qquad \text{if } \ n = 2t-1. \end{cases}$$

In particular, $\pi_n \in \Pi(\mu_n, \nu_n)$ is c-cyclically invariant for all $n \geq 1$.

Proof. Recalling from Lemma 6.7 the formulas for the marginals densities and using

$$\frac{d\pi_n}{d(\mu_n \otimes \nu_n)} = \frac{d\pi_n}{d(\mu \otimes \nu)} \frac{d(\mu \otimes \nu)}{d(\mu_n \otimes \nu_n)},$$

we see that

$$\frac{d\pi_{2t}}{d(\mu_{2t} \otimes \nu_{2t})} = \frac{d\pi_{2t}}{d(\mu \otimes \nu)} \frac{d\mu}{d\mu_{2t}} = e^{\varphi_t \oplus \psi_t - c} e^{\varphi_{t+1} - \varphi_t} = e^{\varphi_{t+1} \oplus \psi_t - c}$$

and similarly for $2t-1$. The last conclusion follows from Lemma 2.7. □

If $\mathcal{C}(\mu_n, \nu_n) < \infty$, the c-cyclical invariance in Lemma 6.22 means that π_n is the minimizer. Depending on the assumptions on c, it may or may not be clear that the Sinkhorn marginals satisfy $\mathcal{C}(\mu_n, \nu_n) < \infty$. One convenience of working with c-cyclical invariance is that finiteness is not important in the first place.

Theorem 6.23 (Weak Convergence of Sinkhorn)**.** *Let* X, Y *be Euclidean spaces, or more generally Polish metric spaces satisfying Assumption 6.20. Let* c *be continuous and bounded from below, and let* $\mathcal{C}_1(\mu,\nu) < \infty$*. Then the Sinkhorn iterates* (π_n) *converge weakly to the optimizer* $\pi_* \in \Pi(\mu,\nu)$ *of* (6.10)*.*

Proof. As $\mu_n \to \mu$ and $\nu_n \to \nu$ in variation by Corollary 6.6, Theorem 6.21 immediately tells us that $\pi_n \to \pi_0 \in \Pi(\mu,\nu)$ weakly for a c-cyclically invariant coupling π_0, and as $\mathcal{C}_1(\mu,\nu) < \infty$, it follows that π_0 is the optimizer π_* (Corollary 2.9). □

A different application of stability to Sinkhorn's algorithm can be found in [16], where a rate of convergence in Wasserstein metric is provided.

6.5 Linear Convergence for Bounded Cost

In this section we give a quite different analysis of Sinkhorn's algorithm. Instead of using probabilistic properties, we study the dual problem as a concave maximization and the algorithm as a coordinate ascent. One purpose of this section is to highlight how strong convexity (a lower bound on the second derivative) leads to linear convergence. This is particularly clear in the analysis of [5] which we follow in this section. Other proofs of linear convergence for bounded cost have been given through the Hilbert–Birkhoff projective metric (see [8, 20]); we do not cover that approach here.

Let c be bounded and measurable (we do not assume $\int e^{-c} d(\mu\otimes\nu) = 1$). For $\varphi \in L^1(\mu)$, $\psi \in L^1(\nu)$, consider the objective function of the dual EOT problem,

$$G(\varphi,\psi) := \mu(\varphi) + \nu(\psi) - \int e^{\varphi\oplus\psi - c}\, d(\mu\otimes\nu) + 1. \tag{6.12}$$

The following algorithm deviates slightly from the one considered above—it centers the first potential—hence we distinguish the notation.

Algorithm 6.24 (Sinkhorn with Centering)**.** Set $\bar\varphi_0 := 0$. For $t \geq 0$,

$$\bar\psi_t(y) := -\log \int_{\mathsf{X}} e^{\bar\varphi_t(x) - c(x,y)}\, \mu(dx), \tag{6.13}$$

$$\bar\varphi_{t+1}(x) := -\log \int_{\mathsf{Y}} e^{\bar\psi_t(y) - c(x,y)}\, \nu(dy) + \lambda_t, \qquad \text{where} \tag{6.14}$$

$$\lambda_t := \int_X \log\left(\int_Y e^{\bar\psi_t(y) - c(x,y)}\, \nu(dy) \right) \mu(dx). \tag{6.15}$$

Comparing with Algorithm 6.1, $\bar{\psi}_t$ is updated like ψ_t, but $\bar{\varphi}_{t+1}$ is centered: λ_t is chosen such that $\mu(\bar{\varphi}_{t+1}) = 0$. While Algorithm 6.1 corresponds to an unconstrained coordinate ascent, Algorithm 6.24 can be expressed as

$$\bar{\psi}_t(y) = \operatorname*{arg\,max}_{\psi \in L^1(\nu)} G(\bar{\varphi}_t, \psi), \qquad \bar{\varphi}_{t+1}(x) = \operatorname*{arg\,max}_{\varphi \in L^1(\mu) \colon \mu(\varphi)=0} G(\varphi, \bar{\psi}_t).$$

This is again a coordinate ascent, but $\bar{\varphi}_{t+1}$ is chosen in a smaller space given by the centering constraint. These iterates are related to the ones in Algorithm 6.1 as follows.

Lemma 6.25. *Let (φ_t, ψ_t) be the usual Sinkhorn iterates as defined in Algorithm 6.1. Then $\mu(\varphi_t) = -(\lambda_0 + \cdots + \lambda_{t-1})$ and*

$$\bar{\varphi}_t = \varphi_t - \mu(\varphi_t), \qquad \bar{\psi}_t = \psi_t + \mu(\varphi_t)$$

for all $t \geq 0$. In particular, $\bar{\varphi}_t \oplus \bar{\psi}_t = \varphi_t \oplus \psi_t$ and $G(\bar{\varphi}_t, \bar{\psi}_t) = G(\varphi_t, \psi_t)$.

Proof. This readily follows by induction. □

As $\bar{\psi}_t$ is defined through the Schrödinger equation,

$$d\pi(\bar{\varphi}_t, \bar{\psi}_t) = e^{\bar{\varphi}_t \oplus \bar{\psi}_t - c} \, d(\mu \otimes \nu) \quad \text{has second marginal } \nu, \tag{6.16}$$

as in (6.5). Or we can argue through Lemma 6.25: $\pi(\bar{\varphi}_t, \bar{\psi}_t) = \pi_{2t}(\varphi_t, \psi_t)$ is as before, hence still has second marginal ν. By contrast, the measure $\pi(\bar{\varphi}_{t+1}, \bar{\psi}_t) = e^{\bar{\varphi}_{t+1} \oplus \bar{\psi}_t - c} \, d(\mu \otimes \nu)$ does not have first marginal μ in general, due to the centering constraint. It is not a probability (unless $\lambda_t = 0$), and we shall not use this measure below.

The main advantage of the centering is that it allows us the separate the two coordinates as follows: for $\varphi \in L^2(\mu)$ and $\psi \in L^2(\nu)$,

$$\|\varphi \oplus \psi\|^2_{L^2(\mu \otimes \nu)} = \|\varphi\|^2_{L^2(\mu)} + \|\psi\|^2_{L^2(\nu)} \qquad \text{if} \quad \mu(\varphi) = 0. \tag{6.17}$$

Next, we check that the modified iterates are still bounded when c is bounded (this could also be inferred from Lemma 6.14 via Lemma 6.25).

Lemma 6.26. *For every $t \geq 0$, we have*

$$\|\bar{\varphi}_t\|_\infty \leq 2\|c\|_\infty, \quad \|\bar{\psi}_t\|_\infty \leq 3\|c\|_\infty.$$

Proof. Using the definition (6.14) of $\bar{\varphi}_t$ and writing $\bar{\psi} := \bar{\psi}_{t-1}$, we find that for all $x_1, x_2 \in \mathsf{X}$,

$$\begin{aligned}
&\bar{\varphi}_t(x_1) - \bar{\varphi}_t(x_2)\\
&= \log \int e^{\bar{\psi}(y)-c(x_2,y)}\,\nu(dy) - \log \int e^{\bar{\psi}(y)-c(x_1,y)}\,\nu(dy)\\
&= \log \int e^{c(x_1,y)-c(x_2,y)+\bar{\psi}(y)-c(x_1,y)}\,\nu(dy) - \log \int e^{\bar{\psi}(y)-c(x_1,y)}\,\nu(dy)\\
&\leq \log \left[e^{\sup_{y\in\mathsf{Y}}|c(x_1,y)-c(x_2,y)|} \int e^{\bar{\psi}(y)-c(x_1,y)}\,\nu(dy)\right] - \log \int e^{\bar{\psi}(y)-c(x_1,y)}\,\nu(dy)\\
&= \sup_{y\in\mathsf{Y}} |c(x_1,y)-c(x_2,y)| \leq 2\|c\|_\infty.
\end{aligned}$$

(This was the same calculation as in the proof of Lemma 4.11, and no particular property of $\bar{\psi}$ was used.) As $\mu(\bar{\varphi}_t) = 0$, we must have $\sup_x \bar{\varphi}_t(x) \geq 0$ and $\inf_x \bar{\varphi}_t(x) \leq 0$, hence the above implies $\|\bar{\varphi}_t\|_\infty \leq 2\|c\|_\infty$. The definition (6.13) of $\bar{\psi}_t$ now directly yields $\|\bar{\psi}_t\|_\infty \leq \|\bar{\varphi}_t\|_\infty + \|c\|_\infty \leq 3\|c\|_\infty$. □

The main result of this section reads as follows.

Theorem 6.27. *Let c be bounded and let $(\bar{\varphi}_*, \bar{\psi}_*)$ be the unique EOT potentials with $\mu(\bar{\varphi}_*) = 0$. The iterates $(\bar{\varphi}_t, \bar{\psi}_t)_{t\geq 0}$ of Algorithm 6.24 satisfy*

$$G(\bar{\varphi}_*, \bar{\psi}_*) - G(\bar{\varphi}_t, \bar{\psi}_t) \leq \beta^t \big(G(\bar{\varphi}_*, \bar{\psi}_*) - G(\bar{\varphi}_0, \bar{\psi}_0)\big), \tag{6.18}$$

$$\|\bar{\varphi}_* - \bar{\varphi}_t\|^2_{L^2(\mu)} + \|\bar{\psi}_* - \bar{\psi}_t\|^2_{L^2(\nu)} \leq \beta_0 \beta^t \big(G(\bar{\varphi}_*, \bar{\psi}_*) - G(\bar{\varphi}_0, \bar{\psi}_0)\big), \tag{6.19}$$

where $\beta := 1 - e^{-24\|c\|_\infty} \in (0,1)$ and $\beta_0 := 2e^{6\|c\|_\infty}$.

Theorem 6.27 carries over to the uncentered Sinkhorn algorithm.

Corollary 6.28. *Let c be bounded and $(\varphi_t, \psi_t)_{t\geq 0}$ the iterates of Algorithm 6.1. Let (φ_*, ψ_*) be the unique EOT potentials with $\mu(\varphi_*) = \lim_t \mu(\varphi_t)$. Then*

$$G(\varphi_*, \psi_*) - G(\varphi_t, \psi_t) \leq \beta^t \big(G(\varphi_*, \psi_*) - G(\varphi_0, \psi_0)\big), \tag{6.20}$$

$$\|\varphi_* - \varphi_t\|^2_{L^2(\mu)} + \|\psi_* - \psi_t\|^2_{L^2(\nu)} \leq \beta_0 \beta^t \big(G(\varphi_*, \psi_*) - G(\varphi_0, \psi_0)\big), \tag{6.21}$$

where $\beta := 1 - e^{-24\|c\|_\infty} \in (0,1)$ and $\beta_0 := 2e^{6\|c\|_\infty}$.

Proof. As $G(\bar{\varphi}_t, \bar{\psi}_t) = G(\varphi_t, \psi_t)$ by Lemma 6.25, the convergence (6.20) is immediate from (6.18). Let $\alpha_t = \nu(\varphi_* - \varphi_t) \geq 0$ and $\beta_t = \nu(\psi_* - \psi_t) \geq 0$, where the sign is due to Lemma 6.4 (iii). Hence, $\Psi_t := \psi_* - \psi_t - \beta_t$ is centered, and we recall that $\bar{\varphi}_* - \bar{\varphi}_t$ is centered as well. Using Lemma 6.25 as well as (6.17) with a centered random variable and a constant,

$$\begin{aligned}
\|\varphi_* - \varphi_t\|^2_{L^2(\mu)} + \|\psi_* - \psi_t\|^2_{L^2(\nu)} &= \|\bar{\varphi}_* - \bar{\varphi}_t + \alpha_t\|^2_{L^2(\mu)} + \|\Psi_t + \beta_t\|^2_{L^2(\nu)} \\
&= \|\bar{\varphi}_* - \bar{\varphi}_t\|^2_{L^2(\mu)} + \alpha_t^2 + \|\Psi_t\|^2_{L^2(\nu)} + \beta_t^2 \\
&\leq \|\bar{\varphi}_* - \bar{\varphi}_t\|^2_{L^2(\mu)} + \|\Psi_t\|^2_{L^2(\nu)} + (\alpha_t + \beta_t)^2 \\
&= \|\bar{\varphi}_* - \bar{\varphi}_t\|^2_{L^2(\mu)} + \|\Psi_t + \alpha_t + \beta_t\|^2_{L^2(\nu)} \\
&= \|\bar{\varphi}_* - \bar{\varphi}_t\|^2_{L^2(\mu)} + \|\bar{\psi}_* - \bar{\psi}_t\|^2_{L^2(\nu)}.
\end{aligned}$$

Therefore, (6.21) follows from (6.19) □

Remark 6.29. As the iterates are uniformly bounded by Lemma 6.26, the linear convergences (6.19), (6.21) in L^2 already imply the corresponding linear convergences in L^p for any $p \in [1, \infty)$.

While the convergence in Theorem 6.27 and Corollary 6.28 is linear, one can observe that the constants β, β_0 may be very close to one and very large, respectively, especially in the context of ($\varepsilon\mathsf{EOT}$) where c is replaced by c/ε. There are other techniques to show linear convergence, but they seem to share the issue of yielding poor constants, contrasting with the fast convergence of Sinkhorn's algorithm typically seen in computational practice.

6.5.1 Proof of Theorem 6.27

The basic idea is to use the strong convexity of the exponential function on an interval $[-\alpha, \infty)$,

$$e^b - e^a \geq (b-a)e^a + \frac{e^{-\alpha}}{2}|b-a|^2 \quad \text{for} \quad a, b \in [-\alpha, \infty). \tag{6.22}$$

Lemma 6.30. *Consider $\varphi, \varphi' \in L^2(\mu)$ and $\psi, \psi' \in L^2(\nu)$, and define*

$$\begin{aligned}
\partial_1 G(\varphi', \psi')(x) &= 1 - \int_{\mathsf{Y}} e^{\varphi'(x)+\psi'(y)-c(x,y)}\,\nu(dy), \\
\partial_2 G(\varphi', \psi')(y) &= 1 - \int_{\mathsf{X}} e^{\varphi'(x)+\psi'(y)-c(x,y)}\,\mu(dx).
\end{aligned}$$

If $\varphi \oplus \psi - c \geq -\alpha$ and $\varphi' \oplus \psi' - c \geq -\alpha$ for some $\alpha \in \mathbb{R}$, then

$$\begin{aligned}
G(\varphi', \psi') - G(\varphi, \psi) \geq& \int_{\mathsf{X}} \partial_1 G(\varphi', \psi')(x)\,[\varphi'(x) - \varphi(x)]\,\mu(dx) \\
&+ \int_{\mathsf{Y}} \partial_2 G(\varphi', \psi')(y)\,[\psi'(y) - \psi(y)]\,\nu(dy) \\
&+ \frac{e^{-\alpha}}{2}\|(\varphi - \varphi') \oplus (\psi - \psi')\|^2_{L^2(\mu\otimes\nu)}.
\end{aligned}$$

Proof. We use (6.22) to obtain the inequality in

$$\begin{aligned}
&G(\varphi',\psi') - G(\varphi,\psi)\\
&= \mu(\varphi'-\varphi) + \nu(\psi'-\psi) + \int e^{\varphi\oplus\psi-c} - e^{\varphi'\oplus\psi'-c}\, d(\mu\otimes\nu)\\
&\geq \mu(\varphi'-\varphi) + \nu(\psi'-\psi) + \int (\varphi\oplus\psi-\varphi'\oplus\psi')e^{\varphi'\oplus\psi'-c}\, d(\mu\otimes\nu)\\
&\quad + \frac{e^{-\alpha}}{2}\int |\varphi\oplus\psi-\varphi'\oplus\psi'|^2\, d(\mu\otimes\nu)\\
&= \int_{\mathsf{X}} \partial_1 G(\varphi',\psi')(x)\,[\varphi'(x)-\varphi(x)]\,\mu(dx)\\
&\quad + \int_{\mathsf{Y}} \partial_2 G(\varphi',\psi')(y)\,[\psi'(y)-\psi(y)]\,\nu(dy)\\
&\quad + \frac{e^{-\alpha}}{2}\|(\varphi-\varphi')\oplus(\psi-\psi')\|^2_{L^2(\mu\otimes\nu)}.
\end{aligned}$$

□

Lemma 6.31. *With $\sigma := e^{-6\|c\|_\infty}$, we have*

$$G(\bar\varphi_{t+1},\bar\psi_{t+1}) - G(\bar\varphi_t,\bar\psi_t) \geq \frac{\sigma}{2}\left(\|\bar\varphi_{t+1}-\bar\varphi_t\|^2_{L^2(\mu)} + \|\bar\psi_{t+1}-\bar\psi_t\|^2_{L^2(\nu)}\right).$$

Proof. We write the left-hand side as

$$\big(G(\bar\varphi_{t+1},\bar\psi_{t+1}) - G(\bar\varphi_{t+1},\bar\psi_t)\big) \quad + \quad \big(G(\bar\varphi_{t+1},\bar\psi_t) - G(\bar\varphi_t,\bar\psi_t)\big)$$

and estimate separately these two steps of the algorithm. For the first part, Lemma 6.30 with $\alpha = 6\|c\|_\infty$ yields

$$\begin{aligned}
&G(\bar\varphi_{t+1},\bar\psi_{t+1}) - G(\bar\varphi_{t+1},\bar\psi_t)\\
&\geq \int_{\mathsf{Y}} \partial_2 G(\bar\varphi_{t+1},\bar\psi_{t+1})(y)\,[\bar\psi_{t+1}(y)-\bar\psi_t(y)]\,\nu(dy) + \frac{\sigma}{2}\|\bar\psi_t-\bar\psi_{t+1}\|^2_{L^2(\nu)}.
\end{aligned}$$

Here the integral vanishes as the second marginal of π_{2t+2} is ν; cf. (6.16):

$$\begin{aligned}
\partial_2 G(\bar\varphi_{t+1},\bar\psi_{t+1})(y)\,\nu(dy) &= \nu(dy) - \int_{\mathsf{X}} e^{\bar\varphi_{t+1}(x)+\bar\psi_{t+1}(y)-c(x,y)}\,\mu(dx)\nu(dy)\\
&= \nu(dy) - \int_{\mathsf{X}} \pi_{2t+2}(dx,dy) = \nu(dy) - \nu(dy) = 0. \qquad (6.23)
\end{aligned}$$

For the second part, Lemma 6.30 yields

$$\begin{aligned}
&G(\bar\varphi_{t+1},\bar\psi_t) - G(\bar\varphi_t,\bar\psi_t)\\
&\geq \int_{\mathsf{X}} \partial_1 G(\bar\varphi_{t+1},\bar\psi_t)(x)\,[\bar\varphi_{t+1}(x)-\bar\varphi_t(x)]\,\mu(dx) + \frac{\sigma}{2}\|\bar\varphi_t-\bar\varphi_{t+1}\|^2_{L^2(\mu)}.
\end{aligned}$$

Here the integral vanishes for a different reason: the definition (6.14) of $\bar\varphi_{t+1}$ states that $\int_Y e^{\bar\psi_t(y)-c(x,y)}\,\nu(dy) = e^{-\bar\varphi_{t+1}(x)+\lambda_t}$; thus

$$\partial_1 G(\bar\varphi_{t+1},\bar\psi_t)(x) = 1 - e^{\bar\varphi_{t+1}(x)} \int_{\mathsf{Y}} e^{\bar\psi_t(y)-c(x,y)}\,\nu(dy) = 1 - e^{\lambda_t}$$

is deterministic and the centering $\mu(\bar\varphi_{t+1}) = \mu(\bar\varphi_t) = 0$ implies

$$\begin{aligned}\int_{\mathsf{X}} \partial_1 G(\bar\varphi_{t+1},\bar\psi_t)(x)\,[\bar\varphi_{t+1}(x) - \bar\varphi_t(x)]\,\mu(dx)& \\ = (1-e^{\lambda_t}) \int_{\mathsf{X}} [\bar\varphi_{t+1}(x) - \bar\varphi_t(x)]\,\mu(dx) = 0.& \qquad (6.24)\end{aligned}$$

Combining the estimates for the two parts completes the proof. □

Proof of Theorem 6.27. Recall the bounds for $\bar\varphi_t, \bar\psi_t$ from Lemma 6.26 and note that $\bar\varphi_*, \bar\psi_*$ satisfy the same bounds (either by following the proof of Lemma 6.26 or by an application of Lemma 6.14). We can then apply Lemma 6.30 with $\alpha = 6\|c\|_\infty$ to obtain

$$\begin{aligned}G(\bar\varphi_t,\bar\psi_t) - G(\bar\varphi_*,\bar\psi_*) \geq & \int_{\mathsf{X}} \partial_1 G(\bar\varphi_t,\bar\psi_t)(x)\,[\bar\varphi_t(x) - \bar\varphi_*(x)]\,\mu(dx) \\ &+ \int_{\mathsf{Y}} \partial_2 G(\bar\varphi_t,\bar\psi_t)(y)\,[\bar\psi_t(y) - \bar\psi_*(y)]\,\nu(dy) \\ &+ \frac{\sigma}{2}\left(\|\bar\varphi_t - \bar\varphi_*\|^2_{L^2(\mu)} + \|\bar\psi_t - \bar\psi_*\|^2_{L^2(\nu)}\right), \qquad (6.25)\end{aligned}$$

where $\sigma := e^{-\alpha} = e^{-6\|c\|_\infty}$ and (6.17) was used in the last line. For the second integral, we have

$$\int_{\mathsf{Y}} \partial_2 G(\bar\varphi_t,\bar\psi_t)(y)\,[\bar\psi_t(y) - \bar\psi_*(y)]\,\nu(dy) = 0 \qquad (6.26)$$

as in (6.23). To estimate the first integral, we first note that as in (6.24),

$$\int_{\mathsf{X}} \partial_1 G(\bar\varphi_{t+1},\bar\psi_t)(x)\,[\bar\varphi_t(x) - \bar\varphi_*(x)]\,\mu(dx) = 0.$$

Hence

$$\begin{aligned}&\int_{\mathsf{X}} \partial_1 G(\bar\varphi_t,\bar\psi_t)(x)\,[\bar\varphi_t(x) - \bar\varphi_*(x)]\,\mu(dx) \\ &= \int_{\mathsf{X}} [\partial_1 G(\bar\varphi_t,\bar\psi_t)(x) - \partial_1 G(\bar\varphi_{t+1},\bar\psi_t)(x)]\,[\bar\varphi_t(x) - \bar\varphi_*(x)]\,\mu(dx) \\ &\geq -\frac{1}{2\sigma}\|\partial_1 G(\bar\varphi_t,\bar\psi_t) - \partial_1 G(\bar\varphi_{t+1},\bar\psi_t)\|^2_{L^2(\mu)} - \frac{\sigma}{2}\|\bar\varphi_t - \bar\varphi_*\|^2_{L^2(\mu)} \qquad (6.27)\end{aligned}$$

where the inequality follows from Hölder's and Young's inequality: apply

$$\int gh\,d\mu \geq -\frac{1}{\sigma^{1/2}}\|g\|_{L^2(\mu)}\sigma^{1/2}\|h\|_{L^2(\mu)} \geq -\frac{1}{2\sigma}\|g\|^2_{L^2(\mu)} - \frac{\sigma}{2}\|h\|^2_{L^2(\mu)}$$

to $g(x) := \partial_1 G(\bar\varphi_t, \bar\psi_t)(x) - \partial_1 G(\bar\varphi_{t+1}, \bar\psi_t)(x)$ and $h(x) := \bar\varphi_t(x) - \bar\varphi_*(x)$. We use (6.27) and (6.26) in (6.25) to find

$$G(\bar\varphi_*, \bar\psi_*) - G(\bar\varphi_t, \bar\psi_t) \leq \frac{1}{2\sigma}\|\partial_1 G(\bar\varphi_t, \bar\psi_t) - \partial_1 G(\bar\varphi_{t+1}, \bar\psi_t)\|^2_{L^2(\mu)}. \tag{6.28}$$

Suppressing the argument for brevity,

$$\begin{aligned}
|\partial_1 G(\bar\varphi_t, \bar\psi_t)(x) - \partial_1 G(\bar\varphi_{t+1}, \bar\psi_t)(x)| &\leq \int_{\mathsf{Y}} \left| e^{\bar\varphi_{t+1}\oplus\bar\psi_t - c} - e^{\bar\varphi_t\oplus\bar\psi_t - c} \right| \nu(dy) \\
&\leq e^{6\|c\|_\infty} \int_{\mathsf{Y}} |\bar\varphi_{t+1}\oplus\bar\psi_t - \bar\varphi_t\oplus\bar\psi_t|\,\nu(dy) \\
&= \frac{1}{\sigma}|\bar\varphi_{t+1}(x) - \bar\varphi_t(x)|
\end{aligned}$$

where the second inequality used Lemma 6.26 and the Lipschitz continuity of the exponential: $|e^b - e^a| \leq e^M|b-a|$ for $a, b \leq M$. As a result,

$$\|\partial_1 G(\bar\varphi_t, \bar\psi_t) - \partial_1 G(\bar\varphi_{t+1}, \bar\psi_t)\|^2_{L^2(\mu)} \leq \frac{1}{\sigma^2}\|\bar\varphi_{t+1} - \bar\varphi_t\|^2_{L^2(\mu)}.$$

In view of (6.28), we conclude that

$$G(\bar\varphi_*, \bar\psi_*) - G(\bar\varphi_t, \bar\psi_t) \leq \frac{1}{2\sigma^3}\|\bar\varphi_{t+1} - \bar\varphi_t\|^2_{L^2(\mu)}.$$

Now using Lemma 6.31 on the right-hand side yields

$$G(\bar\varphi_*, \bar\psi_*) - G(\bar\varphi_t, \bar\psi_t) \leq \frac{1}{\sigma^4}\left(G(\bar\varphi_{t+1}, \bar\psi_{t+1}) - G(\bar\varphi_t, \bar\psi_t)\right).$$

Writing $\Delta_t = G(\bar\varphi_*, \bar\psi_*) - G(\bar\varphi_t, \bar\psi_t)$, this can be expressed as

$$\Delta_t \leq \frac{1}{\sigma^4}(\Delta_t - \Delta_{t+1})$$

or $\Delta_{t+1} \leq (1-\sigma^4)\Delta_t \leq \cdots \leq (1-\sigma^4)^{t+1}\Delta_0$, which was the first claim of the theorem.

As $(\bar\varphi_*, \bar\psi_*)$ solve the Schrödinger equations, we can follow the proof of Lemma 6.31 to obtain

$$G(\bar\varphi_*, \bar\psi_*) - G(\bar\varphi_t, \bar\psi_t) \geq \frac{\sigma}{2}\left(\|\bar\varphi_* - \bar\varphi_t\|^2_{L^2(\mu)} + \|\bar\psi_* - \bar\psi_t\|^2_{L^2(\nu)}\right)$$

and thus $\|\bar\varphi_* - \bar\varphi_t\|^2_{L^2(\mu)} + \|\bar\psi_* - \bar\psi_t\|^2_{L^2(\nu)} \leq \frac{2}{\sigma}\Delta_t$. The second claim follows. □

References

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

[2] V. I. Bogachev. *Measure theory. Vol. I.* Springer-Verlag, Berlin, 2007.

[3] J. M. Borwein and A. S. Lewis. Decomposition of multivariate functions. *Canad. J. Math.*, 44(3):463–482, 1992.

[4] J. M. Borwein, A. S. Lewis, and R. D. Nussbaum. Entropy minimization, *DAD* problems, and doubly stochastic kernels. *J. Funct. Anal.*, 123(2):264–307, 1994.

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

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

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

[8] 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.

[9] 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.

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

[11] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Preprint arXiv:2106.01357v4*, 2021.

[12] G. Deligiannidis, V. De Bortoli, and A. Doucet. Quantitative uniform stability of the iterative proportional fitting procedure. *Preprint arXiv:2108.08129v2*, 2021.

[13] 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.

[14] S. Di Marino and J. Louet. The entropic regularization of the Monge problem on the real line. *SIAM J. Math. Anal.*, 50(4):3451–3477, 2018.

[15] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math.*, 29(4):389–461, 1976.

[16] 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.

[17] 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.

[18] H. Föllmer and N. Gantert. Entropy minimization and Schrödinger processes in infinite dimensions. *Ann. Probab.*, 25(2):901–926, 1997.

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

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

[21] 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.

[22] A. Gerolin, A. Kausamo, and T. Rajala. Multi-marginal entropy-transport with repulsive cost. *Calc. Var. Partial Differential Equations*, 59(3):Paper No. 90, 20, 2020.

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

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] W. Rudin. *Functional Analysis.* McGraw-Hill, New York, 2nd edition, 1991.

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

[33] L. Rüschendorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.*, 17(5):369–375, 1993.

[34] L. Rüschendorf and W. Thomsen. Closedness of sum spaces and the generalized Schrödinger problem. *Teor. Veroyatnost. i Primenen.*, 42(3):576–590, 1997.

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