# Entropic Optimal Transport: Geometry and Large Deviations[*]

Espen Bernton[†] Promit Ghosal[‡] Marcel Nutz[§]

January 23, 2022

**Abstract**

We study the convergence of entropically regularized optimal transport to optimal transport. The main result is concerned with the convergence of the associated optimizers and takes the form of a large deviations principle quantifying the local exponential convergence rate as the regularization parameter vanishes. The exact rate function is determined in a general setting and linked to the Kantorovich potential of optimal transport. Our arguments are based on the geometry of the optimizers and inspired by the use of $c$-cyclical monotonicity in classical transport theory. The results can also be phrased in terms of Schrödinger bridges.

*Keywords* Optimal Transport; Entropic Penalization; Schrödinger Bridge; Large Deviations; Cyclical Invariance

*AMS 2010 Subject Classification* 90C25; 60F10; 49N05

## 1 Introduction

Over the last three decades, optimal transport theory has flourished due to its connections with geometry, analysis, probability theory, and other fields in mathematics; see for instance [53, 54, 58]. Following computational advances which have enabled high-dimensional applications, a renewed interest comes from applied fields such as machine learning, image processing and statistics. Popularized in this area by Cuturi [21], entropic regularization is

[*] The authors are grateful to Julio Backhoff-Veraguas, Guillaume Carlier, Wilfried Gangbo and Jon Niles-Weed for insightful discussions that greatly contributed to this research.

[†] Department of Statistics, Columbia University, eb3311@columbia.edu.

[‡] Department of Mathematics, Massachusetts Institute of Technology, promit@mit.edu.

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

a key computational approach for high-dimensional problems. The resulting *entropic optimal transport* problem provides an approximate optimal transport when solved for small regularization parameter $\varepsilon > 0$ while admitting much more efficient algorithms than the unregularized problem, in addition to having other desirable properties. We defer the discussion of related literature to Section 1.1 below and proceed with a synopsis of the present study.

Given a continuous cost function $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ on Polish probability spaces $(\mathsf{X}, \mu)$ and $(\mathsf{Y}, \nu)$, we consider the entropic optimal transport problem

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

where $\Pi(\mu, \nu)$ is the set of couplings and $H(\cdot|\mu \otimes \nu)$ denotes relative entropy (or Kullback–Leibler divergence) with respect to the product of the marginals; see Section 2 for the formal definitions. The constant $\varepsilon > 0$ acts as a regularization parameter; $\varepsilon = 0$ recovers the (unregularized) optimal transport problem. Under mild conditions detailed in Sections 2 and 3, respectively, the entropic optimal transport problem admits a unique solution $\pi_\varepsilon \in \Pi(\mu, \nu)$ and $\pi_\varepsilon$ converges weakly to a solution $\pi_*$ of the unregularized problem. Our main interest is to quantify the *speed* of this convergence $\pi_\varepsilon \to \pi_*$.

For finite-dimensional linear programs—including optimal transport problems with marginals supported by finite sets—the solution of the entropic regularization is known to converge exponentially fast to a solution of the original problem (in total variation, say). In transport problems with continuous marginals, the situation is quite different even in the most regular examples. For Gaussian marginals on $\mathbb{R}$ and quadratic costs $c(x, y) = |x - y|^2$, direct computation shows that $\pi_\varepsilon$ is Gaussian and $\pi_*$ is given by a linear transport (Monge) map $T$. One finds that the transport cost converges only linearly, $\int c\, d\pi_\varepsilon - \int c\, d\pi_* = \varepsilon/2 + o(\varepsilon)$. The culprit for this slowdown is easily spotted by inspecting the closed-form solution: the leading term in the cost difference stems from the mass $\pi_\varepsilon$ places at a distance of approximately $\sqrt{\varepsilon}$ to the support $\Gamma$ of $\pi_*$ (that is, the graph of $T$). See Section 1.1 for further discussion on the asymptotics of transport costs and value functions, which have been the main focus of the extant literature on the convergence as $\varepsilon \to 0$.

In the present study, we adopt a different, more *local* perspective, from which the Gaussian example is actually encouraging: the density of $\pi_\varepsilon$ decays *exponentially* away from $\Gamma$. Indeed, it is proportional to $e^{-\alpha|y-T(x)|^2/\varepsilon}$, where $\alpha > 0$ is the quotient of the marginal variances.

The main result of this paper is a comparable statement in a remarkably general setting; it takes the form of a large deviations principle. We define a function $I(x, y)$ through the following optimization. In addition to the given point $(x, y) =: (x_1, y_1)$, choose finitely many points $(x_2, y_2), \dots, (x_k, y_k)$ from the support $\Gamma$ of the limiting optimal transport $\pi_*$, as well as a permutation $\sigma \in \Sigma(k)$. Then, consider the difference

$$\sum_{i=1}^{k} c(x_i, y_i) - \sum_{i=1}^{k} c(x_i, y_{\sigma(i)}) \tag{1.2}$$

between the pointwise transport costs from $x_i$ to $y_i$ with the costs for the permuted destinations $y_{\sigma(i)}$. The optimization is to maximize this difference, and we define $I(x, y)$ as the supremum value of (1.2) over all choices of points and permutations. For $(x, y) \in \Gamma$, the optimality of $\pi_*$ implies that $I(x, y) = 0$, because $\Gamma$ is $c$-cyclically monotone. But outside $\Gamma$, we may typically expect that $I(x, y) > 0$. Part (a) of our theorem below, the large deviations upper bound, shows that $I$ is a lower bound for the rate function in the general Polish setting. The matching bound (b) necessitates a condition on the optimal transport problem that is being approximated—but still holds for the majority of continuous or semi-discrete transport problems of interest. We mainly discuss the uniqueness of Kantorovich potentials (Assumption 4.4) as a sufficient condition; it also gives rise to an insightful representation of $I$ as $I(x, y) = c(x, y) - \psi^c(y) + \psi(x)$, the difference between the cost $c(x, y)$ and the solution of the dual optimal transport problem (see Proposition 4.5). An alternative condition imposing regularity of the optimal transport (Assumption 4.9) is also considered. Tacitly assuming the existence of $\pi_\varepsilon$ and its weak limit (cf. Sections 2 and 3), the main result reads as follows.

**Theorem 1.1.** *Let $\Gamma = \operatorname{spt} \pi_*$ where $\pi_* = \lim_{\varepsilon \to 0} \pi_\varepsilon$ is the limiting optimal transport and define $I : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$ by (4.3).*

*(a) For any compact set $C \subset \mathsf{X} \times \mathsf{Y}$,*

$$\limsup_{\varepsilon \to 0} \varepsilon \log \pi_\varepsilon(C) \leq - \inf_{(x,y) \in C} I(x, y).$$

*(b) Let Assumption 4.4 or Assumption 4.9 hold, and consider the sets $\mathsf{X}_0 = \operatorname{proj}_{\mathsf{X}} \Gamma$ and $\mathsf{Y}_0 = \operatorname{proj}_{\mathsf{Y}} \Gamma$ of full marginal measure. For any open set $U \subset \mathsf{X}_0 \times \mathsf{Y}_0$,*

$$\liminf_{\varepsilon \to 0} \varepsilon \log \pi_\varepsilon(U) \geq - \inf_{(x,y) \in U} I(x, y).$$

The theorem shows in particular that the rate depends (only) on the geometry of $\pi_*$, which does not seem to be clear a priori. We mention that our result can also be stated in terms of (static) Schrödinger bridges. In this context, it is a large deviations principle for the small-noise (or small-time) limit; cf. Section 1.1.

For finitely supported marginals, the density of $\pi_\varepsilon$ converges exponentially for any cost function; that is, the rate function is strictly positive outside $\Gamma$. We shall see that the analogue may fail in the continuous case. Rather, positivity depends on the geometry of the cost. The twist condition (injectivity of $\nabla_x c(x,\cdot)$) plays an important role, like in many results on optimal transport. We include affirmative positivity results in particular for quadratic costs, which is the most important case for applications. While not pursued in the present paper, our results should also be useful to derive detailed quantitative bounds on the rate in more specific settings. We may also hope to gain insights into how the rate depends on the dimension.

Geometry is a cornerstone in the now-classical theory of optimal transport, where optimality is captured geometrically by the $c$-cyclical monotonicity of a transport's support. Defined by comparing costs at finitely many points, it yields a powerful tool to derive fundamental results such as stability of optimal transports under weak limits or existence of dual potentials. We are not aware of a comparable technique in the literature on entropic optimal transport (or on Schrödinger bridges). In this paper, we exploit a cyclical invariance property satisfied by the density of $\pi_\varepsilon$. The invariance itself can be understood as a reformulation of a classical characterization for $\pi_\varepsilon$ through the solution of the dual problem, the Schrödinger potentials. The novelty here lies in exploiting the geometric aspect and working on the primal side, following the spirit of $c$-cyclical monotonicity. Like in classical optimal transport, the arguments are remarkably simple and general once the correct notions are in place. Our technique is a departure from the control-theoretic methods in the related literature. Case in point, the geometric proof that a weak limit $\pi = \lim_{\varepsilon\to 0} \pi_\varepsilon$ is an optimal transport (cf. Proposition 3.2), is nearly trivial compared to the Gamma-convergence technique, even in the general Polish context. (Of course, Gamma-convergence is applicable to many other problems where our technique has no analogue.)

We also emphasize another benefit which may illustrate that cyclical invariance is in fact more than just a reformulation of control theory or convex analysis: the geometry singles out a unique coupling $\pi_\varepsilon$ even if the value function (1.1) is infinite and hence the usual notion of solution as a minimizer is not meaningful. This is crucial for instance if costs are quadratic but one of the marginal distributions does not have a finite second moment. Our arguments for the large deviations result apply in that setting without any added difficulty, paralleling the geometric insights in classical optimal transport. (On the other hand, the *existence* of $\pi_\varepsilon$ in the case of infinite value functions is not immediate. We establish it in [32], together with a stability theorem for $\pi_\varepsilon$, using the same geometric standpoint.) Indeed, we expect the technique to be useful in several other aspects of entropic optimal transport and Schrödinger bridges, and thus the technique may be as important a contribution as the main theorem.

The present paper is organized as follows. After reviewing motivations for our research and related literature in the remainder of this Introduction, Section 2 details the basic definitions and introduces cyclical invariance. In Section 3, this notion is used to prove that cluster points of $\pi_\varepsilon$ as $\varepsilon \to 0$ have $c$-cyclically monotone support, hence are optimal transports. The main result on large deviations is obtained in Section 4: part (a) of Theorem 1.1 is stated as Corollary 4.3 whereas (b) is split into Corollaries 4.7 and 4.12, each covering one of the two alternative assumptions. Section 5 gives examples of settings where the rate function $I$ is strictly positive outside the support $\Gamma$, with a focus on quadratic costs. Appendix A contains facts about Schrödinger bridges and a derivation of the cyclical invariance property. In Appendix B, we detail two general settings where Assumption 4.4 on the uniqueness of Kantorovich potentials is satisfied. Finally, Appendix C shows how to translate the results on the positivity of $I$ in Section 5 from quadratic costs to more general cost functions by means of $c$-convex analysis.

## 1.1 Related Literature

In the literature on finite-dimensional linear programs and their entropic regularization, the early work [17] contains a very detailed study of primal and dual convergence, expansion of the value function, and characterizations of the rates. Their setting includes discrete optimal transport problems with marginals supported by finitely many points, and in that case the pointwise results in [17] certainly include the large deviations result for $\varepsilon \to 0$. On the other hand, our main theorem is most relevant when at least one marginal support is connected, hence is complementary to the discrete case. More recently, [59] proved an exponential convergence bound for finite-dimensional linear programs. While the bound is not sharp in a pointwise sense, the result is non-asymptotic; i.e., holds for all $\varepsilon$ below a known threshold. Moreover, the constants are known in terms of the data, which provided valuable intuition for our construction of the rate function $I$. One may also observe how the constants in [59] blow up as the cardinality of the support increases.

In the last decade, optimal transport has found myriad applications in machine learning, statistics, image processing, language processing, and other areas. The literature in the computational area has expanded very quickly and our account is highly incomplete; see [51] for a recent monograph with extensive references. Exact computation of an optimal transport between marginals with $n$ atoms costs $O(n^3 \log n)$, prohibitive for modern applications with large data sets. The recent success of applied optimal transport is enabled by the advent of fast approximate solvers, and entropic regularization is among the most influential schemes for high-dimensional problems. Popularized by Cuturi [21] in this domain, it allows for the application of Sinkhorn's algorithm (also called iterative proportional fitting, and also due to Deming, Stephan, Fortet, Knopp and others) where each iteration is a matrix-vector multiplication costing $O(n^2)$. Importantly for modern applications, it is highly parallelizable on GPUs; a number of further advantages are highlighted in [8]. The convergence of this algorithm was rigorously discussed in [35, 56], among others. More recently, it was shown that $\delta$-accurate approximations of the transport cost can be obtained in $\tilde{O}(n^2/\delta)$ operations via entropic regularization; cf. [10, 41] and the references therein. In addition to computation accuracy, a second error in practice stems from sampling the marginals. For entropic optimal transport (with $\varepsilon > 0$ fixed), the rate of convergence of the empirical cost towards its population limit does not depend on the dimension, in contrast to the curse of dimensionality suffered by its unregularized counterpart [31, 45]. Addressing the combined problem, [9] studies the convergence of the discrete Sinkhorn algorithm to an optimal transport potential in the joint limit when $\varepsilon_n \to 0$ and the marginals $\mu, \nu$ are approximated by discretizations $\mu_n, \nu_n$ satisfying a certain density property. Explicit error bounds are derived, for instance for quadratic cost on the torus, yielding important insights into the optimal trade-off between $n$ and $\varepsilon$. In the present study, we focus on the discrepancy between the entropic optimizer $\pi_\varepsilon$ and the optimal transport $\pi_*$ in a general setting and adopt a local point of view.

Continuing with a different branch of related literature, recall that entropic optimal transport can also be phrased as the (static) Schrödinger bridge problem. Informally stated, consider a system of diffusing particles from time $t_0$ to $t_1$ in thermal equilibrium, and a given joint "reference" law $R$ for its configuration at those times. If marginals $(\mu, \nu)$ differing from the ones of $R$ are observed, what is the most likely evolution (joint law of $\mu, \nu$) of the system conditional on $R$? Schrödinger's answer amounts to $\pi^* = \arg\min_{\Pi(\mu,\nu)} H(\cdot|R)$; see [28, 39] for extensive surveys including historical accounts. (This is the static formulation. Given the origins in physics, it is natural that much of the literature focuses on the *dynamic* Schrödinger bridge problem, which asks for the dynamic evolution of the particle system over time $t \in [t_0, t_1]$. The static problem is recovered by projecting to the marginals.)

The minimization of $H(\cdot|R)$ over $\Pi(\mu, \nu)$ coincides with the entropic optimal transport problem (1.1) if we introduce the cost function $c := -\varepsilon \log(\alpha^{-1} dR/d(\mu \otimes \nu))$, where the parameter $\varepsilon > 0$ is arbitrary and $\alpha$ is a normalizing constant (we tacitly assume that $R \sim \mu \otimes \nu$). Conversely, taking (1.1) as the starting point, defining $R(\varepsilon)$ by $dR(\varepsilon)/d(\mu \otimes \nu) = \alpha e^{-c/\varepsilon}$

yields the associated Schrödinger bridge problem. Assuming for simplicity that $\{c = 0\}$ is the graph of a function $f : \mathsf{X} \to \mathsf{Y}$, Theorem 1.1 is then a large deviations principle as the reference measure $R(\varepsilon)$ degenerates to a deterministic coupling (meaning that a particle with given origin $x$ travels to the predetermined destination $f(x)$).[1] This is also called the small-noise or small-time limit. While not pursued here, it seems plausible that a similar principle could be established for more general sequences $R(\varepsilon)$. From the point of view of Schrödinger bridges, another interesting follow-up question is whether a comparable large deviations result can be stated for the dynamic problem on path space.

Mikami [46, 47] first highlighted the connection between Schrödinger equations and optimal transport in the small-noise limit; see also [14] for a connection through a fluid dynamic formulation. Léonard studied Schrödinger bridges in a series of works starting with [36, 37]; see [39] for further references. In [38], he established convergence of the value function to an optimal transport problem in the sense of Gamma-convergence for a general formulation of the problem. See also [13] where a very accessible proof of the Gamma-convergence is presented for quadratic costs. More recently, [18, 50] study the limit in specific settings and determine higher-order terms in the expansion of the Schrödinger (or entropic) value function around the optimal transport cost. These works complement earlier results of [1, 26, 27] showing that the large deviation rate function for the empirical distribution of independent Brownian particles with drift is asymptotically equivalent to the Jordan–Kinderlehrer–Otto functional arising in the Wasserstein gradient flow. We mention that [18] also considers the large-time limit (corresponding to $\varepsilon \to \infty$); cf. [16] for recent developments. The setup in [50] is closest to ours in that the entropic penalty and the limit $\varepsilon \to 0$ are formulated in the same way, whereas the literature on Schrödinger bridges often formulates the zero-noise limit through a vanishing Laplacian. We also mention [34], establishing convergence of the dual potentials for compact marginals (see [49] for a follow-up and more on the relation to the present work).

While the focus of the aforementioned works is on value functions and global quantities, the present study focuses on the local geometry and convergence. The value functions are not used at all, and so it is quite natural that the results hold even when costs are infinite. We are not aware of a large deviations principle similar to ours in the extant literature. One concrete example where these aspects are of interest, are the multidimensional ranks and quantiles that have been introduced in statistics to extend the usual scalar notions and familiar nonparametric tests; see [15, 23, 24, 33]. Here

[1] Schrödinger's ideas about the "most likely evolution" are usually presented as a large deviations result in the modern literature. That result is very different from the one just discussed.

Brenier's map is fundamental, but like in the scalar case, moment conditions are not natural. McCann's geometric extension [44] of Brenier's map (see also [58, pp. 249–258]) can be used to provide a definition irrespectively of the finiteness of the value function. Unlike their scalar counterparts, the ranks defined through optimal transport are computationally expensive. Entropic optimal transport resolves that issue and provides an approximate Brenier's map. Leveraging this idea, a notion of "differentiable ranks" based on entropic optimal transport was recently proposed in [22]. We expect that our results can be used to study the local deviations of these differentiable ranks from the unregularized ones.

Related to our technique in a broader sense, there have been recent works successfully using ideas of $c$-cyclical monotonicity outside the setting of classical optimal transport. Examples include martingale optimal transport [6] and optimal Skorokhod embeddings [5, 7]. Finally, we mention the intriguing "optimal entropy-transport problem" studied in [40]. Here, the usual optimal transport problem is relaxed in that the marginal constraints are replaced by an entropic penalty relative to a given pair of measures. While similar in name, this problem is quite different from ours, where the marginal constraints are strictly enforced and the entropy of the joint distribution is used as penalty.

## 2 Cyclical Invariance

Let $(\mathsf{X},\mu)$ and $(\mathsf{Y},\nu)$ be Polish probability spaces endowed with their Borel $\sigma$-fields and let $c : \mathsf{X}\times\mathsf{Y}\to\mathbb{R}_+$ be a measurable (cost) function. The associated optimal transport problem is

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathsf{X}\times\mathsf{Y}} c\,d\pi \tag{2.1}$$

where $\Pi(\mu,\nu)$ is the set of all couplings; that is, probability measures $\pi$ on $\mathsf{X}\times\mathsf{Y}$ with marginals $\mu=(\mathrm{proj}_{\mathsf{X}})_{\#}\pi$ and $\nu=(\mathrm{proj}_{\mathsf{Y}})_{\#}\pi$. Given a constant $\varepsilon>0$, the entropic optimal transport problem is

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathsf{X}\times\mathsf{Y}} c\,d\pi+\varepsilon H(\pi|P), \quad P:=\mu\otimes\nu, \tag{2.2}$$

where $H$ denotes the relative entropy or Kullback–Leibler divergence,

$$H(\pi|P):=\begin{cases}\int\log(\frac{d\pi}{dP})\,d\pi, & \pi\ll P,\\ \infty, & \pi\not\ll P.\end{cases}$$

As detailed in Proposition A.1 of Appendix A, this problem admits a unique minimizer $\pi_\varepsilon$ whenever the value (2.2) is finite; i.e., whenever

$$\text{there exists } \pi\in\Pi(\mu,\nu) \text{ with } \textstyle\int c\,d\pi+H(\pi|P)<\infty. \tag{2.3}$$

Moreover, we then have $\pi_\varepsilon \sim P$.

**Definition 2.1.** A coupling $\pi \in \Pi(\mu, \nu)$ is called $(c, \varepsilon)$-*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

$$\prod_{i=1}^{k} \frac{d\pi}{dP}(x_i, y_i) = \exp\left( -\frac{1}{\varepsilon}\left[ \sum_{i=1}^{k} c(x_i, y_i) - \sum_{i=1}^{k} c(x_i, y_{i+1}) \right] \right) \prod_{i=1}^{k} \frac{d\pi}{dP}(x_i, y_{i+1}) \tag{2.4}$$

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 omit the qualifier $(c, \varepsilon)$ when there is no ambiguity. One elementary way to motivate Definition 2.1 is to derive a first-order condition of optimality for (2.2) through variational arguments in the case of discrete marginals, which indeed yields (2.4). Cyclical invariance can be phrased more succinctly using the auxiliary reference measure $R = R(\varepsilon)$ defined by the Gibbs kernel

$$\frac{dR}{dP} = \alpha e^{-c/\varepsilon}, \tag{2.5}$$

where $\alpha = (\int e^{-c/\varepsilon}\, dP)^{-1}$ is the normalizing constant. As $R \sim P$, we can state (2.4) as

$$\prod_{i=1}^{k} \frac{d\pi}{dR}(x_i, y_i) = \prod_{i=1}^{k} \frac{d\pi}{dR}(x_i, y_{i+1}). \tag{2.6}$$

This condition, in turn, is related to a multiplicative decomposition of the density $d\pi/dR$; cf. Appendix A. For our analysis of the limit $\varepsilon \to 0$, the less elegant definition (2.4) will be the more useful one, as it makes explicit the role of $\varepsilon$ and links directly to the $c$-cyclical monotonicity condition of optimal transport.

**Proposition 2.2.** *(a) There is at most one* $(c, \varepsilon)$*-cyclically invariant coupling* $\pi \in \Pi(\mu, \nu)$.

*(b) Let* (2.3) *hold. Then* $\pi \in \Pi(\mu, \nu)$ *is* $(c, \varepsilon)$*-cyclically invariant if and only if it minimizes* (2.2)*. Moreover, there exists a unique such coupling.*

The proof is detailed in Appendix A. Under Condition (2.3), Proposition 2.2 shows the equivalence between minimality and cyclical invariance. The notion of minimality is meaningful only under (2.3), otherwise all couplings have infinite cost. By contrast, we show in [32] that the notion of cyclical invariance remains meaningful in this context of infinite costs: existence and uniqueness hold under mild regularity conditions; e.g., when $\mathsf{X}, \mathsf{Y}$ are Euclidean spaces and $c$ is continuous.

In the remainder of this paper, we simply *assume* that a $(c, \varepsilon)$-cyclically invariant coupling $\pi_\varepsilon \in \Pi(\mu, \nu)$ exists for every $\varepsilon > 0$, rather than imposing

Condition (2.3) as in much of the literature. One reason is that this condition precludes some applications of interest to us. In any event, the arguments in this paper do not simplify if (2.3) is assumed.

## 3 Cluster Points as $\varepsilon \to 0$

Denote by $\pi_\varepsilon$ the unique $(c, \varepsilon)$-cyclically invariant coupling. In this section we show that cluster points of $\pi_\varepsilon$ as $\varepsilon \to 0$ are $c$-cyclically monotone. The estimates leading to that conclusion are obtained by simply integrating the cyclical invariance condition.

**Lemma 3.1.** *Let $k \geq 2$ and $0 \leq \delta \leq \delta' \leq \infty$. Define*

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

*and let $A \subset A_k(\delta, \delta')$ be Borel. Then $\pi_\varepsilon^k := \prod_{i=1}^k \pi_\varepsilon(dx_i, dy_i)$ satisfies*

$$\pi_\varepsilon^k(A) \leq e^{-\delta/\varepsilon} \quad \textit{for all} \quad \varepsilon > 0. \tag{3.1}$$

*Suppose in addition that $\bar{A} := \big\{(x_i, y_{i+1})_{i=1}^k : (x_i, y_i)_{i=1}^k \in A\big\}$ satisfies $\liminf_{\varepsilon \to 0} \varepsilon \log \pi_\varepsilon^k(\bar{A}) = 0$. Then*

$$\liminf_{\varepsilon \to 0} \varepsilon \log \pi_\varepsilon^k(A) \geq -\delta'. \tag{3.2}$$

*Proof.* Set $Z = d\pi_\varepsilon/dP$. Using (2.4), we have for $P^k$-a.e. $(x_i, y_i)_{i=1}^k \in A$ that

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

Integrating over $A$ with respect to $P^k = \prod P(dx_i, dy_i) = \prod P(dx_i, dy_{i+1})$ yields

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

which is (3.1). Analogously, $\pi_\varepsilon^k(A) \geq e^{-\delta'/\varepsilon} \pi_\varepsilon^k(\bar{A})$ and hence

$$\varepsilon \log \pi_\varepsilon^k(A) \geq -\delta' + \varepsilon \log \pi_\varepsilon^k(\bar{A}),$$

so that (3.2) follows under the stated condition on $\bar{A}$. □

In all that follows, probability measures are considered with weak convergence; i.e., the topology induced by bounded continuous functions. We recall that $\Pi(\mu, \nu)$ is weakly compact; cf. [58, p. 45]. As a consequence, any sequence of couplings admits at least one cluster point, and any cluster point is a coupling. A set $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ is called $c$-cyclically monotone if $\sum_{i=1}^k c(x_i, y_i) \leq \sum_{i=1}^k c(x_i, y_{i+1})$ for all $k \geq 1$ and $(x_i, y_i) \in \Gamma$, $1 \leq i \leq k$.

**Proposition 3.2.** *Let $c$ be continuous and let $\pi$ be a cluster point of $(\pi_\varepsilon)$ as $\varepsilon \to 0$. Then $\operatorname{spt}\pi$ is $c$-cyclically monotone, hence $\pi$ is an optimal transport as soon as the optimal transport problem* (2.1) *is finite. If* (2.1) *admits a unique $c$-cyclically monotone coupling $\pi_* \in \Pi(\mu,\nu)$, then $\pi_\varepsilon \to \pi_*$ as $\varepsilon \to 0$.*

*Proof.* Let $\varepsilon_n \to 0$ and $\pi_{\varepsilon_n} \to \pi$. 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 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, $U_1 \times \cdots \times U_k \subset A_k(\delta, \infty)$ implies $\pi^k_{\varepsilon_n}(U_1 \times \cdots \times U_k) \to 0$ by Lemma 3.1, a contradiction. This shows that $\operatorname{spt}\pi$ is $c$-cyclically monotone. It is well known that cyclical monotonicity and optimality are equivalent when (2.1) is finite; cf. [58, Theorem 5.10, p. 57]. As $\Pi(\mu,\nu)$ is compact, $\pi_\varepsilon$ must have cluster points as $\varepsilon \to 0$, so that uniqueness implies convergence. □

**Remark 3.3.** For the particular case of quadratic cost on $\mathbb{R}^d$ and marginals satisfying certain integrability conditions, the conclusion of Proposition 3.2 is obtained in [13] by (arguably much more involved) Gamma-convergence arguments. That line of argument focuses on the properties of the value function, hence cannot be applied when the value function is infinite. A related but slightly different convergence result, also obtained by Gamma-convergence, is stated in [38, Theorem 2.4] and includes lower semicontinuous cost functions. On the other hand, the convergence in Proposition 3.2 may fail if continuity is relaxed to lower semicontinuity: one example, discussed in more detail in [48, Remark 4.3], is $c(x,y) = \mathbf{1}_{\{x \neq y\}}$ and $\mu = \nu = \mathrm{Unif}[0,1]$.

Uniqueness of $c$-cyclically monotone transports is known for many examples of continuous or semi-discrete optimal transport problems—arguably for most of the important examples except distance costs—and then Proposition 3.2 shows the convergence of $\pi_\varepsilon$ as $\varepsilon \to 0$. See, e.g., [58, Theorem 5.30, p. 84]. When the transport problem admits multiple solutions, it is not obvious whether $\pi_\varepsilon$ converges. If there exists an optimal transport $\pi$ with $H(\pi|P) < \infty$, one can show that $\pi_\varepsilon$ converges to the unique optimal transport $\pi_*$ with minimal relative entropy $H(\cdot|P)$; cf. [48, Theorem 5.1]. This includes the discrete case with finitely supported marginals as analyzed in [17], but also the semi-discrete case (where one marginal is continuous) under minor integrability conditions. Convergence is also known for the scalar Monge problem where $c(x,y) = |x-y|$ on $\mathsf{X} = \mathsf{Y} = \mathbb{R}$ and the marginals are absolutely continuous; here a relatively explicit analysis is possible [25]. It has been conjectured that convergence holds in a general setting.

# 4 Rate Function

Throughout this section, the cost function $c$ is assumed to be continuous. For simplicity of exposition, we shall also assume that

$$\pi_\varepsilon \to \pi_* \quad \text{as} \quad \varepsilon \to 0, \tag{4.1}$$

for some (necessarily $c$-cyclically monotone) transport $\pi_* \in \Pi(\mu, \nu)$. However, if it is merely known that $\pi_{\varepsilon_n} \to \pi_*$ along a specific sequence $\varepsilon_n \to 0$, then all of our results hold along that sequence, regardless of whether $(\pi_\varepsilon)$ has other cluster points. In fact, the arguments in this paper are *complementary* to the question of convergence discussed in the preceding paragraph: *given* the convergence of a sequence, we describe the large deviations.

## 4.1 Large Deviations Upper Bound

In this subsection we introduce the function $I$ and show the large deviations upper bound; i.e., that $I$ provides a lower bound for the large deviations rate. With the definitions in place, the arguments are straightforward and apply in great generality. We write $B_r(z)$ for the open ball of radius $r$ around $z$, in any metric space. The first lemma is a way to bound the decay of a ball in $\mathsf{X} \times \mathsf{Y}$ based on the estimate for subsets of $(\mathsf{X} \times \mathsf{Y})^k$ in Lemma 3.1.

**Lemma 4.1.** *Let $(x, y) \in \mathsf{X} \times \mathsf{Y}$. Suppose there exist $(x_i, y_i)_{2 \le i \le k} \subset \operatorname{spt} \pi_*$ with $k \ge 2$ such that*

$$\delta_0 := \sum_{i=1}^{k} c(x_i, y_i) - \sum_{i=1}^{k} c(x_i, y_{i+1}) > 0, \quad \textit{where} \quad (x_1, y_1) := (x, y).$$

*Given $\delta < \delta_0$, there exist $\alpha, r, \varepsilon_0 > 0$ such that*

$$\pi_\varepsilon(B_r(x, y)) \le \alpha e^{-\delta/\varepsilon} \quad \textit{for} \quad \varepsilon \le \varepsilon_0.$$

*Proof.* Once again, continuity of $c$ implies that for $r > 0$ small enough, $\sum c(\tilde{x}_i, \tilde{y}_i) - \sum c(\tilde{x}_i, \tilde{y}_{i+1}) \ge \delta$ for all $(\tilde{x}_i, \tilde{y}_i) \in B_i := B_r(x_i, y_i)$, and then $B_1 \times \cdots \times B_k \subset A_k(\delta, \infty)$ in Lemma 3.1 yields

$$\pi_\varepsilon(B_1) \cdots \pi_\varepsilon(B_k) \le e^{-\delta/\varepsilon}. \tag{4.2}$$

For $i \ge 2$ we have $\liminf \pi_\varepsilon(B_i) \ge \pi_*(B_i)$ due to the weak convergence $\pi_\varepsilon \to \pi_*$, and $\beta_i := \pi_*(B_i) > 0$ as $(x_i, y_i) \in \operatorname{spt} \pi_*$. Let $\beta = \min_{i \ge 2} \beta_i$. Then $\pi_\varepsilon(B_i) \ge \beta/2$ for $i \ge 2$ and $\varepsilon$ small, and thus (4.2) yields $\pi_\varepsilon(B_1) \le (\beta/2)^{1-k} e^{-\delta/\varepsilon}$. □

Denote by $\Sigma(k)$ the set of permutations of $\{1,\dots,k\}$. Next, we state the definition of $I(x,y)$; it is designed to capture the rate $\delta$ in Lemma 4.1 and optimize it over the choice of $(x_i,y_i)_{2\leq i\leq k}$.

**Lemma 4.2.** *Given a $c$-cyclically monotone set $\emptyset\neq\Gamma\subseteq\mathsf{X}\times\mathsf{Y}$, define*

$$I(x,y):=\sup_{k\geq 2}\sup_{(x_i,y_i)_{i=2}^k\subset\Gamma}\sup_{\sigma\in\Sigma(k)}\sum_{i=1}^k c(x_i,y_i)-\sum_{i=1}^k c(x_i,y_{\sigma(i)}) \tag{4.3}$$

*where $(x_1,y_1):=(x,y)$. Then $I:\mathsf{X}\times\mathsf{Y}\to[0,\infty]$ is lower semicontinuous and $I=0$ on $\Gamma$. We have*

$$I(x,y)\geq\sup_{k\geq 2}\sup_{(x_i,y_i)_{i=2}^k\subset\Gamma}\sum_{i=1}^k c(x_i,y_i)-\sum_{i=1}^k c(x_i,y_{i+1}), \tag{4.4}$$

*and equality holds as soon as $x\in\mathsf{X}_0:=\operatorname{proj}_{\mathsf{X}}\Gamma$ or $y\in\mathsf{Y}_0:=\operatorname{proj}_{\mathsf{Y}}\Gamma$.*

*Proof.* We have $I\geq 0$ as $\sigma=\mathrm{Id}$ is a possible choice in (4.3). For $(x,y)\in\Gamma$, the difference of sums in (4.3) is nonpositive by cyclical monotonicity. The semicontinuity follows from the continuity of $c$.

Let $I'(x,y)$ be the right-hand side of (4.4). As the pairs $(x_i,y_i)_{i=2}^k$ can be relabeled arbitrarily, this is the same as (4.3) except that the last supremum in (4.4) is taken over $\sigma\in\Sigma(k)\setminus\{\mathrm{Id}\}$. If $I(x,y)>0$, the identity permutation is not optimal for the relevant pairs $(x_i,y_i)_{i=2}^k$ and equality must hold in (4.4). Thus, if equality fails, then $I(x,y)=0$ whereas $I'(x,y)<0$. Let $x\in\mathsf{X}_0$, then we can choose $k=2$ and $(x_2,y_2)\in\Gamma$ with $x_2=x$, which yields $\sum_{i=1}^2 c(x_i,y_i)-\sum_{i=1}^2 c(x_i,y_{i+1})=0$ and hence $I'(x,y)\geq 0$. The argument for $y\in\mathsf{Y}_0$ is symmetric. □

The reader may ignore the difference between (4.3) and (4.4); it is merely a notational nuisance. We have the following result for the $c$-cyclically monotone set $\Gamma:=\operatorname{spt}\pi_*$, also stated as Theorem 1.1 (a) in the Introduction.

**Corollary 4.3.** *For any compact set $C\subset\mathsf{X}\times\mathsf{Y}$,*

$$\limsup_{\varepsilon\to 0}\varepsilon\log\pi_\varepsilon(C)\leq-\inf_{(x,y)\in C}I(x,y).$$

*Proof.* Fix $\eta>0$ and $(x,y)\in C$. By the definition of $I(x,y)$ there are $k\geq 1$ and $(x_i,y_i)_{i=2}^k\subset\Gamma$ such that

$$\sum_{i=1}^k c(x_i,y_i)-\sum_{i=1}^k c(x_i,y_{i+1})>I_\eta(x,y)-\eta/2,$$

where $(x_1, y_1) := (x, y)$ and $I_\eta(x, y) := I(x, y) \wedge \eta^{-1}$. (The truncation is needed only if $I(x, y) = \infty$.) Lemma 4.1 thus yields a ball $B_r(x, y)$ with

$$\limsup \varepsilon \log \pi_\varepsilon(B_r(x, y)) \leq -I_\eta(x, y) + \eta. \tag{4.5}$$

This holds for every $(x, y) \in C$, and as $C$ is covered by finitely many such balls, we deduce that

$$\limsup \varepsilon \log \pi_\varepsilon(C) \leq - \inf_{(x,y)\in C} I_\eta(x, y) + \eta.$$

Recalling that $\eta > 0$ was arbitrary, the claim follows. □

We note that the measure $\pi_* = \lim_\varepsilon \pi_\varepsilon$ is not compactly supported in general. It is then an open problem how to relax the compactness condition in Corollary 4.3 and hence obtain a “stronger” version of the large deviations principle.

## 4.2 Large Deviations Lower Bound

Our next aim is to show that $I$ is also an upper bound for the large deviations rate, thus matching the bound in Corollary 4.3. This will be accomplished in two slightly different settings and approaches. The dual approach expresses $I$ as the gap (4.6) between the cost $c$ and the solution of the dual optimal transport problem, whereas the primal directly uses the definition (4.3) of $I$ and imposes regularity conditions. The results correspond to Theorem 1.1 (b) in the Introduction.

### 4.2.1 Bound via Kantorovich Potential

We start with the dual approach, first recalling some standard notions of optimal transport—we have tried to consistently use the notation of [58]. A proper function $\psi : \mathsf{X} \to (-\infty, \infty]$ is called $c$-convex if there exists some $\zeta : \mathsf{Y} \to [-\infty, \infty]$ such that $\psi(x) = \sup_{y\in\mathsf{Y}}[\zeta(y) - c(x, y)]$ for all $x \in \mathsf{X}$. Its $c$-conjugate is defined by $\psi^c(y) := \inf_{x\in\mathsf{X}}[\psi(x) + c(x, y)]$ for $y \in \mathsf{Y}$, and its $c$-subdifferential is

$$\partial_c\psi = \{(x, y) \in \mathsf{X} \times \mathsf{Y} : \psi^c(y) - \psi(x) = c(x, y)\}.$$

Given a $c$-cyclically monotone set $\Gamma$, a $c$-convex function $\psi$ is called a Kantorovich potential if $\Gamma \subset \partial_c\psi$; that is, if $\psi^c(y) - \psi(x) = c(x, y)$ on $\Gamma$. This implies in particular that $\psi, \psi^c$ are finite on

$$\mathsf{X}_0 := \mathrm{proj}_{\mathsf{X}}\,\Gamma, \quad \mathsf{Y}_0 := \mathrm{proj}_{\mathsf{Y}}\,\Gamma.$$

In the context of optimal transport, $\operatorname{spt}\pi \subset \partial_c\psi$ for some optimal $\pi \in \Pi(\mu,\nu)$ implies that $\partial_c\psi$ contains the support of any optimal transport. Indeed, $\partial_c\psi$ is a maximal $c$-monotone set for inclusion. In what follows, the cyclically monotone set of interest is $\Gamma = \operatorname{spt}\pi_*$, where $\pi_*$ is the limiting optimal transport (4.1).

**Assumption 4.4.** Uniqueness of Kantorovich potentials holds on $\mathsf{X}_0$; that is, for any $c$-convex functions $\psi_1, \psi_2$ on $\mathsf{X}$ with $\Gamma \subset \partial_c\psi_i$, it holds that $\psi_1 - \psi_2$ is constant on $\mathsf{X}_0$.

This is often considered a fairly weak assumption, at least for differentiable cost functions, and we detail sufficient conditions in Proposition B.2 of Appendix B. However, we emphasize that connectedness of at least one marginal support is crucial (cf. Example 4.8 below).

As announced, Assumption 4.4 allows us to express $I$ through the Kantorovich potential; see (4.6). For our present purpose, the key consequence is (4.7). It is worth noting that (4.6) also allows us to translate a large body of known results about $c$-convex functions, such as regularity results, into statements about $I$. Finally, the gap (4.6) also plays a role in the regularity theory of optimal transport maps (especially in [42]), thus relating to the second approach in Section 4.2.2 below.

**Proposition 4.5.** *Let Assumption 4.4 hold. Then*

$$I(x,y) = c(x,y) - \psi^c(y) + \psi(x), \quad (x,y) \in \mathsf{X}_0 \times \mathsf{Y}_0 \tag{4.6}$$

*for any Kantorovich potential $\psi$. In particular, $I < \infty$ on $\mathsf{X}_0 \times \mathsf{Y}_0$. If $(x,y),(x',y') \in \mathsf{X}_0 \times \mathsf{Y}_0$ are such that $(x',y),(x,y') \in \Gamma$, then*

$$I(x,y) + I(x',y') = c(x,y) + c(x',y') - c(x,y') - c(x',y). \tag{4.7}$$

*Proof.* We first elaborate on Assumption 4.4. A particular family of Kantorovich potentials, sometimes called Rockafellar antiderivatives of $\Gamma$, is defined as follows (cf. [58, Equation (5.17), p. 65]): fix $(x_0,y_0) \in \Gamma$ and set

$$\psi_{(x_0,y_0)}(x) := \sup_{k\geq 1} \sup_{(x_i,y_i)_{i=1}^k \in \Gamma} \sum_{i=0}^{k} [c(x_i,y_i) - c(x_{i+1},y_i)], \quad \text{where } x_{k+1} := x. \tag{4.8}$$

It then holds that $\psi_{(x_0,y_0)}(x) = 0$ for $x = x_0$. Clearly Assumption 4.4 implies that changing the reference point $(x_0,y_0)$ only changes this potential by a constant. In particular,

$$\Psi_{(x_0,y_0)}(x,y) := \psi^c_{(x_0,y_0)}(y) - \psi_{(x_0,y_0)}(x), \quad (x,y) \in \mathsf{X}_0 \times \mathsf{Y}_0 \tag{4.9}$$

does not depend on $(x_0, y_0) \in \Gamma$, and we may simply write $\Psi := \Psi_{(x_0,y_0)}$. Indeed, under Assumption 4.4, $\Psi$ is even the same for any potential $\psi$.

We now use this independence to prove the lemma. To avoid notational conflict, we first rewrite the definition (4.8) as

$$\psi_{(\bar{x},\bar{y})}(x) = \sup_{k\geq 2} \sup_{(x_i,y_i)_{i=2}^k \in \Gamma} c(\bar{x},\bar{y}) + \sum_{i=2}^{k} [c(x_i,y_i) - c(x_{i+1},y_i)] - c(x_2,\bar{y}), \tag{4.10}$$

where we have avoided the subscript $i = 1$. Fix $(x, y) \in \mathsf{X}_0 \times \mathsf{Y}_0$. Writing $(x_1, y_1) := (x, y)$ as in Lemma 4.2, the definition $x_{k+1} := x$ of (4.8) becomes our usual cyclical convention $x_{k+1} = x_1$. As $y \in \mathsf{Y}_0$, there exists $\bar{x} \in \mathsf{X}_0$ such that $(\bar{x}, y) \in \Gamma$. Using (4.10) with $\bar{y} := y$ then yields

$$\begin{aligned}
\psi_{(\bar{x},y)}(x) &= \sup_{k\geq 2} \sup_{(x_i,y_i)_{i=2}^k \in \Gamma} c(\bar{x},y) + \sum_{i=2}^{k} c(x_i,y_i) - \sum_{i=2}^{k} c(x_{i+1},y_i) - c(x_2,y) \\
&= \sup_{k\geq 2} \sup_{(x_i,y_i)_{i=2}^k \in \Gamma} c(\bar{x},y) - c(x,y) + \sum_{i=1}^{k} c(x_i,y_i) - \sum_{i=1}^{k} c(x_{i+1},y_i) \\
&= c(\bar{x},y) - c(x,y) + \sup_{k\geq 2} \sup_{(x_i,y_i)_{i=2}^k \in \Gamma} \sum_{i=1}^{k} c(x_i,y_i) - \sum_{i=1}^{k} c(x_i,y_{i+1}) \\
&= c(\bar{x},y) - c(x,y) + I(x,y),
\end{aligned}$$

where we have used the last part of Lemma 4.2. In view of $\psi_{(\bar{x},y)}(\bar{x}) = 0$, the fact that $\Psi_{(\bar{x},y)} = c$ on $\Gamma$ shows in particular that $c(\bar{x}, y) = \psi^c_{(\bar{x},y)}(y)$, and hence the preceding display yields

$$I(x,y) = c(x,y) + \psi_{(\bar{x},y)}(x) - \psi^c_{(\bar{x},y)}(y) = c(x,y) - \Psi_{(\bar{x},y)}(x,y).$$

By the first part of the proof, $\Psi_{(\bar{x},y)}(\cdot) = \Psi(\cdot)$ does not depend on $(\bar{x}, y)$ and the above is precisely (4.6).

To see (4.7), let $(x', y), (x, y') \in \Gamma$. Using that $I = 0$ on $\Gamma$ by Lemma 4.2 and then (4.6),

$$\begin{aligned}
I(x,y) + I(x',y') &= I(x,y) + I(x',y') - I(x,y') - I(x',y) \\
&= c(x,y) + c(x',y') - c(x,y') - c(x',y) \\
&\quad - \Psi(x,y) - \Psi(x',y') + \Psi(x,y') + \Psi(x',y),
\end{aligned}$$

where the last line vanishes as $\Psi$ is a sum of marginal functions (4.9). □

**Remark 4.6.** The proof of Proposition 4.5 is based on the condition that

$$\Psi_{(x_0,y_0)} \text{ of (4.9) does not depend on } (x_0,y_0) \in \Gamma, \tag{4.11}$$

which may seem weaker than Assumption 4.4. However, Assumption 4.4 is in fact equivalent to (4.11); the proof is stated below. As a direct consequence, another equivalent condition is that the Rockafellar antiderivative (4.8) be independent of $(x_0, y_0)$. The symmetry of (4.11) shows that it is further equivalent to impose the analogue of Assumption 4.4 on $\mathsf{Y}$ instead of $\mathsf{X}$.

*Proof that* (4.11) *implies Assumption 4.4.* By construction, the Rockafellar antiderivative $\psi_0 := \psi_{(x_0,y_0)}$ of (4.8) has the minimality property $\psi_0 \leq \xi$ on $\mathsf{X}_0$ whenever $\xi$ is a potential with $\xi(x_0) = 0 = \psi_0(x_0)$. (See [58, p. 62], or [4] for a more general result and further context.) Consider another point $(x_1, y_1) \in \Gamma$, let $\psi_1 := \psi_{(x_1,y_1)}$ and let $\xi$ be any potential. Using the minimality twice,

$$\psi_0(x_1) - \psi_0(x_0) \leq \xi(x_1) - \xi(x_0) \leq \psi_1(x_1) - \psi_1(x_0).$$

Given (4.11), the right-hand side can be expressed as

$$\begin{aligned}\psi_1(x_1) - \psi_1(x_0) &= \psi_1(x_1) - \psi_1^c(y_0) - \psi_1(x_0) + \psi_1^c(y_0)\\ &= \psi_0(x_1) - \psi_0^c(y_0) - \psi_0(x_0) + \psi_0^c(y_0)\\ &= \psi_0(x_1) - \psi_0(x_0),\end{aligned}$$

which is the left-hand side. It follows that $\psi_0(x_1)-\psi_0(x_0) = \xi(x_1)-\xi(x_0)$ for any potential $\xi$, and as $x_0, x_1 \in \mathsf{X}_0$ were arbitrary, Assumption 4.4 holds. □

We can now show the large deviations lower bound.

**Corollary 4.7.** *Let Assumption 4.4 hold. For any open set $U \subset \mathsf{X}_0 \times \mathsf{Y}_0$,*

$$\liminf_{\varepsilon\to 0} \varepsilon \log \pi_\varepsilon(U) \geq - \inf_{(x,y)\in U} I(x,y).$$

*Proof.* It suffices to show that given $(x, y) \in U$ and $\eta > 0$, there exists $r_0 > 0$ such that for all $r < r_0$,

$$\limsup_{\varepsilon\to 0} -\varepsilon \log \pi_\varepsilon(B_r(x,y)) \leq I(x,y) + \eta.$$

Let $\eta > 0$, pick any $(x', y') \in \mathsf{X}_0 \times \mathsf{Y}_0$ such that $(x', y), (x, y') \in \Gamma$, and set

$$\alpha := c(x,y) + c(x',y') - c(x,y') - c(x',y).$$

We have $I(x, y) < \infty$ and $I(x', y') < \infty$ by Proposition 4.5. For $r > 0$ small enough we may use Lemma 3.1 with $\delta' := \alpha+\eta/2$ and $B_r(x, y)\times B_r(x', y') \subset A_2(0, \delta')$ to obtain

$$\begin{aligned}\limsup\ &- \varepsilon \left[\log \pi_\varepsilon(B_r(x,y)) + \log \pi_\varepsilon(B_r(x',y')\right]\\ &= \limsup -\varepsilon \log \pi_\varepsilon^2\big(B_r(x,y) \times B_r(x',y')\big)\\ &\leq \alpha + \eta/2. \end{aligned}\tag{4.12}$$

On the other hand, for $r$ small enough, Lemma 4.1 yields as in (4.5) that

$$\liminf -\varepsilon \log \pi_\varepsilon(B_r(x',y')) \geq I(x',y') - \eta/2. \tag{4.13}$$

Using (4.12), then (4.7) and finally (4.13),

$$\begin{aligned}\limsup - \varepsilon \log \pi_\varepsilon(B_r(x,y)) &+ \liminf -\varepsilon \log \pi_\varepsilon(B_r(x',y')) \\ &\leq \limsup -\varepsilon \left[\log \pi_\varepsilon(B_r(x,y)) + \log \pi_\varepsilon(B_r(x',y')\right] \\ &\leq \alpha + \eta/2 \\ &= I(x,y) + I(x',y') + \eta/2 \\ &\leq I(x,y) + \liminf -\varepsilon \log \pi_\varepsilon(B_r(x',y')) + \eta\end{aligned}$$

and the claim follows. □

The following simple example shows that if both marginals supports are disconnected (and Assumption 4.4 is violated), $I$ may fail to be an upper bound for the rate function.

**Example 4.8** (Disconnected Supports)**.** Consider the normalized $2 \times 2$ assignment problem: $\mathsf{X} = \mathsf{Y} = \{1,2\}$ and $\mu = \nu = (\delta_{\{1\}} + \delta_{\{2\}})/2$. Here $\Pi(\mu,\nu)$ is the convex hull of the two couplings

$$\pi_* = (\delta_{\{(1,1)\}} + \delta_{\{(2,2)\}})/2, \quad \pi_0 = (\delta_{\{(1,2)\}} + \delta_{\{(2,1)\}})/2.$$

In particular, every $\pi \in \Pi(\mu,\nu)$ is symmetric: $\pi\{(1,2)\} = \pi\{(2,1)\}$. Consider a cost function $c$ with $c(1,1) = c(2,2) = 0$ and $c(1,2) + c(2,1) > 0$. Then $\pi_*$ is the unique optimal transport and we know that $\pi_\varepsilon \to \pi_*$. Let

$$r(i,j) := \lim_{\varepsilon \to 0} -\varepsilon \log \pi_\varepsilon(\{i,j\}) \tag{4.14}$$

be the exponential rate of convergence. Using Lemma 3.1 with

$$A = \{(1,2),(2,1)\} \subset A_2(\delta,\delta)$$

for $\delta := c(1,2) + c(2,1) > 0$ shows $r(1,2) + r(2,1) = \delta$. As $\pi_\varepsilon$ must be symmetric, we conclude that the true exponential rate is

$$r(1,2) = r(2,1) = \delta/2.$$

(A priori, it may not be obvious that the limit (4.14) exists, but a posteriori, this is justified as every subsequential limit leads to the same value.) On the other hand, the definition (4.3) of $I$ readily yields that $I \equiv 0$.

### 4.2.2 Bound via Regularity

In the remainder of the section we present an alternative approach to the large deviations lower bound which does not (directly) refer to potentials but instead employs a continuity condition for the limiting optimal transport $\pi_*$. We call a subset of a metric space arcwise connected if any two points are connected by a continuous curve of finite length.

**Assumption 4.9.** (a) $\Gamma = \operatorname{graph} T$ for a map $T : \mathsf{X}_0 \to \mathsf{Y}$.

(b) $\mathsf{X}_0$ is arcwise connected.

(c) The function $c(\cdot, T(\cdot))$ has the following continuity property: given a compact $K \subset \mathsf{X}_0$, we have uniformly over $x_1, x_2 \in K$ that

$$|c(x_1, T(x_1)) + c(x_2, T(x_2)) - c(x_1, T(x_2)) - c(x_2, T(x_1))| = o(d(x_1, x_2)). \tag{4.15}$$

As an example, consider $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$ with cost $c(x,y) = \|x-y\|^2/2$ and an optimal transport $\pi$ given by a continuous transport map $T$ on the arcwise connected support $\operatorname{spt}\mu$. Then Assumption 4.9 holds with $\mathsf{X}_0 = \operatorname{spt}\mu$, as (4.15) equals

$$|\langle x_1 - x_2, T(x_1) - T(x_2)\rangle| \le \|x_1 - x_2\| \|T(x_1) - T(x_2)\|$$

and $T$ is uniformly continuous on compact sets. General sufficient conditions for the continuity of $T$ can be found in [19, Theorem 1].

Next, we show how to establish the key half of (4.7) under Assumption 4.9.

**Lemma 4.10.** *Let Assumption 4.9 hold. If $(x,y), (x',y') \in \mathsf{X}_0 \times \mathsf{Y}_0$ are such that $(x',y), (x,y') \in \Gamma$, then*

$$I(x,y) + I(x',y') \ge c(x,y) + c(x',y') - c(x,y') - c(x',y). \tag{4.16}$$

*Proof.* Set $(x_1, y_1) := (x,y)$ and $(x_1', y_1') := (x', y')$. Let $k \ge 2$ and consider arbitrary $(x_i, y_i), (x_i', y_i') \in \Gamma$ for $2 \le i \le k$. The definition of $I$ yields that

$$I(x,y) + I(x',y') \ge \sum_{i=1}^{k} [c(x_i, y_i) + c(x_i', y_i')] - \sum_{i=1}^{k} [c(x_i, y_{i+1}) + c(x_i', y_{i+1}')].$$

This holds in particular for the choices $x_k := x'$ and $x_k' := x$, which entail that $y_k = T(x') = y$ and $y_k' = T(x) = y'$. Moreover, we have $y_i = T(x_i)$ and $y_i' = T(x_i')$ for $i \ge 2$. Separating the first term of the first sum and the last term of the second sum, we obtain that

$$\begin{aligned} I(x,y) + I(x',y') \ge{}& c(x,y) + c(x',y') - c(x,y') - c(x',y) \\ &+ \sum_{i=2}^{k} [c(x_i, y_i) + c(x_i', y_i')] - \sum_{i=1}^{k-1} [c(x_i, y_{i+1}) + c(x_i', y_{i+1}')]. \end{aligned}$$

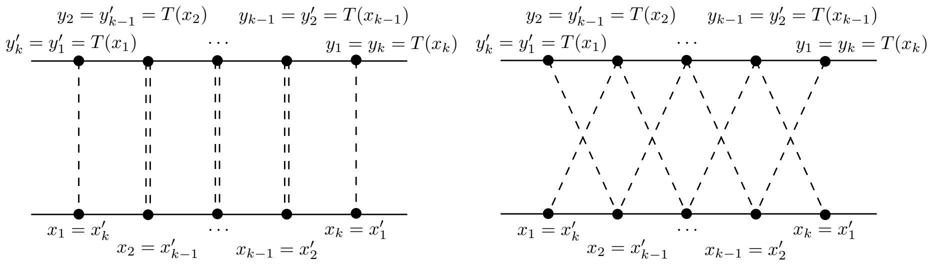

Figure 1: Schematic representation of the sums (4.17) (left) and (4.18) (right). Each dashed line stands for a term $c(\cdot,\cdot)$.

We further choose $x'_i := x_{k-i+1}$ for $i = 2, \dots, k-1$, which implies $y'_i = y_{k-i+1}$ for $i = 2, \dots, k-1$. Then the first sum can be rearranged as

$$\sum_{i=2}^{k} c(x_i, y_i) + c(x'_i, y'_i) = \sum_{i=1}^{k-1} c(x_i, T(x_i)) + c(x_{i+1}, T(x_{i+1})) \tag{4.17}$$

and the second sum can be rearranged as

$$\sum_{i=1}^{k-1} c(x_i, y_{i+1}) + c(x'_i, y'_{i+1}) = \sum_{i=1}^{k-1} c(x_i, T(x_{i+1})) + c(x_{i+1}, T(x_i)); \tag{4.18}$$

(These rearrangements are elementary if tedious; Figure 1 may be helpful to complete them.) In summary, we have

$$I(x, y) + I(x', y') \geq c(x, y) + c(x', y') - c(x, y') - c(x', y) + \Xi$$

where, always with the conventions $x_1 = x$ and $x_k = x'$,

$$\Xi := \sup_{k \geq 2} \sup_{x_2, \dots, x_{k-1} \in \operatorname{spt} \mu} \Xi_k \qquad \text{for}$$

$$\Xi_k := \sum_{i=1}^{k-1} c(x_i, T(x_i)) + c(x_{i+1}, T(x_{i+1})) - c(x_i, T(x_{i+1})) - c(x_{i+1}, T(x_i)).$$

It remains to show that given $\eta > 0$, we can achieve $\Xi_k \geq -\eta$ by a suitable choice of $k$ and $x_2, \dots, x_{k-1}$. Fix a continuous, rectifiable curve $\varphi : [0, 1] \to \mathsf{X}_0$ with $\varphi(0) = x$ and $\varphi(1) = x'$, and denote its length by $C$. For each $k \geq 2$ there exist $0 = t_1 < t_2 < \cdots < t_{k-1} < t_k = 1$ such that $x_i := \varphi(t_i)$ satisfy $d(x_i, x_{i+1}) \leq C/(k-1)$ for all $1 \leq i \leq k-1$. Applying

Assumption 4.9 on the compact set $\varphi([0,1])$, we have that

$$\sum_{i=1}^{k-1} |c(x_i, T(x_i)) + c(x_{i+1}, T(x_{i+1})) - c(x_i, T(x_{i+1})) - c(x_{i+1}, T(x_i))| \\ \leq (k-1)o(C/(k-1)) = o(1) \tag{4.19}$$

as $k \to \infty$. □

**Remark 4.11.** The preceding arguments can be generalized to handle certain discontinuities in $T$, even though at a discontinuity, (4.15) can only be expected with $o(1)$ rather than $o(d(x_1, x_2))$. Indeed, the conclusion of (4.19) still holds if for a bounded number of $i$'s, the term under the sum is only $o(1)$. For instance, this can be used to handle the case of semi-discrete transport with quadratic cost, where $\nu$ has finite support and hence the transport map is necessarily discontinuous.

**Corollary 4.12.** *Let Assumption 4.9 hold. For any open set $U \subset \mathsf{X}_0 \times \mathsf{Y}_0$,*

$$\liminf_{\varepsilon \to 0} \varepsilon \log \pi_\varepsilon(U) \geq - \inf_{(x,y) \in U} I(x,y).$$

*Proof.* The argument is similar to the proof of Corollary 4.7, using the inequality (4.16) instead of the equality (4.7). In the course of the argument one also obtains that (4.16) already implies (4.7). We omit the details. □

# 5 Positivity of the Rate Function

The aim of this section is to establish that, under certain conditions, $I(x,y)$ of (4.3) is strictly positive for $(x,y) \in \mathsf{X}_0 \times \mathsf{Y}_0$ outside the support $\Gamma$ of the limiting optimal transport $\pi_*$. In view of Corollary 4.7, this implies that that the mass of $\pi_\varepsilon$ around $(x,y)$ converges exponentially fast.

When both marginals are supported by finitely many points, it is known that exponential convergence holds for any cost function [17, 59]. We shall see that in the continuum case, such a statement must depend on the *geometry* of the cost. Throughout this section, we assume that $\mathsf{X} = \mathbb{R}^d$ (while $\mathsf{Y}$ is Polish). The cost $c$ is continuous and differentiable in $x$ with $\nabla_x c$ continuous, and there exists an optimal transport $\pi_*$ as in (4.1). We recall the *twist condition* of optimal transport (e.g., [58, p. 234]) which requires that $\nabla_x c(x, \cdot)$ be injective; it holds in particular for the quadratic cost. Example 5.7 below shows that $I$ may vanish on a set of large measure $\mu \otimes \nu$ when the twist condition does not hold.

Similarly to the preceding section, we present a primal and a dual approach. The direct approach proceeds as follows. Given $(x,y) \notin \Gamma$, we use the geometry of $c$ and regularity of the optimal transport to find an auxiliary pair $(\tilde{x}, \tilde{y}) \in \Gamma$ such that $c(x,y) - c(\tilde{x}, y) + c(\tilde{x}, \tilde{y}) - c(x, \tilde{y}) > 0$. Then, the definition (4.3) of $I$ (with $k = 2$) shows that $I(x,y) > 0$. The following is one possible implementation.

**Lemma 5.1.** *Fix $(x, y') \in \Gamma$ and $y \in \mathsf{Y}$. Suppose that*

$$v := \nabla_x c(x,y) - \nabla_x c(x,y') \neq 0 \tag{5.1}$$

*and that there exist $(x_n, y_n) \in \Gamma$ such that $(x_n, y_n) \to (x, y')$ and*

$$\liminf \cos \alpha_n > 0 \quad \textit{for} \quad \alpha_n := \angle(v, x - x_n). \tag{5.2}$$

*Then $I(x,y) > 0$.*

*Proof.* Set $\Delta(x, y', y_n) := \nabla_x c(x,y') - \nabla_x c(x, y_n)$; then $\Delta(x, y', y_n) \to 0$ as $d(y', y_n) \to 0$ and we have

$$\begin{aligned}
\delta_n :&= c(x,y) - c(x_n, y) + c(x_n, y_n) - c(x, y_n) \\
&= \langle \nabla_x c(x,y) - \nabla_x c(x, y_n), x - x_n \rangle + o(\|x - x_n\|) \\
&= \langle \nabla_x c(x,y) - \nabla_x c(x,y') + \Delta(x, y', y_n), x - x_n \rangle + o(\|x - x_n\|) \\
&= \langle v, x - x_n \rangle + o(\|x - x_n\|) + O(d(y', y_n))\|x - x_n\| \\
&= \cos(\alpha_n)\|v\|\|x - x_n\| + o(\|x - x_n\|) + O(d(y', y_n))\|x - x_n\|.
\end{aligned}$$

As $v \neq 0$ and $\liminf_n \cos \alpha_n > 0$, it follows that $\delta_n > 0$ for $n$ large enough. Fix such an $n$, then choosing $k = 2$ and $(x_2, y_2) := (x_n, y_n)$ in (4.3) shows that $I(x,y) \geq \delta_n > 0$. □

Recall the notation $\mathsf{X}_0 = \operatorname{proj}_{\mathsf{X}} \Gamma$ and $\Gamma = \operatorname{spt} \pi_*$. If $x$ is interior in $\mathsf{X}_0$, we can choose auxiliary points in any direction from $x$ and Lemma 5.1 yields a positivity result for $I(x,y)$ as follows.

**Lemma 5.2.** *Let $x \in \operatorname{int} \mathsf{X}_0$ and $y \in \mathsf{Y}$. Let $\pi_*$ be given by a transport map $T$ which is continuous at $x$. If $\nabla_x c(x,y) - \nabla_x c(x, T(x)) \neq 0$, then $I(x,y) > 0$.*

*Proof.* For $n$ large we can uniquely define a point $x_n \in \partial B_{1/n}(x) \subset \mathsf{X}_0$ by the requirement that $x - x_n$ be parallel to $v := \nabla_x c(x,y) - \nabla_x c(x, T(x))$ (here $\partial B$ denotes the boundary). Then $\cos \alpha_n = 1$ in the notation of (5.2) and we conclude using Lemma 5.1 with $(x, y') = (x, T(x))$ and $(x_n, y_n) = (x_n, T(x_n))$. □

Sufficient conditions for the continuity (and higher regularity) of the transport map have been studied extensively; see [58, Section 12] for an overview of now-classical results and, among others, [19] for recent results including unbounded domains.

The situation is more delicate if $x$ is a boundary point of $\mathsf{X}_0$ or a point of discontinuity of the transport map, as that restricts the viable choices for approximating sequences. We provide some examples of possible results; for simplicity of exposition, they are stated for the quadratic cost on $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$. The extension of such arguments to a general class of cost functions is discussed in Appendix C.

**Lemma 5.3.** *Let $c(x, y) = \|x - y\|^2$, let $\mathsf{X}_0$ be strictly convex*[2] *and consider $(x, y) \in (\mathsf{X}_0 \times \mathsf{Y}_0) \setminus \Gamma$ with $x \in \partial \mathsf{X}_0$. Suppose that $\pi_*$ is given by a transport map $T$ which is continuous on a neighborhood $B_r(x) \cap \mathsf{X}_0$ for some $r > 0$. Then $I(x, y) > 0$.*

*Proof.* The main step is to find a point $x'' \in \mathsf{X}_0$ such that

$$\langle v, x - x'' \rangle > 0. \tag{5.3}$$

Once that is achieved, we may choose a sequence $x_n \to x$ in the open segment $(x'', x)$ which is contained in $\operatorname{int} \mathsf{X}_0$ due to strict convexity. As $(x_n, T(x_n)) \to (x, T(x))$ by continuity and $\alpha_n = \angle(v, x - x'')$ for all $n$, we conclude by Lemma 5.1 with $(x, y') := (x, T(x))$.

To find $x''$ satisfying (5.3), we first fix $x' \in \mathsf{X}_0$ such that $(x', y) \in \Gamma$. As $c$ is quadratic, we have $v = y' - y$ in (5.1) and the cyclical monotonicity of $\Gamma$ yields

$$\langle v, x - x' \rangle = \langle y' - y, x - x' \rangle \geq 0.$$

If this inequality is strict, we choose $x'' := x'$. Whereas if $\langle v, x - x' \rangle = 0$, we consider the mid-point $\bar{x} = (x' - x)/2$ which satisfies $\bar{x} \in \operatorname{int} \mathsf{X}_0$ by strict convexity as well as $\langle v, x - \bar{x} \rangle = 0$. After choosing $\rho > 0$ small enough such that $\partial B_\rho(\bar{x}) \subset \mathsf{X}_0$, we can find a point $x'' \in \partial B_\rho(\bar{x}) \subset \mathsf{X}_0$ such that $\langle v, x - x'' \rangle > 0$, completing the proof. □

Next, we illustrate the dual approach in a problem with *discontinuous* optimal transport map. For the remainder of the section, we assume that there exists a Kantorovich potential $\psi$ such that

$$I(x, y) = c(x, y) - \psi^c(y) + \psi(x), \quad (x, y) \in \mathsf{X}_0 \times \mathsf{Y}_0. \tag{5.4}$$

As seen in Proposition 4.5, a sufficient condition is Assumption 4.4 (uniqueness of potentials). If we assume that $\mu \sim \mathcal{L}^d$ on its support, the quadratic cost and the convexity condition in the below results already guarantee that Assumption 4.4 holds; cf. Proposition B.2. The relevance of (5.4) is that it yields the representation

$$\{I = 0\} \cap (\mathsf{X}_0 \times \mathsf{Y}_0) = \partial_c \psi \cap (\mathsf{X}_0 \times \mathsf{Y}_0), \tag{5.5}$$

so that our question regarding exponential convergence can be phrased as:

[2] In the sense that the open segment $(x, x')$ is contained in $\operatorname{int} \mathsf{X}_0$ for distinct $x, x' \in \mathsf{X}_0$.

does $\Gamma$ fill the entire set $\partial_c\psi \cap (\mathsf{X}_0 \times \mathsf{Y}_0)$?

The intersection with $\mathsf{X}_0 \times \mathsf{Y}_0$ is crucial to avoid a negative answer in many cases with discontinuous transport (see also the proof of Proposition 5.5 below). On the other hand, the intersection is justified because the interpretation of $I$ as rate of convergence is meaningless outside $\operatorname{spt}\pi_\varepsilon$.

We first state the following continuation argument similar to Lemma 5.3.

**Lemma 5.4.** *Let $c(x,y) = \|x-y\|^2$, let $\mathsf{X}_0$ be strictly convex and consider $(x,y) \in (\mathsf{X}_0 \times \mathsf{Y}_0) \setminus \Gamma$ with $x \in \partial\mathsf{X}_0$. Suppose that $I(\tilde{x},y) > 0$ for all $\tilde{x} \in \operatorname{int}\mathsf{X}_0 \cap B_r(x)$, for some $r>0$. Then $I(x,y)>0$.*

*Proof.* We may state the proof with the equivalent cost $c(x,y) = -\langle x,y\rangle/2$, so that the notions of $c$-convex analysis and convex analysis coincide. Suppose for contradiction that $I(x,y)=0$. Fix $x' \in \mathsf{X}_0$ such that $(x',y) \in \Gamma$ and denote $\phi := -\psi^c$ for $\psi$ as in (5.4), then both $x$ and $x'$ are in the set

$$\{I(\cdot,y)=0\} = \partial_c\phi(y) = \partial\phi(y),$$

where $\partial\phi(y)$ denotes the subdifferential of the convex function $\phi$ in the usual sense. The latter set being convex, it must include the whole segment $[x,x']$, meaning that $I(\tilde{x},y)=0$ for all $\tilde{x} \in [x,x']$. The interior of the segment is included in $\operatorname{int}\mathsf{X}_0$ by strict convexity, contradicting the hypothesis. □

**Proposition 5.5** (Semidiscrete Transport)**.** *Let $c(x,y) = \|x-y\|^2$ on $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$, let $\mathsf{X}_0$ be strictly convex, let $\mu \ll \mathcal{L}^d$ and let $\operatorname{spt}\nu$ be at most countable, with no accumulation points. Then $\{I=0\} \cap (\mathsf{X}_0 \times \mathsf{Y}_0) = \Gamma$.*

*Proof.* Again, we may state the proof with the equivalent cost $c(x,y) = -\langle x,y\rangle/2$. Let $(x,y) \in \mathsf{X}_0 \times \mathsf{Y}_0$. In view of Lemma 5.4, it suffices to treat the case $x \in \operatorname{int}\mathsf{X}_0$. Denote by $\operatorname{dom}\nabla\psi$ the set of points where $\psi$ is differentiable and assume that $I(x,y)=0$; that is, $y \in \partial_c\psi(x) = \partial\psi(x)$. The (ordinary) subdifferential $\partial\psi(x)$ equals $\{\nabla\psi(x)\}$ if $x \in \operatorname{dom}\nabla\psi$, whereas in general, it can be described (cf. [55, Theorem 25.6, p. 246]) as the closed convex hull of

$$S(x) = \Big\{ \lim_{n\to\infty} \nabla\psi(x_n) : x_n \to x,\, x_n \in \operatorname{dom}\nabla\psi,\, \lim_{n\to\infty}\nabla\psi(x_n) \text{ exists}\Big\}. \tag{5.6}$$

*Case 1: $x \in \operatorname{dom}\nabla\psi$.* As $\Gamma \subset \partial\psi$ and $\partial\psi(x)$ is a singleton, it follows that $(x,y) = (x,\nabla\psi(x)) \in \Gamma$.

*Case 2: $y \in S(x)$.* Let $x_n \to x$ be as in (5.6). Recalling that $x \in \operatorname{int}\mathsf{X}_0$, we have $x_n \in \mathsf{X}_0$ for $n$ large. Thus $(x_n,\nabla\psi(x_n)) \in \Gamma$ by Case 1 and closedness entails that the limit $(x,y)$ pertains to $\Gamma$ as well.

*Case 3: $y \in \partial\psi(x) \setminus S(x)$.* We shall show that this case does not occur. As a first step, we argue that

$$\partial\psi(x) = \operatorname{conv} S(x) \tag{5.7}$$

in the present context (without taking closure). As $x \in \operatorname{int} \mathsf{X}_0 \subset \operatorname{int}\{\psi < \infty\}$, the subdifferential $\partial\psi(x)$ is bounded [55, Theorem 23.4, p. 217]. Let $U$ be a bounded neighborhood of $\partial\psi(x)$. The discreteness assumption on $\operatorname{spt}\nu$ entails that $U \cap \mathsf{Y}_0$ is a finite set (and that $\mathsf{Y}_0 = \operatorname{spt}\nu$). Let $x_n \to x$ be as in (5.6). For $x_n$ close to $x$ we have $\nabla\psi(x_n) \in U$, but also $\nabla\psi(x_n) \in \mathsf{Y}_0$ by Case 1. As a result, the set $S(x)$ of limits is finite. In particular, its convex hull is already closed, and (5.7) follows.

Now let $y \in \partial\psi(x) \setminus S(x)$. By (5.7), $y$ is a nontrivial convex combination $y = \sum_{i=1}^k \theta_i y_i$ for some distinct $y_i \in S(x)$ and $\theta_i \in (0,1)$ with $\sum \theta_i = 1$. Let $\phi := -\psi^c$ (which is the Legendre–Fenchel transform of $\psi$ in this context) and $x' \in \partial\phi(y)$. Then cyclical monotonicity of $\partial\phi$ implies $\langle x' - x, y - y_i\rangle \geq 0$ for all $i$ and as

$$\sum_{i=1}^k \theta_i \langle x' - x, y - y_i\rangle = \langle x' - x, 0\rangle = 0, \tag{5.8}$$

it follows that $\langle x' - x, y - y_i\rangle = 0$ for all $i$. That is, we have

$$\partial\phi(y) - \{x\} \;\perp\; y - y_i \quad \text{for all} \quad 1 \leq i \leq k,$$

which implies in particular $\dim \partial\phi(y) < d$. On the other hand, $\nu(\{y\}) > 0$ by the discreteness of $\mathsf{Y}_0$. Thus $\mu(\partial\phi(y)) = \nu(\{y\}) > 0$, contradicting that $\mu \ll \mathcal{L}^d$ does not charge lower dimensional sets. This shows that Case 3 does not occur and completes the proof. □

The preceding arguments can be extended to a class of cost functions satisfying a Ma-Trudinger-Wang condition. This is detailed in Appendix C.

**Proposition 5.6.** *After replacing convexity by c-convexity, Lemma 5.4 and Proposition 5.5 extend to cost functions c satisfying Assumption C.1.*

We conclude with a simple example illustrating the relevance of the twist condition. Here, $\nabla_x c(x,y)$ vanishes below the diagonal, so that the condition fails, and indeed the convergence $\pi_\varepsilon \to \pi_*$ is sub-exponential in that region.

**Example 5.7** (No Twist)**.** Consider $\mathsf{X} = \mathsf{Y} = \mathbb{R}$ with identical marginals $\mu = \nu$ having support $[0,1]$ and the cost function

$$c(x,y) = \begin{cases} (y-x)^2, & y \geq x, \\ 0, & y < x. \end{cases}$$

As $\nabla_x c(x,y) = 0$ for all $y < x$, this cost does not satisfy the twist condition. Clearly there is a unique optimal transport $\pi_* \in \Pi(\mu,\nu)$, given by the Monge map $T(x) = x$. Its support is $\Gamma = \{(x,x) : 0 \leq x \leq 1\}$ and one can check by direct calculation based on (4.3) that $I = c$ on $\mathsf{X}_0 \times \mathsf{Y}_0 = [0,1]^2$. Assumption 4.9 is readily verified, hence Corollary 4.12 shows that $I$ is indeed the rate function in this context. We can obtain the same conclusion from Corollary 4.7, at least if we also suppose that $\mu$ is equivalent to the Lebesgue measure on $[0,1]$: then, Proposition B.2 shows that Assumption 4.4 holds. Or as a third option, we may verify directly that $I$ satisfies (4.7), and then conclude as in the proof of Corollary 4.7. In any event, we see that $I = 0$ on $\{y < x\}$, indicating sub-exponential decay of the weight of $\pi_\varepsilon$.

# A Cyclical Invariance and Factorization

In this section we detail some classical facts about static Schrödinger bridges as well as the proof of Proposition 2.2. Let $(\mathsf{X},\mu)$ and $(\mathsf{Y},\nu)$ be Polish probability spaces; as before, we denote by $\Pi(\mu,\nu)$ the set of couplings.

**Proposition A.1.** *Let $R$ be a probability measure on $\mathsf{X}\times\mathsf{Y}$ and suppose that*

$$\textit{there exists } \pi \in \Pi(\mu,\nu) \textit{ with } H(\pi|R) < \infty. \tag{A.1}$$

*Then there is a unique minimizer $\pi^* \in \Pi(\mu,\nu)$ for $\inf_{\pi\in\Pi(\mu,\nu)} H(\pi|R)$. Assume in addition that $R \sim \mu\otimes\nu$. Then $\pi_* \sim \mu\otimes\nu$ and there exist measurable functions $Z : \mathsf{X}\times\mathsf{Y} \to (0,\infty)$, $f : \mathsf{X} \to (0,\infty)$, $g : \mathsf{Y} \to (0,\infty)$ such that*

$$Z(x,y) = f(x)g(y), \quad (x,y) \in \mathsf{X}\times\mathsf{Y} \tag{A.2}$$

*and $Z$ is a version of the Radon–Nikodym density $d\pi^*/dR$. Conversely, if $\pi \in \Pi(\mu,\nu)$ and a version of its density has the form* (A.2) *on a set of full $\mu\otimes\nu$-measure, where $f$ and $g$ are arbitrary $[-\infty,\infty]$-valued functions, then $\pi = \pi^*$.*

*The uniqueness result also holds without Assumption* (A.1)*, if stated as follows. Let $\pi,\pi' \in \Pi(\mu,\nu)$ and $\pi,\pi',R \sim \mu\otimes\nu$. If versions of $d\pi/dR$ and $d\pi'/dR$ both admit factorizations as above, then $\pi = \pi'$.*

*Proof.* The result under (A.1) can be found in [48, Theorem 2.1] in the stated form (where we do not assume a priori that one can choose $\pi \sim R$ in (A.1)).

For the final generalization on the uniqueness claim, let $\pi,\pi'$ be as stated and note that a version of the density $d\pi/d\pi'$ then admits a factorization. We consider $\pi'$ as an auxiliary reference measure, instead of $R$. Then the analogue of (A.1) holds as $\pi'$ is itself a coupling and clearly $\pi'$ is the unique minimizer of $H(\cdot|\pi')$. We can now apply the above results. □

We mention that the existence and uniqueness of $\pi_*$ are due to [20], and that the factorization of the density and its measurability are delicate in general (see [11, 12, 29, 57], among others) but less so under our condition that $R \sim \mu \otimes \nu$. An insightful approach with a direct construction of the factorization was recently proposed in [2]; it yields similar results for the entropic function $h(x) = x \log x$ considered here but also allows for a generalization to nonconvex penalties $h$. In addition, it portrays what we called cyclical invariance as the cyclical monotonicity of an optimal transport problem arising from the linearization of the static Schrödinger bridge problem. Another recent work, [3], uses Markovian methods to obtain a generalized factorization result for Schrödinger bridges with additional constraints.

*Proof of Proposition 2.2.* Recall the definition (2.5) of $R$ and note that $R \sim P = \mu \otimes \nu$. The entropic optimal transport problem (2.2) can we rewritten as $\inf_{\pi \in \Pi(\mu,\nu)} \varepsilon H(\pi|R)$, putting it in the realm of Proposition A.1. Similarly, (2.3) is equivalent to (A.1). Let $Z$ be as in (A.2), then (2.6) follows, and hence also (2.4).

Conversely, if $\pi \in \Pi(\mu,\nu)$ is cyclically invariant, then $\pi \sim P$ and (2.6) holds for its density $Z$. Fix an arbitrary $x_0 \in \mathsf{X}$ and note that $f(x) := Z(x,y)/Z(x_0,y)$ is independent of $y$ due to (2.6) with $k = 2$. Setting $g(y) = Z(x_0,y)/f(x_0)$ then yields the (measurable) factorization $Z(x,y) = f(x)g(y)$, and we conclude by Proposition A.1. Alternately, the existence of a factorization can be deduced from (2.6) by the general result of [11, Theorem 3.3]. □

**Remark A.2.** The above proof shows that if the cyclical invariance condition (2.4) holds for $k = 2$, then it already holds for arbitrary $k \geq 2$.

# B Uniqueness of Potentials

**Definition B.1.** Let $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ and $\Lambda \subset \mathsf{X}$. We say that *uniqueness of potentials holds on* $\Lambda$ if for any $c$-convex functions $\psi_1, \psi_2$ on $\mathsf{X}$ with $\Gamma \subset \partial_c \psi_i$, it holds that $\psi_1 - \psi_2$ is constant on $\Lambda$.

We detail two classes of optimal transport problems where uniqueness of potentials holds. Connectedness of at least one marginal support is essential—uniqueness fails even for the simplest discrete problem, $\mu = \nu = (\delta_{\{1\}} + \delta_{\{2\}})/2$ with cost $c(\{i\},\{j\}) = \mathbf{1}_{i \neq j}$.

**Proposition B.2.** *Let* $\mathsf{X} = \mathbb{R}^d$ *and* $\mu \sim \mathcal{L}^d$ *on* $\operatorname{spt} \mu$, *where* $\mathcal{L}^d(\partial \operatorname{spt} \mu) = 0$ *and* $\operatorname{int} \operatorname{spt} \mu$ *is connected. Let* $\Gamma = \operatorname{spt} \pi$ *where* $\pi \in \Pi(\mu,\nu)$ *is an optimal transport for the continuous cost function c.*

*(a)* Lipschitz cost: *Suppose that*

*$c(\cdot, y)$ is differentiable for all $y$, and locally Lipschitz uniformly in $y$.*

*Then uniqueness of potentials holds on* $\operatorname{spt}\mu$, *and in particular on* $\operatorname{proj}_{\mathsf{X}}\Gamma$.

*(b)* Convex, superlinear cost: *Let $\mathsf{Y} = \mathbb{R}^d$ and $c(x, y) = h(y - x)$, where*

*(i) $h : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable,*

*(ii) $h$ has superlinear growth: $h(x)/\|x\| \to \infty$ whenever $\|x\| \to \infty$,*

*(iii) given $r < \infty$ and $\theta \in (0, \pi)$, and for $p \in \mathbb{R}^d$ sufficiently far from the origin, there is a cone of the form $\{x \in \mathbb{R}^d : \|x - p\| \, \|z\| \cos(\theta/2) \leq \langle z, x - p\rangle \leq r\|z\|\}$ for some $z \in \mathbb{R}^d \setminus \{0\}$ on which $h$ assumes its maximum at $p$.*

*Then uniqueness of potentials holds on* $\operatorname{proj}_{\mathsf{X}}\Gamma$.

**Remark B.3.** (a) If $c \in C^1(\mathbb{R}^d \times \mathbb{R}^{d'})$ and $\nu$ is compactly supported, we can always change $c$ outside a neighborhood of $\operatorname{spt}\mu \times \operatorname{spt}\nu$ to satisfy the condition of (a), without affecting the set of optimal transports.

(b) The convex cost with superlinear growth is essentially the well-known setting of Gangbo and McCann [30]; cf. their hypotheses (H2)–(H4). The technical condition (iii) is implied by (ii) in the radial case $h(x) = \tilde{h}(\|x\|)$; in particular, all the conditions are satisfied for $c(x, y) = \|y - x\|^p$ with $p \in (1, \infty)$. In contrast to the main result of [30], $h$ is not assumed to be strictly convex—strictness is required for uniqueness of optimal transports, but not for uniqueness of potentials. For instance, the "parabola with an affine piece," given by $h(x) = \tilde{h}(\|x\|)$ with $\tilde{h}(t) = t^2\mathbf{1}_{[0,1]} + (2t - 1)\mathbf{1}_{(1,2)} + (t^2 - 2t + 3)\mathbf{1}_{[2,\infty)}$, satisfies all the assumptions in (b). The affine piece will lead to non-uniqueness of optimal transports for a large class of marginals in the one-dimensional case.

(c) Dual uniqueness may fail if $c$ is not differentiable. For $c(x, y) = |y - x|$ on $\mathbb{R} \times \mathbb{R}$, the $c$-convex functions are exactly the 1-Lipschitz functions. If $\mu = \nu$ is the Lebesgue measure on $[0, 1]$, the identical transport $\pi$ is optimal and any 1-Lipschitz function $\psi$ satisfies $\Gamma = \{(x, x) : x \in [0, 1]\} \subset \partial_c\psi$.

The proof of the proposition is based on the following standard consideration (e.g., [30, Lemma 3.1]).

**Lemma B.4.** *Let $\Gamma \subset \mathsf{X} \times \mathsf{Y}$ and let $\psi, \phi$ be $\overline{\mathbb{R}}$-valued functions such that $\phi(y) - \psi(x) \leq c(x, y)$ on $\mathsf{X} \times \mathsf{Y}$ and $\phi(y) - \psi(x) = c(x, y)$ on $\Gamma$. If $\mathsf{X} = \mathbb{R}^d$ and $(x, y) \in \Gamma$ are such that $\psi$ and $c(\cdot, y)$ are differentiable at $x$, then $\nabla\psi(x) = -\nabla_x c(x, y)$. In particular, if $c(\cdot, y)$ is differentiable for all $y \in \mathsf{Y}$, then $\nabla\psi(x)$ is uniquely determined for $x \in$* $\operatorname{proj}_{\mathsf{X}}\Gamma \cap \operatorname{dom}\nabla\psi$.

*Proof.* Let $(x, y) \in \Gamma$ be as stated. Then

$$\begin{aligned}\psi(x) + \nabla\psi(x) \cdot h + o(h) = \psi(x+h) &\geq \phi(y) - c(x+h, y) \\ &= \psi(x) + c(x,y) - c(x+h,y) \\ &= \psi(x) - \nabla_x c(x,y) \cdot h + o(h)\end{aligned}$$

and hence $\nabla\psi(x) = -\nabla_x c(x,y)$ as the direction of $h$ is arbitrary. □

**Lemma B.5.** *Let* $\Gamma = \operatorname{spt}\pi$ *for some* $\pi \in \Pi(\mu, \nu)$*. Then* $\operatorname{spt}\mu = \overline{\operatorname{proj}_{\mathsf{X}}\Gamma}$.

*Proof.* Let $(x, y) \in \Gamma$, then $\mu(B_r(x)) = \pi(B_r(x) \times \mathsf{Y}) > 0$ for all $r > 0$. This shows $\operatorname{proj}_{\mathsf{X}}\Gamma \subset \operatorname{spt}\mu$. Let $x \in \operatorname{spt}\mu$. As $\mu(B_r(x)) > 0$, there must be some $x' \in B_r(x)$ with $x' \in \operatorname{proj}_{\mathsf{X}}\Gamma$, and this holds for all $r > 0$. Hence, $\operatorname{spt}\mu \subset \overline{\operatorname{proj}_{\mathsf{X}}\Gamma}$. □

*Proof of Proposition B.2.* We denote by $\operatorname{dom}\psi$ the set where a function $\psi$ is finite and by $\operatorname{dom}\nabla\psi$ the subset where it is differentiable.

(a) Let $\psi$ be a $c$-convex function on $\mathsf{X} = \mathbb{R}^d$ with $\Gamma \subset \partial_c\psi$. The local Lipschitz bound of $c(\cdot, y)$ implies the same bound for $\psi$. In particular, $\psi$ is continuous and $\mathcal{L}^d$-a.e. differentiable on $\operatorname{dom}\psi = \mathbb{R}^d$. The coupling property guarantees that $\operatorname{proj}_{\mathsf{X}}\Gamma \subset \operatorname{spt}\mu$ has full $\mu$-measure, hence also full $\mathcal{L}^d$-measure. It follows that $\Lambda := \operatorname{dom}\nabla\psi \cap \operatorname{proj}_{\mathsf{X}}\Gamma \subset \operatorname{spt}\mu$ has full $\mathcal{L}^d$-measure, and $\nabla\psi$ is uniquely determined on $\Lambda$ by Lemma B.4. As $\psi$ is locally Lipschitz and $\operatorname{int}\operatorname{spt}\mu$ is open and connected, this implies that $\psi$ is uniquely determined (up to constant) on $\operatorname{int}\operatorname{spt}\mu$ (see, e.g., [52, Formula 2]). By continuity on $\mathbb{R}^d$, it is also determined on the closure, which equals $\operatorname{spt}\mu$ due to $\mathcal{L}^d(\partial\operatorname{spt}\mu) = 0$.

(b) In this setting, the local Lipschitz property will only hold within $\operatorname{int}\operatorname{spt}\mu$ and $\psi$ need not be continuous (or even finite) up to the boundary. As we require uniqueness at all (rather than almost all) points $x$, we argue the boundary case in a second step.

*Step 1.* We first show that uniqueness of potentials holds on $\operatorname{int}\operatorname{spt}\mu$. It is proved in [30, Proposition C.3 and Corollary C.5] that for any $c$-convex function $\psi$ there is a convex set $K$ with $\operatorname{int}K \subset \operatorname{dom}\psi \subset K$ and that $\psi$ is locally Lipschitz (hence $\mathcal{L}^d$-a.e. differentiable) within $\operatorname{int}\operatorname{dom}\psi$. By convexity, $\operatorname{int}\overline{K} = \operatorname{int}K = \operatorname{int}\operatorname{dom}\psi$. If $\Gamma \subset \partial_c\psi$, then $\operatorname{proj}_{\mathsf{X}}\Gamma \subset \operatorname{dom}\psi$ and hence $\operatorname{spt}\mu = \overline{\operatorname{proj}_{\mathsf{X}}\Gamma} \subset \overline{\operatorname{dom}\psi} \subset \overline{K}$, showing that

$$\operatorname{int}\operatorname{spt}\mu \subset \operatorname{int}\overline{K} = \operatorname{int}\operatorname{dom}\psi.$$

It follows that $\psi$ is locally Lipschitz and $\mathcal{L}^d$-a.e. differentiable on $\operatorname{int}\operatorname{spt}\mu$. On the other hand, $\operatorname{proj}_{\mathsf{X}}\Gamma$ has full $\mu$-measure in $\operatorname{int}\operatorname{spt}\mu$ by the coupling property, hence also full $\mathcal{L}^d$-measure. Thus $\Lambda := \operatorname{dom}\nabla\psi \cap \operatorname{proj}_{\mathsf{X}}\Gamma$ has full $\mathcal{L}^d$-measure within $\operatorname{int}\operatorname{spt}\mu$ and we conclude as in (a).

*Step 2.* Define $\mathsf{X}_1 := \operatorname{proj}_{\mathsf{X}} \Gamma \cap \operatorname{int} \operatorname{spt} \mu$. Then

$$\Gamma = \overline{\Gamma}_1 \quad \text{for} \quad \Gamma_1 := \{(x,y) : x \in \mathsf{X}_1,\, y \in \Gamma_x\}, \tag{B.1}$$

where $\Gamma_x$ denotes the section $\{y \in \mathsf{Y} : (x,y) \in \Gamma\}$. Indeed, $\mu(\mathsf{X}_1) = 1$ as stated in Step 1, which implies $\pi(\Gamma_1) = 1$ and hence $\Gamma \subset \overline{\Gamma}_1$ by the definition of $\Gamma = \operatorname{spt} \pi$. Conversely, $\Gamma_1 \subset \Gamma$ is clear, and then $\overline{\Gamma}_1 \subset \Gamma$ by closedness.

Fix $(x,y) \in \Gamma$. By (B.1) we can find $(x_n, y_n) \in \Gamma_1$ with $(x_n, y_n) \to (x,y)$ and in particular

$$\psi^c(y) - \psi(x) = c(x,y) = \lim c(x_n, y_n) = \lim\,[\psi^c(y_n) - \psi(x_n)].$$

The $c$-convex functions $-\psi^c$ and $\psi$ are lower semicontinuous thanks to the continuity of $c$, so that $\psi^c(y) \geq \limsup \psi^c(y_n)$ and $-\psi(x) \geq \limsup -\psi(x_n)$. Together, it follows that $\psi^c(y) = \lim \psi^c(y_n)$ and $\psi(x) = \lim \psi(x_n)$. As $x_n \in \operatorname{int} \operatorname{spt} \mu$, we know from Step 1 that $\psi(x_n)$ is uniquely determined, and then so is $\psi(x)$. □

# C Proof of Proposition 5.6

In this section, we discuss how to extend Proposition 5.5 to a general class of cost functions $c$ satisfying the Ma-Trudinger-Wang condition "(Aw)" introduced in [43]; we use Loeper's equivalent geometric characterization [42] to generalize from the quadratic case. We recall that the dual representation (5.4) of $I$ has been assumed.

A number of terms from $c$-convex analysis are needed. For ease of reference, we (mostly) follow the notation of [42], whose Section 2 also provides an excellent introduction to the notions used below. Consider a $C^1$ function $c(x,y)$ on the product of two domains $\Omega, \Omega' \subset \mathbb{R}^d$ and suppose that $c$ satisfies the twist condition in both variables; i.e., $\nabla_x c(x,\cdot)$ and $\nabla_y c(\cdot,y)$ are injective. Given $x \in \Omega$, the *$c$-exponential* map $\mathfrak{T}_x$ is defined by $\mathfrak{T}_x = -\nabla_x c(x,\cdot)^{-1}$. A *$c$-segment wrt. $x$* is the image of a segment (in the usual sense) under the map $\mathfrak{T}_x$. The *$c$-segment of $y_1, y_2 \in \Omega'$ wrt. $x$* is the image of the segment joining $-\nabla_x c(x,y_1)$ and $-\nabla_x c(x,y_2)$ under $\mathfrak{T}_x$. The set $\Omega'$ is *$c$-convex wrt. $\Omega$* if the $c$-segment of $y_1, y_2$ wrt. $x$ is contained in $\Omega'$ for all $y_1, y_2 \in \Omega'$ and $x \in \Omega$, or equivalently, if $-\nabla_x c(x,\Omega')$ is convex for $x \in \Omega$. *Strict* $c$-convexity means that, in addition, the interior of the $c$-segment is in the interior of $\Omega'$. A proper function $\psi : \Omega \to \mathbb{R} \cup \{+\infty\}$ is *$c$-convex* if if there exists $\zeta : \Omega' \to [-\infty, \infty]$ such that $\psi(x) = \sup_{y \in \Omega'}[\zeta(y) - c(x,y)]$. The *$c$-transform* of $\psi$ is defined by $\psi^c(y) := \inf_{x \in \Omega}[c(x,y) + \psi(x)]$ for $y \in \Omega'$ and its *$c$-subdifferential* at $x$ is the set $\partial_c \psi(x) = \{y \in \Omega' : \psi^c(y) - \psi(x) = c(x,y)\}$.

The function $\psi$ is *semiconvex* if it is the sum of a convex function and a function of class $C^{1,1}$. Its (ordinary) subdifferential $\partial\psi(x)$ at $x \in \Omega$ is

$$\partial\psi(x) := \{y \in \mathbb{R}^d : \psi(x') \geq \psi(x) + \langle y, x' - x\rangle + o(\|x - x'\|),\ x' \in \Omega\}.$$

Clearly $\partial\psi(x)$ is convex. Moreover, it coincides with the subdifferential of convex analysis if $\psi$ is convex, and it satisfies an analogue of the cyclical monotonicity of convex analysis: adding up the defining inequalities shows

$$\langle y - y', x - x'\rangle \geq o(\|x - x'\|) \quad \text{for} \quad y \in \partial\psi(x), \quad y' \in \partial\psi(x'). \tag{C.1}$$

We shall use analogous notation for functions on $\Omega'$ instead of $\Omega$ (a minor abuse of notation since $c$ is then used with its variables exchanged).

**Assumption C.1.** Let $\Omega, \Omega'$ be domains in $\mathbb{R}^d$ with $\mathsf{X}_0 \subset \Omega$ and $\mathsf{Y}_0 \subset \Omega'$, and let $c \in C^1$ satisfy the twist condition in both variables. Moreover, let $\mathsf{X}_0$ be strictly $c$-convex wrt. $\mathsf{Y}_0$ and let $\Omega'$ be $c$-convex wrt. $\Omega$. Finally, we assume that any $c$-convex function $\psi$ on $\Omega$ is locally semiconvex and satisfies

$$-\nabla_x c(x, \partial_c\psi(x)) = \partial\psi(x), \tag{C.2}$$

and that the analogue holds for functions on $\Omega'$.

The main condition is (C.2). As $\partial\psi(x)$ is convex, it implies in particular that $\partial_c\psi(x)$ is $c$-convex. (The converse implications also holds; see [42]. Note that our notation differs slightly from [42], where $\partial_c\psi(x)$ denotes what is $-\nabla_x c(x, \partial_c\psi(x))$ in our notation.) It is shown in [42] how (C.2) can be deduced from the (Aw) condition when the the domains are bounded, sufficiently $c$-convex and $c \in C^4$. Local semiconvexity of $c$-convex functions can be ensured by comparably mild conditions on the data, see for instance [42, Proposition 2.2] or [30, Corollary C.5]. Apart from the quadratic cost, another classical example treated in [42] is the reflector-antenna cost $c(x, y) = -\log\|x - y\|$. See also [58] for further background.

*Proof of Proposition 5.6. Step 1: Generalization of Lemma 5.4.* This extension is straightforward: using the same notation as in the proof of Lemma 5.4, we again have $x, x' \in \{I(\cdot, y) = 0\} = \partial_c(-\psi^c)(y)$. The latter set is $c$-convex by Assumption C.1, hence contains the $c$-segment of $x, x'$ wrt. $y$. The interior of the segment is contained in $\operatorname{int}\mathsf{X}_0$ by strict $c$-convexity, and it includes points from the neighborhood were $I$ was assumed to be positive—a contradiction.

*Step 2: Generalization of Proposition 5.5.* Let $(x, y) \in \mathsf{X}_0 \times \mathsf{Y}_0$ be such that $I(x, y) = 0$. In view of Step 1, it again suffices to treat the case $x \in \operatorname{int}\mathsf{X}_0$. Moreover, as the $c$-convex function $\psi$ is semiconvex by our assumption, it still holds that $\partial\psi(x)$ is the closed convex hull of $S(x)$ as defined in (5.6). The proofs for Case 1 and Case 2 carry over by simply replacing $\partial\psi(x)$ with $\partial_c\psi(x)$ and $\nabla\psi(x)$ with $\mathfrak{T}_x(\nabla\psi(x))$. In Case 3, the proof of (5.7) also carries over using semiconvexity. The arguments around (5.8) can be adapted as follows: Let $\phi := -\psi^c$ and $x' \in \partial\phi(y)$. Then the cyclical monotonicity property (C.1) of $\partial\phi$ implies $\langle x'-x, y-y_i\rangle \geq o(\|x'-x\|)$ for all $i$. In view of (5.8), it now follows that $\langle x'-x, y-y_i\rangle = o(\|x'-x\|)$ for all $i$, but noting that the convex set $\partial\phi(y)$ contains the segment $[x', x]$, this already implies that $\langle x'-x, y-y_i\rangle = 0$ for all $i$. The remainder of the proof is identical. □

# References

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

[2] J. Backhoff-Veraguas, M. Beiglböck, and G. Conforti. A non-linear monotonicity principle and application to the Schrödinger problem. *Preprint arXiv:2101.09975v1*, 2021.

[3] A. Baradat and C. Léonard. Minimizing relative entropy of path measures under marginal constraints. *Preprint arXiv:2001.10920v1*, 2020.

[4] S. Bartz and S. Reich. Abstract convex optimal antiderivatives. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 29(3):435–454, 2012.

[5] M. Beiglböck, A. M. G. Cox, and M. Huesmann. Optimal transport and Skorokhod embedding. *Invent. Math.*, 208(2):327–400, 2017.

[6] M. Beiglböck and N. Juillet. On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.*, 44(1):42–106, 2016.

[7] M. Beiglböck, M. Nutz, and F. Stebegg. Fine properties of the optimal Skorokhod embedding problem. *To appear in J. Eur. Math. Soc. (JEMS).*

[8] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.

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

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

[11] J. M. Borwein and A. S. Lewis. Decomposition of multivariate functions. *Canad. J. Math.*, 44(3):463–482, 1992.

[12] 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.

[13] 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.

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

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

[16] G. Clerc, G. Conforti, and I. Gentil. Long-time behaviour of entropic interpolations. *Preprint arXiv:2007.07594v1*, 2020.

[17] 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.

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

[19] D. Cordero-Erausquin and A. Figalli. Regularity of monotone transport maps between unbounded domains. *Discrete Contin. Dyn. Syst.*, 39(12):7101–7112, 2019.

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

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

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

[23] N. Deb and B. Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Preprint arXiv:1909.08733v1*, 2019.

[24] E. del Barrio, J. A. Cuesta-Albertos, M. Hallin, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension $d$: A measure transportation approach. *Ann. Statist.*, 49(2):1139–1165, 2021.

[25] 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.

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

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

[28] 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.

[29] H. Föllmer and N. Gantert. Entropy minimization and Schrödinger processes

in infinite dimensions. *Ann. Probab.*, 25(2):901–926, 1997.

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

[31] 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.

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

[33] P. Ghosal and B. Sen. Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. *Preprint arXiv:1905.05340v1*, 2019.

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

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

[36] C. Léonard. Minimization of energy functionals applied to some inverse problems. *Appl. Math. Optim.*, 44(3):273–297, 2001.

[37] C. Léonard. Minimizers of energy functionals. *Acta Math. Hungar.*, 93(4):281–325, 2001.

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

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

[40] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117, 2018.

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

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

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

[44] R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 1995.

[45] 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.

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

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

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

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

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

[51] 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.

[52] L. Qi. The maximal normal operator space and integration of subdifferentials of nonconvex functions. *Nonlinear Anal.*, 13(9):1003–1011, 1989.

[53] S. T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, 1998. Theory.

[54] S. T. Rachev and L. Rüschendorf. *Mass transportation problems. Vol. II.* Probability and its Applications (New York). Springer-Verlag, New York, 1998. Applications.

[55] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

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

[57] L. Rüschendorf and W. Thomsen. Closedness of sum spaces and the generalized Schrödinger problem. *Teor. Veroyatnost. i Primenen.*, 42(3):576–590, 1997.

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

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