ON THE DIFFERENCE BETWEEN ENTROPIC COST AND THE OPTIMAL TRANSPORT COST

SOUMIK PAL

Abstract. Consider the Monge-Kantorovich problem of transporting densities $\rho_0$ to $\rho_1$ on $\mathbb{R}^d$ with a strictly convex cost function. A popular relaxation of the problem is the one-parameter family called the entropic cost problem. The entropic cost $K_h$, $h > 0$, is significantly faster to compute and $hK_h$ is known to converge to the optimal transport cost as $h$ goes to zero. We are interested in the rate of convergence. We show that the difference between $K_h$ and $1/h$ times the optimal cost of transport has a pointwise limit when transporting a compactly supported density to another that satisfies a few other technical restrictions. This limit is the relative entropy of $\rho_1$ with respect to a Riemannian volume measure on $\mathbb{R}^d$ that measures the local sensitivity of the transport map. For the quadratic Wasserstein transport, this relative entropy is exactly one half of the difference of entropies of $\rho_1$ and $\rho_0$. In that case we complement the results of Adams et al. and others [1, 7, 12] who use gamma convergence. More surprisingly, we demonstrate that this difference of two entropies (plus the cost) is also the limit for the Dirichlet transport introduced by Pal and Wong [24]. The latter can be thought of as a multiplicative analog of the Wasserstein transport and corresponds to a non-local operator. It hints at an underlying gradient flow of entropy, in the sense of Jordan-Kinderlehrer-Otto, even when the cost function is not a metric. The proofs are based on Gaussian approximations to Schrödinger bridges as $h$ approaches zero.

1. Introduction.

Let $\rho_0$ and $\rho_1$ be Borel probability density functions on $\mathbb{R}^d$. We will assume throughout this paper that these are continuous and compactly supported. Following a standard abuse of notation, we will make no notational distinction between absolutely continuous measures and their densities.

Consider the Monge-Kantorovich problem of transporting $\rho_0$ to $\rho_1$ optimally with respect to a cost function $c(x, y) = g(x - y)$ where $g: \mathbb{R}^d \to [0, \infty)$ is strictly convex that satisfies the assumptions in the seminar paper by Gangbo and McCann [14, Theorem 1.2]. In fact, we are going to further assume that $g(0) = 0$, $g(x) > 0$ for $x \neq 0$ and the Hessian of $g$ at the origin is invertible.

More formally, let $\Sigma = \mathbb{R}^d \times \mathbb{R}^d$ and let $M_1(\Sigma)$ denote the set of Borel probability measures on $\Sigma$ equipped with the Lévy metric of weak convergence. We will throughout denote $(X, Y)$ as a pair of $d$-dimensional random variables with law in $M_1(\Sigma)$. For $\nu \in M_1(\Sigma)$, define the two projections, $\pi^0(\nu)$ and $\pi^1(\nu)$ to be the
The Monge-Kantorovich optimal transport cost for the initial density $\rho_0$ and the target density $\rho_1$ with cost $c$ is given by
\[ \mathbb{W}_c(\rho_0, \rho_1) := \inf_{\nu \in \Pi(\rho_0, \rho_1)} \nu(g(x - y)). \]

Here and throughout, if $\nu$ is a measure and $f$ on $\Sigma$ is $\nu$-integrable, the integral will be denoted by $\nu(f)$. In other words, we want the coupling $\nu$ of $(\rho_0, \rho_1)$ that minimizes the expected cost $\nu(g(x - y))$. The coupling that achieves this optimal cost is called the optimal coupling. We say that the optimal coupling solves the Monge problem if it is of the form $(X, T(X))$, where $X \sim \rho_0$ and $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable map that push-forwards $\rho_0$ to $\rho_1$. The particularly well-known case is when $g(x - y) = \frac{1}{2} \|x - y\|^2$, for which we will denote $\mathbb{W}_c(\rho_0, \rho_1)$ by $\mathbb{W}_2^2(\rho_0, \rho_1)$.

For a probability density $\rho$ (or, its corresponding measure) define its entropy by
\[ \text{Ent}(\rho) := \int \rho(x) \log \rho(x) dx. \]

This definition is the negative of the usual Shannon differential entropy. Recall the notion of relative entropy of two densities $\rho$ with respect to $\rho'$,
\[ H(\rho \mid \rho') = \int \rho(y) \log \frac{\rho(y)}{\rho'(y)} dy. \]

The above is nonnegative if both $\rho, \rho'$ are probability densities, but not in general. The following entropic relaxation of the optimal transport problem \[1\] has become quite popular, especially in connection to efficient computational algorithms \[5, 25\].

**Definition 1 (Entropic relaxation).** Fix a parameter $h > 0$ and define the entropic regularization of optimally transporting $\rho_0$ to $\rho_1$ with parameter $h$ as
\[ K'_h(\rho_0, \rho_1) := \inf_{\nu \in \Pi(\rho_0, \rho_1)} \left[ \frac{1}{h} \nu(g(x - y)) + \text{Ent}(\nu) \right]. \]

The quantity $K'_h$ is closely related to minimizing relative entropy. Since $\rho_0$ and $\rho_1$ are compactly supported, assume that $g(z) = \infty$ outside a compact ball. Then consider the joint probability density on $\Sigma$,
\[ \mu_h(x, y) = \rho_0(x) \frac{1}{\Lambda_h(x)} \exp \left( -\frac{1}{h} g(x - y) \right), \]

where $\Lambda_h(x)$ is the normalizing function $\int_{\mathbb{R}^d} \exp \left( -\frac{1}{h} g(x - y) \right) dy$. It is easily verifiable that, if $\nu \in \Pi(\rho_0, \rho_1)$, then
\[ H(\nu \mid \mu_h) = \frac{1}{h} \nu(g(x - y)) + \text{Ent}(\nu) - \text{Ent}(\rho_0) + \int \log \Lambda_h(x) \rho_0(x) dx. \]

**Definition 2 (Entropic cost).** Fix a parameter $h > 0$ and define the entropic cost of optimally transporting $\rho_0$ to $\rho_1$ with parameter $h$ as
\[ K_h(\rho_0, \rho_1) := \inf_{\nu \in \Pi(\rho_0, \rho_1)} H(\nu \mid \mu_h). \]

Since the last two terms in \[4\] do not depend on the coupling,
\[ K_h = K'_h - \text{Ent}(\rho_0) + \int \log \Lambda_h(x) \rho_0(x) dx, \]

while the minimizers of both $K_h$ and $K'_h$ are the same.
As \( h \to 0^+ \) one expects that the minimizer of the entropic problem will “converge” to the minimizer of \((1)\). Such results are known under regularity assumptions (see \cite{18} Theorem 3.3) and are made precise in terms of gamma convergence. In passing, let us mention that the entropic relaxation is frequently written in terms a different parameter \( \lambda = 1/h \): \( hK'_h(\rho_0, \rho_1) = \inf_{\nu \in \Pi(\rho_0, \rho_1)} \left[ \nu(g(X - Y)) + \frac{1}{h} \Ent(\nu) \right] \).

Of course, this does not affect the minimizers.

In this paper we are interested in the limit \( \lim_{h \to 0^+} \left[ K_h(\rho_0, \rho_1) - \frac{1}{h} \mathcal{W}_2^2(\rho_0, \rho_1) \right] \). Since the entropic cost is easier to compute, it is natural to ask for error bounds from the optimal cost.

In the case of the quadratic Wasserstein cost, it is now a famous result (see \cite{1}) that \( x^* := x - (\nabla g)^{-1} \circ \nabla \psi \) is the push-forward of \( \rho_0 \) to \( \rho_1 \). Additionally, the coupling \((X, X^*) \), \( X \sim \rho_0 \), achieves the infimum in \((1)\). This solution is unique, almost surely, on the support of \( \rho_0 \). When \( g(z) = \frac{1}{2} \|z\|^2 \), \( \nabla g \) is the identity map and \( \psi \) is c-concave function is differentiable, wherever it is finite, and it is twice differentiable almost everywhere in the sense of Alexandrov. Theorem 1.2 from \cite{14} shows that there is a c-concave function \( \psi \) on \( \mathbb{R}^d \) such that the map

\[ x^* := x - (\nabla g)^{-1} \circ \nabla \psi \]

push-forwards \( \rho_0 \) to \( \rho_1 \). Additionally, the coupling \((X, X^*) \), \( X \sim \rho_0 \), achieves the infimum in \((1)\). This solution is unique, almost surely, on the support of \( \rho_0 \). When \( g(z) = \frac{1}{2} \|z\|^2 \), \( \nabla g \) is the identity map and \( \psi \) is c-concave function is differentiable, wherever it is finite, and it is twice differentiable almost everywhere in the sense of Alexandrov. Theorem 1.2 from \cite{14} shows that there is a c-concave function \( \psi \) on \( \mathbb{R}^d \) such that the map

\[ x^* := x - (\nabla g)^{-1} \circ \nabla \psi \]

It is proved in \cite{14} that for a cost function as above the optimal coupling is also the solution of the Monge problem, and, in fact, the following description of the optimal solution exists. For \( c(x, y) = g(x - y) \) as above, define the class of c-concave functions as a function \( \psi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\} \) for which there exists a function \( \xi : \mathbb{R}^d \to \mathbb{R} \) such that \( \psi(x) = \inf_{y \in \mathbb{R}^d} \left[ g(x - y) + \xi(y) \right] \).

It has been shown in \cite{14} that, except on a set of dimension \((d - 1)\), the c-concave function is differentiable, wherever it is finite, and it is twice differentiable almost everywhere in the sense of Alexandrov. Theorem 1.2 from \cite{14} shows that there is a c-concave function \( \psi \) on \( \mathbb{R}^d \) such that the map

\[ x^* := x - (\nabla g)^{-1} \circ \nabla \psi \]
with equality precisely when \((x, y) \in \partial \psi\). Finally from \cite[Proposition 3.4]{14}, we get that for \(g\) satisfying our assumptions, \(\partial^c \psi(x) = \{x^\star\}\) whenever \(\psi\) is differentiable at \(x\), which happens Lebesgue almost everywhere on the domain of \(\psi\).

**Definition 3** (\(c\)-divergence). For a \(c\) concave function \(\psi\), the divergence is a function on \(\Sigma\) denoted by \(D[y \mid x^\star] := g(x - y) - \psi(x) - \psi^*(y)\).

Hence, divergence measures the so-called “slackness” in Linear Programming. A transport interpretation of divergence is provided in \cite{23} where it is shown to be a measure of the error in transporting \(x\) to \(y\) instead of the optimal \(x\) to \(x^\star\). See a more geometric analysis in \cite{27}. For the quadratic cost, this notion of divergence coincides with the well-known Bregman divergence and is a fundamental quantity in Information Geometry \cite{2}.

Clearly \(D[y \mid x^\star] \geq 0\) by \cite{8} and, if \(\partial^c \psi(x) = \{x^\star\}\), then \(D[y \mid x^\star]\) is exactly zero if and only if \(y = x^\star\). It follows that \(D[y \mid x^\star]\) must be locally quadratic in a neighborhood of \(x^\star\). A similar conclusion holds for the the convex function \(g\) at the origin since \(g(0) = 0\) and \(g\) is otherwise positive. We are going to assume the following regularity.

**Assumption 1.** Assume that \(\rho_0, \rho_1\) are continuous and compactly supported. Let \(S := \text{spt}(\rho_1)\) denote the support of \(\rho_1\). Assume \(\rho_1\) is smooth (at least \(C^5\)) up to the boundary and \(\inf_{y \in S} \rho_1(y) > 0\). Moreover, \(S\) has a nice boundary in the sense that, if we consider its \(c\)-expansion \(S^\epsilon := \{y : \inf_{z \in S} \|y - z\| \leq \epsilon\}\), then \(\text{Vol}(S^\epsilon) - \text{Vol}(S) = O(\epsilon), \) as \(\epsilon \to 0^+\). This is true, for example, if \(S\) is convex.

**Assumption 2.** Assume that every \(y\) on the support of \(\rho_1\) the following quadratic approximation to the divergence holds uniformly on compact sets.

\[
(9) \quad R[z \mid y] := D[z \mid y] - \frac{1}{2}(z - y)^T A(y)(z - y) = o\left(\|z - y\|^2\right), \quad \text{as } z \to y,
\]

for some smooth (at least \(C^5\)), up to the boundary family, of matrices \(A(y), y \in \text{spt}(\rho_1)\). The family \(A(\cdot)\) is assumed to be uniformly elliptic in the sense that there exists an \(\epsilon > 0\) such that the eigenvalues of \(A(y)\) lie in \((\epsilon, 1/\epsilon)\) for all \(y \in \text{spt}(\rho_1)\). \(1\) We will consider the inverse matrix \(A^{-1}(y)\) as a Riemannian matrix derived from the divergence.

We also assume that a similar Taylor approximation holds for the convex function \(g\) at the origin:

\[
(10) \quad r[y \mid x] := g(x - y) - \frac{1}{2}(y - x)^T \nabla^2 g(0)(y - x) = o\left(\|y - x\|^2\right), \quad \text{as } y \to x,
\]

where the matrix \(\nabla^2 g(0)\) is invertible.

Define the Riemannian volume measure on \(\mathbb{R}^d\) by

\[
(11) \quad \frac{d\mu_g(y)}{dy} = \rho_1(y) \sqrt{\det(A^{-1}(y))}, \quad y \in \mathbb{R}^d,
\]

then \(H(\rho_1 \mid \mu_g) = \frac{1}{2} \int \rho_1(y) \log \det (A(y)) \, dy\).

\(1\) For the quadratic Wasserstein transport the assumption is asking for uniform ellipticity of the Hessian matrix of the Brenier map.
Theorem 1. Suppose Assumptions 1 and 2 are satisfied. Then, the following convergences hold.

\[
\lim_{h \to 0^+} \left( K_h(\rho_0, \rho_1) - \frac{1}{2h} \mathbb{W}_g(\rho_0, \rho_1) \right) = -\frac{1}{2} \log \det \nabla^2 g(0) + H(\rho_1 | \mu_g).
\]

(12)

\[
\lim_{h \to 0^+} \left( K_h'(\rho_0, \rho_1) - \frac{1}{2h} \mathbb{W}_g(\rho_0, \rho_1) + \frac{d}{2} \log(2\pi h) \right) = \text{Ent}(\rho_0) + H(\rho_1 | \mu_g).
\]

In particular, for the quadratic Wasserstein cost \( g(x-y) = \frac{1}{2} \|x-y\|^2 \),

\[
\lim_{h \to 0^+} \left( K_h(\rho_0, \rho_1) - \frac{1}{2h} \mathbb{W}_2^2(\rho_0, \rho_1) \right) = \frac{1}{2} \text{Ent}(\rho_1) - \text{Ent}(\rho_0).
\]

(13)

\[
\lim_{h \to 0^+} \left( K_h'(\rho_0, \rho_1) - \frac{1}{2h} \mathbb{W}_2^2(\rho_0, \rho_1) + \frac{d}{2} \log(2\pi h) \right) = \frac{1}{2} \left( \text{Ent}(\rho_1) + \text{Ent}(\rho_0) \right).
\]

The proof of this theorem follows from the following observation which is best stated in the Wasserstein case. Consider Brownian motion \( \{X_t, 0 \leq t \leq h\} \) in \( \mathbb{R}^d \) “conditioned” to have initial distribution \( \rho_0 \) and terminal distribution \( \rho_1 \) at time \( h \). Such a process is called a Schrödinger bridge between \( \rho_0 \) and \( \rho_1 \). See [19] and [4].

It is intuitive that the law of the vector \( (X_0, X_h) \) is the minimizer of the entropic cost \( K_h \). Hence, if we can exactly describe its joint density, we can compute \( K_h \).

Exactly describing such a bridge is very hard although can be done as an \((f,g)\) transform. See [19, Section 3]. The core of our proof is to show that, for \( h \approx 0 \), the bridge is approximately the following. Sample \( X_0 \) from \( \rho_0 \). Given \( X_0 = x \), sample \( X_h \) from a Gaussian distribution with mean \( x^* \) and covariance \( hA^{-1}(x^*) \) and then join the two points by a straight line traveling at unit speed. Of course, the distribution of \( X_h \), say \( \rho^h_1 \), obtained this way is not exactly \( \rho_1 \). However, as \( h \to 0^+ \), the approximation is tight: not only \( \rho^h_1 \) converges weakly to \( \rho_1 \) but also \( \text{Ent}(\rho^h_1) \) converges to \( \text{Ent}(\rho_1) \). This Gaussian approximation also gives an interpretation of the matrix \( A^{-1}(x^*) \). If \( A^{-1}(x^*) \) is large, one can perturb \( X_h \) more, relatively speaking, around \( x^* \) without paying too much cost. However, if \( A^{-1}(x^*) \) is small, \( X_h \) has a relatively smaller fluctuation around \( x^* \). Hence the Riemannian volume measure \( \mu_g \) can be thought of as a measure of sensitivity of the Monge map.

1.1. The Dirichlet transport. The quadratic Wasserstein case (13) is very special since the limit is the difference of the two entropies. See the next section for its connection to heat equation and gradient flow of entropy. Rather surprisingly, a similar limit appears for a completely different transport cost which. This is the Dirichlet transport problem which we discuss below. More details can be found in [21]. Let \( n \geq 2 \) be an integer and consider the open simplex

\[
\Delta_n = \{(p_1, p_2, \ldots, p_n) : p_i > 0, \forall i, p_1 + \ldots + p_n = 1\}.
\]

Its closure in \( \mathbb{R}^n \) is denoted by \( \overline{\Delta}_n \). The unit simplex has an abelian group structure with a multiplicative group operation

\[
p \odot q := \left( \frac{p \cdot q}{\sum_{j=1}^{n} p_j q_j} \right)_{1 \leq i \leq n}, \quad p, q \in \Delta_n, \quad p^{-1} := \left( \frac{1/p_i}{\sum_{j=1}^{n} 1/p_j} \right)_{1 \leq i \leq n}.
\]

(14)

The identity element is the barycenter \( \tau := \left( \frac{1}{n}, \ldots, \frac{1}{n} \right) \). The \( \odot \) operation plays an analogous role as vector addition in the Wasserstein transport. In particular, the
sign changes that often appear in standard optimal transport definitions will be
replaced here by inversion $p \mapsto p^{-1}$ which takes a little bit of time to get used to.

Let $c : \Delta_n \times \Delta_n \to [0, \infty)$ be the cost function defined (see \cite{24} Lemma 6))

$$
(15) \quad c(p, q) = \log \left( \frac{1}{n} \sum_{i=1}^{n} \frac{q_i}{p_i} \right) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{q_i}{p_i} = H(p \| q \circ p^{-1})
$$

where $H$ is the discrete relative entropy defined on $\Delta_n \times \Delta_n$ by $H(p \| q) := \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$. Clearly, $c(p, q) \geq 0$ for all $p, q$, and $c(p, q) = 0$ only if $p = q$. It is clear that the cost function is not symmetric in $p$ and $q$, although $c(p, q) = c(q^{-1}, p^{-1})$.

The variable $\pi := q \circ p^{-1}$ will play an important role throughout this paper. Following our previous works \cite{22, 23, 21} we call $\pi$ the portfolio vector. Note that $p = q$ (i.e., $c(p, q) = 0$) if and only if the portfolio vector $\pi$ is equal to the barycenter $\overline{\pi}$. Since $H(p \| q)$ is a convex function of $q$ for fixed $p$, our cost function $H(\pi \| p^{-1} \circ q)$ is similar to the cost $g(y - x)$, for $x, y \in \mathbb{R}^n$, but not exactly so.

As before, given densities $\rho_0, \rho_1$, compactly supported on $\Delta_n$, consider the cost $C(\rho_0, \rho_1)$ of transporting $\rho_0$ to $\rho_1$ with respect to cost $c$. It is clear that $C(\rho_0, \rho_1)$ is not a metric since it is asymmetric in $\rho_0$ and $\rho_1$. The regularity theory for this transport has been recently studied in \cite{17}. It is possible by a change of coordinates (see Section 3) to reduce this cost

The name Dirichlet transport comes from its deep connection to the Dirichlet distribution as shown in \cite{24}. For any $\lambda > 0$, consider the symmetric Dirichlet distribution $\text{Diri}(\lambda)$ with parameters $(\lambda/n, \ldots, \lambda/n)$ on $\Delta_n$ with density

$$
(16) \quad \frac{\Gamma(\lambda)}{(\Gamma(\lambda/n))^n} \prod_{i=1}^{n} p_i^{\lambda/n - 1}, \quad p_n := 1 - \sum_{i=1}^{n-1} p_i.
$$

The density is with respect to the $(n - 1)$ dimensional Lebesgue measure on the open set $\{(p_1, p_2, \ldots, p_{n-1}), p_i > 0, \sum_{i=1}^{n-1} p_i < 1\}$.

To calculate entropy consider the Haar measure on $\Delta_n$ with respect to the group operation $\circ$. This measure is the sigma-finite Dirichlet measure $\text{Diri}(0)$ with a density $v_0(p) = \prod_{i=1}^{n} p_i^{-1}$. For any probability density $\rho$ on $\Delta_n$, define

$$
\text{Ent}_0(\rho) := \rho \left( \log \frac{\rho}{v_0} \right).
$$

Let $\mu_h$ denote the joint density on $\Delta_n \times \Delta_n$ where we sample $X \sim \rho_0$ and, given $X = y$, $Y$ has the distribution of $p \circ G_h$, for an independent Dirichlet$(1/h)$ random variable $G_h$. Define the entropic cost of coupling $(\rho_0, \rho_1)$ with parameter $h$ as

$$
(17) \quad K_h(\rho_0, \rho_1) = \inf_{\nu \in \Pi(\rho_0, \rho_1)} H(\nu \| \mu_h).
$$

Given $\rho_0, \rho_1$, the Monge solution exists and is unique and there is a corresponding notion of divergence. The details can be found in Section 3.

**Theorem 2.** Suppose that Assumption 3 and 4 (similar to Assumption 1 and 2) are satisfied. Then,

$$
\lim_{h \to 0^+} \left( K_h(\rho_0, \rho_1) - \left( \frac{1}{h} - \frac{n}{2} \right) C(\rho_0, \rho_1) \right) = \frac{1}{2} (\text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0)).
$$

The cost $c(p, q)$ is not of the form $g(p - q)$ for a convex function $g$, although $H(\pi \| q \circ p^{-1})$ is a convex function of $q \circ p^{-1}$. However, by a change of coordinates, it can be related. See Section 3 or \cite{23} for details. Unfortunately, this observation
1.2. Discussion on the gradient flow of entropy. Considers the heat equation with initial condition $\rho_0$:

$$\partial_t \rho(t, x) = \Delta_x \rho(t, x), \quad 0 \leq t \leq 1, \quad x \in \mathbb{R}^d, \quad \rho(0, \cdot) = \rho_0.$$  

Here $\Delta_x$ is the Laplacian in $x$. Let $\rho(t)$ denote the density $\rho(t, \cdot)$.

In [15] Jordan, Kinderlehrer, and Otto (JKO) interprets the heat equation as a gradient flow entropy in the $W_2$ metric space. The idea comes from the following discretization scheme that they propose. Let $h > 0$ be the step size. Starting with $\rho(0) = \rho_0$, (18) determine $\rho^{(k)}$ that minimizes \[
\frac{1}{2h} \mathbb{W}_2^2(\rho, \rho^{(k-1)}) + \left( \text{Ent}(\rho) - \text{Ent}(\rho^{(k-1)}) \right) .
\]

Define the piecewise constant interpolation $\rho_h(t) = \rho^{(k)}$, for $t \in [kh, (k+1)h)$ and $k = 0, 1, 2, \ldots$. Then, [15] shows that, as $h \downarrow 0$, $\rho_h(t)$ converges to $\rho(t)$ weakly in $L^1$. Their theorem is somewhat more general since it includes Fokker-Planck equations, but that extension from the heat equation is not very hard. In [1] the authors developed an alternative approach via the entropic cost function. Recall the discussion in the paragraph above Section 1.1. Let $\mu_h(x, y)$ be the joint density of $(X_0, X_h)$. Then it follows that $K_h(\rho_0, \rho_1) := \inf_{\nu \in \Pi(\rho_0, \rho_1)} \mathbb{H}(\nu|\mu_h)$. Then the authors argued (dimension one, compactly supported densities) that (13) holds in the sense of gamma convergence. Since the solution to the heat equation could be seen as successively minimizing the entropic cost over a discrete set of points $h, 2h, 3h, \ldots$, it gives credence to the idea that, for small values of $h$, the Euler discretization scheme (13) converges to the solution of the heat equation. Making a rigorous theory of such gradient flow of entropy is highly non-trivial and can be found in the textbook [3]. Very recently, a pathwise version of gradient flow result has appeared in [16].

Theorem 2 regarding the Dirichlet transport hints at a possible extension of the theory of gradient flow of entropy on metric spaces of probabilities. $\mu_h$ corresponds to a stochastic process that evolves by successive multiplications $\odot$ by independent symmetric Dirichlet distributed random variables. Such an evolution is non Fokker-Planck and, in fact, non local, since it proceeds by jumps. However, the entropic cost admits a similar asymptotic formula as that of quadratic Wasserstein. Thus, one naturally expects a theory of “gradient flow of entropy” even when the cost of transport $C$ is not a metric between probabilities. For non-local operators corresponding to stochastic processes such as jump processes and Markov chains, a rather important theory of gradient flow of entropy has been developing steadily. See articles [20, 10, 8, 11, 6, 9]. However, the cost function considered in these papers are metrics constructed by generalizations of the Benamou-Brenier formula.

As a final comment, both quadratic Wasserstein and the Dirichlet transports are related to probability distributions (Gaussian and Dirichlet, respectively) that form an exponential family [2]. In [21] it is shown that the transport for all exponential families share similar features including the structure of the Kantorovich potential and the Monge solution. It seems worthwhile to extend the current analysis as well.
2. General convex cost and the proof of Theorem 1

Fix $h > 0$. Recall the joint density $\mu_h$ from [13], well-defined under the assumption that $g = \infty$ outside a compact set in $\Sigma$. This joint density is the product of $\rho_0(x)$ and a transition densities on $\Sigma$:

\[ p_h(x, y) := \frac{1}{\Lambda_h(x)} \exp \left( -\frac{1}{h} g(x - y) \right), \quad \Lambda_h(x) := \int_{\mathbb{R}^d} \exp \left( -\frac{1}{h} g(x - y) \right) dy. \]

The following exponential tilting $p_h$ is of central importance.

\[ \tilde{p}_h(x, y) := \frac{1}{Z_h(x)} \exp \left( \frac{1}{h} \left( \psi(x) + \psi^*(y) \right) \right) p_h(x, y) \]

\[ = \frac{1}{Z_h(x) \Lambda_h(x)} \exp \left( -\frac{1}{h} D[y \mid x^*] \right), \quad \text{from Definition [3]} \]

\[ = \frac{1}{\Lambda_h(x)} \exp \left( -\frac{1}{h} D[y \mid x^*] \right), \quad \text{say.} \]

Here $\tilde{\Lambda}_h(x) := Z_h(x) \Lambda_h(x)$ and $Z_h(x)$ is the appropriate normalizing function

\[ Z_h(x) := \frac{1}{\Lambda_h(x)} \int \exp \left( -\frac{1}{h} D[y \mid x^*] \right) dy. \]

The following is the key lemma for our proofs.

**Lemma 3.** Under Assumptions [11] and [2] the following hold.

(i) The normalizing constant $\Lambda_h(x)$ converges to the following limit uniformly on $\text{spt}(\rho_0)$.

\[ \lim_{h \to 0^+} \frac{\Lambda_h(x)}{(2\pi h)^{d/2}} = \frac{1}{\sqrt{\det (\nabla^2 g(0))}}. \]

(ii) The normalizing constant $\tilde{\Lambda}_h(x)$ has the following limit uniformly on $\text{spt}(\rho_0)$.

\[ \lim_{h \to 0^+} \frac{\tilde{\Lambda}_h(x)}{(2\pi h)^{d/2}} = \frac{1}{\sqrt{|J|(x^*)}}, \quad |J|(x^*) := \det (A(x^*)), \]

where the matrix valued function $A(\cdot)$ is defined in Assumption [2].

(iii) Recall the Riemannian volume measure from [11]. Then,

\[ -\lim_{h \to 0^+} \int \log Z_h(x) \rho_0(x) dx = -\frac{1}{2} \log \det \nabla^2 g(0) + H(\rho_1 \mid \mu_y), \]

where $H(\rho_1 \mid \mu_y) = \frac{1}{2} \int \rho_1(y) \log |J|(y) dy$. In particular, when $g(x - y) = \frac{1}{2} \|x - y\|^2$,

\[ \lim_{h \to 0^+} -\int \log Z_h(x) \rho_0(x) dx = \frac{1}{2} (\text{Ent}(\rho_1) - \text{Ent}(\rho_0)). \]

(iv) Let $\tilde{\mu}_h(x, y)$ denote the joint density $\rho_0(x) \tilde{p}_h(x, y)$ on $\Sigma$. Let $\rho_1^n$ denote the law of $Y$ under $\tilde{\mu}_h$, then $\lim_{h \to 0^+} \rho_1^n = \rho_1$ in total variation. Moreover,

\[ \lim_{h \to 0^+} \text{Ent}(\rho_1^n) = \text{Ent}(\rho_1). \]

The long proof of this lemma is broken down in separate parts.
Proof of Lemma 3 (i). For this proof let $V$ temporarily denote the Hessian matrix $\nabla^2 g(0)$ which is assumed to be invertible. Let $q_h(x, y)$ denote the Gaussian transition density with mean $x$ and covariance matrix $hV^{-1}$. That is,

$$q_h(x, y) = \frac{\det(V^{1/2})}{(2\pi h)^{d/2}} \exp\left(-\frac{1}{2h}(y-x)^T V (y-x)\right), \quad y \in \mathbb{R}^d.$$ 

We will show that $p_h$ and $q_h$ are asymptotically, as $h \to 0+$, the same in a certain sense.

To see this, recall the definition of $r[y \mid x]$,

$$r[y \mid x] = g(x-y) - \frac{1}{2}(y-x)^T V (y-x).$$

Hence, by a change of measure,

$$\Lambda_h(x) = \int_{\mathbb{R}^d} \exp\left(-\frac{1}{h} r[y \mid x]\right) q_h(x, y) dy.$$ 

We show that the right side converges to one as $h \to 0+$.

By the strict convexity of $g$ and Taylor approximation, for any $\epsilon \in (0, 1)$, there exists a $\delta > 0$ such that

$$|r[y \mid x]| \leq \frac{\epsilon}{2} (y-x)^T V (y-x),$$

for all $y \in B_\delta(x)$, the Euclidean ball of radius $\delta$ around $x$. Consider the decomposition:

$$\int e^{-r[y|x]/h} q_h(x, y) dy = \int_{B_\delta(x)} e^{-r[y|x]/h} q_h(x, y) dy + \int_{B^c_\delta(x)} e^{-r[y|x]/h} q_h(x, y) dy.$$ 

We will analyze the two terms separately. For the first term,

$$\int_{B_\delta(x)} e^{-r[y|x]/h} q_h(x, y) dy \leq \int_{B_\delta(x)} \exp\left(\frac{\epsilon}{2h} (y-x)^T V (y-x)\right) q_h(x, y) dy \leq \int_{\mathbb{R}^d} \exp\left(\frac{\epsilon}{2h} (y-x)^T V (y-x)\right) q_h(x, y) dy \leq \frac{\det(V^{1/2})}{\sqrt{(2\pi h)^{d}}} \int_{\mathbb{R}^d} \exp\left(-\frac{(1-\epsilon)}{2h} (y-x)^T V (y-x)\right) dy \leq (1-\epsilon)^{-d/2}. $$

For the second term on the right side of (23), reverse the change of measure.

$$\int_{B^c_\delta(x)} e^{-r[y|x]/h} q_h(x, y) dy = \frac{\det(V^{1/2})}{\sqrt{(2\pi h)^d}} \int_{B^c_\delta(x)} e^{-g(x-y)/h} dy$$

Since $g$ is strictly convex, it has a positive infimum on $B_\delta^c$. Moreover, $g = \infty$ outside a bounded domain. Hence, there exists positive constants $M, m$ such that the above expression is bounded above by

$$M \frac{\det(V^{1/2})}{\sqrt{(2\pi h)^d}} e^{-m/h} \to 0, \quad \text{as } h \to 0+.$$

Thus, combining the two bounds,

$$\int e^{-r[y|x]/h} q_h(x, y) dy \leq (1-\epsilon)^{-d/2} + \epsilon, \quad \text{for all small enough } h > 0,$$

which shows an upper bound on the right side of (21).
For the lower bound, discard the integrand outside \( B_h(x) \). For all small enough \( h > 0 \), \( q_h(x, \cdot) \) puts its almost entire mass in \( B_h(x) \). Hence, for all small enough \( h \),

\[
\int e^{-r|y|x|/h} q_h(x, y) dy \geq \int_{B_h(x)} e^{-r|y|x|/h} q_h(x, y) dy \\
\geq \int \exp \left( -\frac{\epsilon}{2h} (y - x)^T V (y - x) \right) q_h(x, y) dy - \epsilon \\
\geq \int \frac{\det(V^{1/2})}{\sqrt{(2\pi h)^d}} \exp \left[ -\frac{(1 + \epsilon)}{2h} (y - x)^T V (y - x) \right] dy - \epsilon \\
\geq (1 + \epsilon)^{-d/2} - \epsilon, \text{ for all small enough } h.
\]

Thus we get that

\[
(1+\epsilon)^{-d/2} - \epsilon \leq \lim_{h \to 0^+} \Lambda_h(x) \frac{\det(V^{1/2})}{(2\pi h)^{d/2}} \leq \limsup_{h \to 0^+} \Lambda_h(x) \frac{\det(V^{1/2})}{(2\pi h)^{d/2}} \leq (1-\epsilon)^{-d/2} + \epsilon.
\]

As we now take \( \epsilon \downarrow 0 \), we recover our claim. \( \square \)

**Proof of Lemma** (ii). The proof of part (ii) is very similar to that of part (i) except at every step we will require a uniform estimate over \( x \in \text{spt}(\rho) \). Suppress the argument \( x^* \) from the matrix \( A(x^*) \) for simplicity. As in (i), define the Gaussian kernel

\[
\tilde{q}_h(x, y) = \frac{\det(A^{1/2})}{(2\pi h)^d} \exp \left( -\frac{1}{2h} (y - x^*)^T A (y - x^*) \right)
\]

which has mean \( x^* \) and covariance \( hA^{-1} \). Then

\[
\tilde{\Lambda}_h(x) = \frac{\det(A^{1/2})}{(2\pi h)^d/2} = \int_{\mathbb{R}^d} \exp \left( -R[y \mid x^*] \right) \tilde{q}_h(x, y) dy,
\]

where the remainder term \( R[y \mid x^*] \) is defined in (ii). We show that the right side converges to one as \( h \to 0^+ \).

The analysis is very similar to the proof of (i) and it suffices to highlight just the main differences. First, by our assumption on locally uniform estimates, one can find a \( \delta > 0 \) such that for all \( x^* \in \text{spt}(\rho_1) \),

\[
|R[y \mid x^*]| \leq \frac{\epsilon}{2} (y - x)^T A (y - x), \quad y \in B_h(x^*).
\]

So the bound

\[
\int_{B_h(x^*)} e^{-R[y|x^*]/h} \tilde{q}_h(x, y) dy \leq (1 - \epsilon)^{-d/2}
\]

holds as before in (i). On \( B_h^*(x^*) \), the infimum of \( D[y \mid x^*] \) is positive and the divergence is infinity whenever \( g \) is infinity. Therefore the complete upper bound holds as in (i). The proof of lower bound is identical to (i). Hence, given \( \epsilon > 0 \), there exists \( h(\epsilon) > 0 \) such that for all \( 0 < h < h(\epsilon) \), the following bound holds uniformly over \( x \in \text{spt}(\rho_0) \):

\[
(1 + \epsilon)^{-d/2} - \epsilon \leq \liminf_{h \to 0^+} \tilde{\Lambda}_h(x) \frac{\det(A^{1/2}(x^*))}{(2\pi h)^{d/2}} \\
\leq \limsup_{h \to 0^+} \tilde{\Lambda}_h(x) \frac{\det(A^{1/2}(x^*))}{(2\pi h)^{d/2}} \leq (1 - \epsilon)^{-d/2} + \epsilon.
\]

Taking \( \epsilon \downarrow 0 \), this proves (ii). \( \square \)
Proof of Lemma 3 (iii). Since \( Z_h(x) = \tilde{A}(x)/A(x) \), from the uniform convergence statements in parts (i) and (ii) it follows that

\[
\lim_{h \to 0^+} \int \log Z_h(x) \rho_0(x) dx = \frac{1}{2} \int \left( \log \det \nabla^2 g(0) - \log |J| (x^*) \right) \rho_0(x) dx
\]

\[
= \frac{1}{2} \log \det \nabla^2 g(0) - \frac{1}{2} \int \log |J| (y) \rho_1(y) dy,
\]

since \( x \to x^* \) push-forwards \( \rho_0 \) to \( \rho_1 \). However, by our definition of \( \mu_g \), \( H (\rho_1 | \mu_g) = \frac{1}{2} \int (\log |J| (y)) \rho_1(y) dy \). This proves our claim that

\[
- \lim_{h \to 0^+} \int \log Z_h(x) \rho_0(x) dx = - \frac{1}{2} \log \det \nabla^2 g(0) + H (\rho_1 | \mu_g).
\]

Now consider the special case \( g(x - y) = \frac{1}{2} \|x - y\|^2 \), the quadratic Wasserstein cost. By Brenier’s theorem, there exists a convex function \( \varphi \) such that \( \nabla \varphi \) push-forwards \( \rho_0 \) to \( \rho_1 \). In terms of this function \( \varphi \) and its convex conjugate \( \varphi^* \), we can write

\[
\psi(x) = \frac{1}{2} \|x\|^2 - \varphi(x), \quad \psi^*(y) = \frac{1}{2} \|y\|^2 - \varphi^*(y),
\]

Thus, \( D[y | x^*] = \varphi(x) + \varphi^*(y) - \langle x, y \rangle \),

is the Bregman divergence of the convex function \( \varphi \).

Hence

\[
\det (\nabla^2 g(0)) = 1, \quad A(y) = \nabla^2 \varphi^*(y), \quad |J| (y) = \det (\nabla^2 \varphi^*(y)).
\]

Use the fact that the push-forward of \( \rho_1 \) by \( \nabla \varphi^* \) gives us \( \rho_0 \). Thus, by the change-of-variable formula

\[
|J| (x^*) = \frac{\rho_1(x^*)}{\rho_0(x)}, \quad \text{or} \quad \log |J| (x^*) = \log \rho_1(x^*) - \log \rho_0(x).
\]

Thus

\[
\operatorname{Ent}(\rho_0) = \int \rho_0(x) \log \rho_0(x) dx = \int \rho_1(y) \log \frac{\rho_1(y)}{|J|(y)} dy
\]

\[
= \operatorname{Ent}(\rho_1) - \int \rho_1(y) \log |J|(y) dy.
\]

Therefore, \( \int \rho_1(y) \log |J|(y) dy = \operatorname{Ent}(\rho_1) - \operatorname{Ent}(\rho_0) \), and, hence, from (25),

\[
\lim_{h \to 0^+} - \int \log Z_h(x) \rho_0(x) dx = - \frac{1}{2} \log(1) + \frac{1}{2} \int \log |J|(y) \rho_1(y) dy
\]

\[
= \frac{1}{2} \left( \operatorname{Ent}(\rho_1) - \operatorname{Ent}(\rho_0) \right).
\]

This completes the proof of the statement. \( \Box \)

Proof of Lemma 3 (iv). We will estimate the entropy and the total variation distance by comparing the densities \( \rho_1 \) and \( \rho_1^n \).

It follows from Assumption 4 that there exists \( \varepsilon > 0 \) such that \( \varepsilon < \rho_1(y) < \varepsilon^{-1} \), for all \( y \in S = \text{spt}(\rho_1) \). By definition \( S \) is a closed set.
Now,

\[ \| \rho_1 - \rho_1^h \|_{TV} = \frac{1}{2} \int | \rho_1(y) - \rho_1^h(y) | \, dy, \quad \text{and} \]

\[ \text{Ent}(\rho_1^h) = \int_{\mathbb{R}^d} \rho_1^h(y) \log \rho_1^h(y) \, dy = \int_S \rho_1^h(y) \log \rho_1^h(y) \, dy + \int_{S^c} \rho_1^h(y) \log \rho_1^h(y) \, dy. \]

Consider the integral of \( \rho_1^h \log \rho_1^h \) over \( S \). By a change of measure,

\[ \int_S \rho_1^h(y) \log \rho_1^h(y) \, dy = \int_S \rho_1^h(y) \left( \log \frac{\rho_1^h(y)}{\rho_1(y)} \right) \rho_1(y) \, dy + \int_S \rho_1^h(y) \log \rho_1(y) \, dy. \]

We claim that

\[ \lim_{h \to 0+} \int_S \rho_1^h(y) \log \rho_1^h(y) \, dy = \text{Ent}(\rho_1), \quad \text{and} \]

\[ \lim_{h \to 0+} \int_{S^c} \rho_1^h(y) \log \rho_1^h(y) \, dy = 0. \]

These two together will prove the statement of the Lemma.

By the Dominated Convergence Theorem, in order to prove the first claim in \(^{27}\) it suffices to show the following uniform bound. There exist bounded functions \( \zeta_1(y), \zeta_2(y), \zeta_3(y), \) \( y \in S \), such that, as \( h \to 0+ \),

\[ \rho_1^h(y) - \rho_1(y) = h \zeta_1(y) + h^{3/2} \zeta_2(y) + h^2 \zeta_3(y) + O(h^{5/2}), \quad \text{and that} \quad \frac{1}{C} \leq \frac{\rho_1^h(y)}{\rho_1(y)} \leq C, \]

for some positive constant \( C \). This will show total variation convergence and the convergence of entropy. Although the above asymptotics is finer than what we require for this argument, we will need these functions \( \zeta_i \)s later.

Let’s prove \(^{28}\). Consider the explicit expressions for \( \rho_1^h(y) \) and change the variable to \( z = x^* \):

\[ \rho_1^h(y) = \int \frac{\rho_0(x)}{\Lambda_h(x)} \exp \left( -\frac{1}{h} D[y \mid x^*] \right) \, dx = \int \frac{\rho_1(z)}{\Lambda_h(z^*)} \exp \left( -\frac{1}{h} D[y \mid z] \right) \, dz \]

\[ = \int \frac{(2\pi h)^{d/2}}{\sqrt{J(z)\Lambda_h(z^*)}} \rho_1(z) \exp \left( -\frac{1}{h} R[y \mid z] \right) \tilde{q}_h(z, y) \, dz, \quad z^* = x. \]

Here, \( \tilde{q}_h(z, y) \) be the Gaussian density with mean \( z \) and covariance \( hA^{-1}(z) \).

By our assumption on the smoothness of \( \rho_1 \), there exists \( \delta > 0 \) such that for \( z \in B_\delta(y) \), the fourth order Taylor expansion holds:

\[ \rho_1(z) = \rho_1(y) + \nabla \rho_1(y) \cdot (z - y) + \frac{1}{2}(z - y)^T \nabla^2 \rho_1(y)(z - y) \]

\[ + \text{third order terms} + \text{fourth order terms} + O(\|z - y\|^{5/2}). \]
Consider the integral \( \int_{B_3(y)} \frac{\rho_1(z)}{\Lambda_h(z^*)} \exp \left( -\frac{1}{h} D[y \mid z] \right) dz = \rho_1(y) \int_{B_3(y)} \frac{1}{\Lambda_h(z^*)} \exp \left( -\frac{1}{h} D[y \mid z] \right) dz \)

\[ + \nabla \rho_1(y) \cdot \int_{B_3(y)} (z - y) \frac{(2\pi h)^{d/2}}{\sqrt{J(z)\Lambda_h(z^*)}} \rho_1(z) \exp \left( -\frac{1}{h} R[y \mid z] \right) \bar{q}_h(z, y) dz \]

\[ + \frac{1}{2} \int_{B_3(y)} (z - y)^T \nabla^2 \rho_1(y) (z - y) \frac{(2\pi h)^{d/2}}{\sqrt{J(z)\Lambda_h(z^*)}} \rho_1(z) \exp \left( -\frac{1}{h} R[y \mid z] \right) \bar{q}_h(z, y) dz \]

\[ + \text{third order terms + fourth order terms} \]

\[ + O \left( \int_{B_3(y)} \|z - y\|^5 \frac{(2\pi h)^{d/2}}{\sqrt{J(z)\Lambda_h(z^*)}} \rho_1(z) \exp \left( -\frac{1}{h} R[y \mid z] \right) \bar{q}_h(z, y) dz \right). \]

The right side above consists of six terms, say \( E_1, \ldots, E_6 \) in order of appearance. For \( h \) sufficiently small, the intuition is that integral outside \( B_3(y) \) is small of the order \( \exp(-1/h) \). Inside \( B_3(y) \), the measure is approximately Gaussian with mean \( y \) and covariance \( hA^{-1}(y) \). Hence the first term \( E_1 \) gives \( \rho_1(y) \) (plus exponentially in \( 1/h \) small errors); the linear terms \( E_2 \) integrates to zero plus \( O(h) \), since the Gaussian measure is centered at \( y \); the quadratic terms \( E_3 \) is order of \( h \) (plus lower order terms), and so on until the final error term \( E_6 \) is \( h^{3/2} \).

The verification of the above is tedious but completely standard analysis. We point out the major steps. For example, we will ignore the term

\[ \frac{(2\pi h)^{d/2}}{\sqrt{J(z)\Lambda_h(z^*)}} \]

which has already been shown to be uniformly close to one for small \( h \) and can be handled by a similar argument. We also need a separate argument on \( S^c \) where \( \rho_1(y) = 0 \). This will be done more carefully later.

Now, to verify that \( E_1 \) is \( \rho_1(y) \) by exponentially small error is easy since it has already been argued before in part (ii) that

\[ \int_{B_3(y)} \frac{1}{\Lambda_h(z^*)} \exp \left( -\frac{1}{h} D[y \mid z] \right) dz = O \left( h^{-d/2} e^{-c/h} \right), \]

which is exponentially small.

Given any \( \epsilon > 0 \), by Assumption \ref{assumption} choose \( h, \delta \) small enough such that for all \( z \in B_3(y) \) the following estimates hold

\[ |R[y \mid z]| \leq \epsilon(z - y)^T A(y)(z - y), \quad \|A(z)A^{-1}(y)\| - 1 \leq \epsilon, \]

\[ \det A^{1/2}(z) = \det A^{1/2}(y) + O(\|y - z\|), \quad \frac{\rho_1(z)}{\rho_1(y)} - 1 \leq \epsilon. \]

Consider the integral for linear term \( E_2 \). By expanding the Gaussian integral we get

\[ \int_{B_3(y)} (z - y) \exp \left( -\frac{1}{h} R[y \mid z] \right) \frac{\det A^{1/2}(z)}{(2\pi h)^{d/2}} \exp \left( -\frac{1}{2h} (z - y)^T A(z)(z - y) \right) dz. \]

By expanding \( A(z) = A(y) + \text{lower order terms} \) and by the symmetry of the Gaussian distribution we obtain that term \( E_2 \) is \( h\zeta_2(y) \), for the some bounded \( \zeta_2 \). By the
same logic as before,
\[ \int_{B^c_{\frac{h}{2}}(y)} (z - y) \exp \left( \frac{1}{h} R[y | z] \right) \frac{\det A^{1/2}(z)}{(\sqrt{2\pi h})^d} \exp \left( \frac{-1}{2h} (z - y)^T A(z)(z - y) \right) dz \]
is exponentially small of order \( e^{-O(1/h)} \).

The same argument holds for terms \( E_3 \) through \( E_6 \) since the Gaussian integrals of the quadratic terms is \( O(h) \) and the cubic terms are \( O(h^{3/2}) \), and so on. Collecting all error terms gives us the \( \zeta_i \)s. The rest is clearly \( O(h^{3/2}) \). This gives us [28].

For \( y \in S^c \), we have the following claim: as \( h \to 0+ \), uniformly in \( y \),
\[ (31) \]
\[ \rho^h_1(y) = O \left( (\log(1/h))^{d/2} \right), \quad \int_{S^c} \rho^h_1(y) \log \rho^h_1(y) dy = O \left( h^{1/2}(\log(1/h))^{d/2+2} \right) \to 0. \]

By ignoring smaller order terms, it is sufficient to prove the above for a simpler expression:
\[ (32) \]
\[ \gamma^h(y) = \int \rho_1(z) \exp \left( -\frac{1}{h} D[y | z] \right) \frac{1}{(\sqrt{2\pi h})^d} dz. \]

Define \( \tau^2(y, S) := \inf_{z \in S} D[y | z] > 0 \). We are using the notation \( \tau^2 \), not because \( \tau \) is a metric (it is not), but because we use its level sets below as if it were a metric. This intuition will hopefully make the proof more transparent.

Consider a positive parameter \( a > 4 \) such that \( d/2 < a < d/2 + 1 \). Define a partition of \( S^c \) given by
\[ A_k := \left\{ y : k a h \log \frac{1}{h} < \tau^2(y, S) \leq (k + 1) a h \log \frac{1}{h} \right\}, \quad k = 0, 1, 2, \ldots. \]

We will separately estimate \( \gamma^h_1 \) above and below on each \( A_k \).

For \( y \in A_0 \), let \( B_{\sqrt{a h \log(1/h)}(y)} \) denote the set
\[ B_{\sqrt{a h \log(1/h)}(y)} := \{ z : D[y | z] \leq a h \log(1/h) \}. \]

Since \( D[y | z] \) is uniformly locally quadratic on \( S \), we get
\[ \text{Vol} \left( B_{\sqrt{a h \log(1/h)}(y)} \right) = O \left( (h \log(1/h))^{d/2} \right), \quad \text{as } h \to 0+. \]

Then, for \( a > (d + 2)/m \), if \( y \in A_0 \), we get
\[ \begin{align*}
\int_S \frac{1}{(\sqrt{2\pi h})^d} \exp \left( -\frac{1}{h} D[y | z] \right) dz & \leq \int_{S \cap B_{\sqrt{a h \log(1/h)}(y)}} \frac{dz}{(\sqrt{2\pi h})^d} \\
+ \int_{S \cap B_{\sqrt{a h \log(1/h)}(y)}} \frac{1}{(\sqrt{2\pi h})^d} \exp \left( -\frac{1}{h} a h \log(1/h) \right) dz \\
& \leq \frac{1}{(\sqrt{2\pi h})^d} \text{Vol} \left( B_{\sqrt{a h \log(1/h)}(y)} \right) + \frac{1}{(\sqrt{2\pi h})^d} \exp \left( -\frac{1}{h} a h \log(1/h) \right) \text{Vol}(S) \\
& \leq O \left( (\log(1/h))^{d/2} \right) + O \left( h^{a - d/2} \right).
\end{align*} \]
Thus $\gamma_h^1(y) \leq C_0(\log(1/h))^{d/2}$ for some constant $C_0 > 0$, as $h \to 0+$. Similarly we get the lower bound.

$$\int_S \frac{1}{\sqrt{2\pi h}^d} \exp\left( -\frac{1}{h} D(y \mid z) \right) dz \geq \int_S \frac{1}{\sqrt{2\pi h}^d} \exp\left( -\frac{1}{h} a h \log(1/h) \right) dz \geq \frac{h^a}{\sqrt{2\pi h}^d} \text{Vol}(S) = \frac{1}{(2\pi)^d} h^{a-d/2} \text{Vol}(S).$$

Since $a - d/2 < 1$, $\gamma_h^1(y) \geq C'_0 h$ for some positive constant $C'_0$. Hence, over $A_0$,

$$\int_{A_0} \left| \log \gamma_h^1(y) \right| \leq \max \left( \left| \log C_0 \right| + \frac{d}{2} \log \log(1/h), \left| \log C'_0 \right| + \log(1/h) \right) \leq D_0 + D'_0 \log(1/h),$$

for some positive constants $D_0, D'_0$. Hence

$$\int_{A_0} \left| \log \gamma_h^1(y) \right| \gamma_h^1(y) dy \leq \int_{A_0} (D_0 + D'_0 \log(1/h)) \gamma_h^1(y) dy \leq (D_0 + D'_0 \log(1/h)) C_0(\log(1/h))^{d/2} \text{Vol}(A_0) = O(\sqrt{h}(\log(1/h))^{d/2+2}) \to 0,$$

as $h \to 0+$. Here we are using the assumption that the volume of $\delta$ expansion of $S$ is $\text{Vol}(S) + O(\delta)$. Since $A_0$ is the $\delta$ expansion of $S$ in $\tau$, which is comparable to a $\delta = O(\sqrt{h}(\log(1/h)))$ Euclidean expansion of $A_0$, we get $\text{Vol}(A_0) = O(\sqrt{h}(\log(1/h))).$

The others are easier to estimate. For $y \in A_k, k \geq 1$, we get

$$\int_S \frac{1}{\sqrt{2\pi h}^d} \exp\left( -\frac{1}{h} D(y \mid z) \right) dz \leq \frac{1}{(\sqrt{2\pi h})^d} e^{-\frac{4}{h} h \log(1/h)} \text{Vol}(S) \leq \frac{\text{Vol}(S)}{(\sqrt{2\pi})^d} h^k. $$

Along the same lines, the lower bound gives

$$\int_S \frac{1}{\sqrt{2\pi h}^d} \exp\left( -\frac{1}{h} D(y \mid z) \right) dz \geq \frac{\text{Vol}(S)}{(\sqrt{2\pi})^d} h^{a(k+1)-d/2} \geq \frac{\text{Vol}(S)}{(\sqrt{2\pi})^d} h^{ak}.$$ 

Hence, for $y \in A_k, k \geq 1$, 

$$\gamma_h^1(y) \leq C_1 h^{k+1}, \quad \left| \log \gamma_h^1(y) \right| \leq D_1 (k + 1) \log(1/h),$$
The final integral is finite since in a small neighborhood around $S$ compact set.

$$\int_{A_k} \gamma^h(y) dy \leq M_0 D_1 \text{Vol}(S) \sum_{k=1}^\infty (k+1) \log(1/h) \int_{A_k} \gamma^h(y) dy$$

$$\leq D_1 \sum_{k=1}^\infty (k+1) \log(1/h) \int_{A_k} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{2h} \tau^2(y,S)\right) dy$$

$$\leq D_1 \sum_{k=1}^\infty (k+1) \log(1/h) e^{-k a h \log(1/h)} \int_{A_k} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{4h} \tau^2(y,S)\right) dy$$

$$\leq D_1 \sum_{k=1}^\infty (k+1) \log(1/h) h^{k a / 2} \int_{A_k} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{4h} \tau^2(y,S)\right) dy$$

$$\leq D_2 h \log(1/h) \sum_{k=1}^\infty (k+1) \int_{A_k} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{4h} \tau^2(y,S)\right) dy$$

$$\leq D_2 h \log(1/h) \int_{S^c} \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{1}{4h} \tau^2(y,S)\right) dy \leq D_2 h \log(1/h).$$

The final integral is finite since in a small neighborhood around $S$, $\tau^2$ is comparable to the squared Euclidean distance, while, further away, it is infinity outside a compact set.

In conclusion, by comparing the largest terms in $A_0$ and $A_k$, $k \geq 1$, we get

$$\int_{S^c} |\log \gamma^h(y)| \gamma^h(y) dy \leq O \left( \sqrt{h} (\log(1/h))^{d/2+2} \right) + O(h \log(1/h)) = O \left( \sqrt{h} (\log(1/h))^{d/2+2} \right),$$

and we have proved (31) and completed the proof of Lemma 3 (iv).

**Proof of the Theorem**

We start by proving the limit of $K_h$.

Consider the exponential tilting $\tilde{\mu}_h$ from Lemma 3 (iv). Recall that $\mu_h$ denotes the joint density $\rho_0(x)p_h(x,y)$. Then, from (20),

$$\frac{d\tilde{\mu}_h}{d\mu_h}(x,y) = \frac{\tilde{\mu}_h(x,y)}{p_h(x,y)} = \frac{1}{Z_h(x)} \exp \left( \frac{1}{h} (\psi(x) + \psi^*(y)) \right)$$

For any $\nu \in M_1(\Sigma)$, absolutely continuous with respect to $\tilde{\mu}_h$,

$$H(\nu \mid \tilde{\mu}_h) = \nu \left( \log \frac{d\nu}{d\tilde{\mu}_h} \right) = \nu \left( \log \frac{d\nu}{d\mu_h} - \log \frac{d\tilde{\mu}_h}{d\mu_h} \right)$$

$$= H(\nu \mid \mu_h) - \frac{1}{h} \nu (\psi(x) + \psi^*(y)) + \nu (\log Z_h(x)) .$$

If moreover $\nu \in \Pi(\rho_0, \rho_1)$, then, by Kantorovich duality,

$$\frac{1}{h} \int (\psi(x) + \psi^*(y)) d\nu = \frac{1}{h} \left[ \int \psi(x) \rho_0(x) dx + \int \psi^*(y) \rho_1(y) dy \right] = \frac{1}{h} \mathbb{W}_g (\rho_0, \rho_1)$$

and $\nu (\log Z_h(x)) = \int \log (Z_h(x)) \rho_0(x) dx < \infty$ by Lemma 3 (iii).

Thus, alternatively, $K_h$ from (33) is given by

$$K_h = \frac{1}{h} \mathbb{W}_g (\rho_0, \rho_1) - \int \log (Z_h(x)) \rho_0(x) dx + \inf_{\nu \in \Pi(\rho_0, \rho_1)} H(\nu \mid \tilde{\mu}_h).$$
Since $H(\nu \mid \hat{\mu}_h) \geq 0$, we immediately get the lower bound
\begin{equation}
K_h - \frac{1}{h} \mathbb{W}_2(g(\rho_0, \rho_1)) \geq - \int \log (Z_h(x)) \rho_0(x) dx = - \frac{1}{2} \log \det \nabla^2 g(0) + H(\rho_1 \mid \mu_y) + o(1),
\end{equation}
as $h \to 0+$ (by Lemma 3 (iii)).

We now obtain a matching upper bound. Let $\rho_1^h$ denote $\pi^1(\hat{\mu}_h)$ as in Lemma 3 (iv). Consider the function $\zeta_1, \zeta_2, \zeta_3$ appearing in (28). Let

\[ \zeta(y) = \zeta_1(y) + h^{1/2} \zeta_2(y) + h^2 \zeta_3(y). \]

Since $\zeta$ is bounded and $\rho_1(y)$ is strictly positive on $S$, for all small enough $h$, the function $\rho_1^h - h \zeta(y)$ is positive for all $y \in S$. Moreover

\[ \int_S \rho_1^h(y) \left( 1 - h \frac{\zeta(y)}{\rho_1^h(y)} \right) dy = \int_S (\rho_1^h - h \zeta(y)) dy = O(h^{5/2}) + \int_S \rho_1(y) dy = 1 + O(h^{5/2}). \]

Consider a modified density $\eta_1^h(y)$ given by

\[ \eta_1^h(y) = \frac{1}{N_h} \rho_1^h(y) \left( 1 - \frac{\zeta(y)}{\rho_1^h(y)} \right) 1\{y \in S\}. \]

Then, $N_h = 1 + O(h^{5/2})$. This density has the property that if $\| \cdot \|_{TV}$ refers to the total variation distance between measures, then, from the expansion (28) it follows that and

\[ \| \eta_1^h - \rho_1 \|_{TV} = \left( 1 - O(h^{5/2}) \right) \left( \rho_1^h - h \zeta(y) \right) - \rho_1(y) = O(h^{5/2}). \]

The second key property is that if we modify $\bar{\mu}_h(x, y)$ by

\[ \bar{\mu}_h(x, y) = \bar{\mu}_h(x, y) \frac{1}{N_h} \rho_0(x) \left( 1 - \frac{\zeta(y)}{\rho_1^h(y)} \right) 1\{y \in S\} \]

\[ = \frac{1}{N_h} \rho_0(x) \exp \left( \frac{1}{h} D[y \mid x^*] \right) \left( 1 - \frac{\zeta(y)}{\rho_1^h(y)} \right) 1\{y \in S\}. \]

Then $\bar{\mu}_h(x, y)$ satisfies all the properties stated in Lemma 3. This is because $\bar{\mu}_h$ and $\bar{\mu}_h$ are exponentially equivalent since $\log N_h = O(h^{5/2}) \to 0$. Notice that $X$, under $\bar{\mu}_h$ is no longer distributed as $\rho_0$. Let this density be denoted by $\rho_0^h(x)$. Define

\[ \bar{\mu}_h(x, y) = \frac{\rho_0(x)}{\rho_0^h(x)} \bar{\mu}_h(x, y). \]

It is easy to see that $\bar{\mu}_h$ satisfies both properties that $X \sim \rho_0$, the total variation distance of $Y$ from $\rho_1$ is $O(h^{5/2})$ and $\bar{\mu}_h$ is exponentially equivalent to $\hat{\mu}_h$ in the sense that $\log (d\bar{\mu}_h / \hat{\mu}_h) \to 0$ uniformly.

Let $\gamma^h$ denote the density of $Y$ sampled from $\bar{\mu}_h$. Since the total variation distance between $\gamma^h$ and $\rho_1$ is $O(h^{5/2})$ and they are compactly supported, the quadratic Wasserstein distance $\mathbb{W}_2^2(\gamma^h, \rho_1) = O(h^{5/2})$. Let $\kappa_h$ denote the Brenier map transporting $\gamma^h$ to $\rho_1$. Thus, if $(X, Y)$ is a sample from $\bar{\mu}_h$, then $(X, Y') := \nabla \kappa(Y)$ is an element in $\Pi(\rho_0, \rho_1)$, and $E \|Y - Y'\|_2 = O(h^{5/4})$. Let $\hat{\mu}_h$ be the density of $(X, Y')$.

We claim that $H(\hat{\mu}_h \mid \bar{\mu}_h) \to 0$, as $h \to 0+$. Let us evaluate explicitly the relative entropy appearing on the right hand side. By the change of variable formula,

\[ \frac{\hat{\mu}_h(u, v)}{\bar{\mu}_h(u, \nabla \kappa(v))} = \frac{1}{|\nabla^2 \kappa_h(v)|}, \]
Thus, asymptotically, 

\[
H(\tilde{\mu}_h | \tilde{\pi}_h) = \mu_h \left( \log \frac{\mu_h(u, v)}{\mu_h(u, \nabla \kappa(v))} \right)
\]

\[
= \mu_h \left( \log \frac{d\tilde{\mu}_h(u, v)}{d\mu_h(u, \nabla \kappa(v))} \right) - \tilde{\mu}_h \left( \log \frac{\tilde{\mu}_h(u, v)}{\mu_h(u, \nabla \kappa(v))} \right)
\]

\[
= - \int \log \left| \nabla^2 \kappa^*_h(y) \right| \rho_1(y)dy - \tilde{\mu}_h \left( \log \frac{\tilde{\mu}_h(u, v)}{\mu_h(u, \nabla \kappa(v))} \right).
\]

By using the same logic as in (26), the first integral on the right is \( \text{Ent}(\rho_1) - \text{Ent}(\gamma_1^h) \) which goes to zero in the same way as Lemma 3 (iv). For the other term, let 

\[
(X, Y, Y' = \nabla \kappa(Y))
\]

be as before, then

\[
\tilde{\mu}_h \left( \log \frac{\tilde{\mu}_h(u, v)}{\mu_h(u, \nabla \kappa(v))} \right) = \frac{1}{h} \mathbb{E} \left[ D[Y' | X] - D[Y | X^*] \right] + O(h)
\]

\[
\leq O \left( \frac{1}{h} \mathbb{E} \left\| Y - Y' \left\|_2 \right. \right) + O(h) = O \left( h^{1/4} \right) + O(h) = O(h^{1/4}).
\]

We are using above the the divergence is Lipschitz which follows from the Lipschitz property of \( g \) and \( c \)-concave functions. Anyway, this proves that \( H(\tilde{\mu}_h | \tilde{\pi}_h) \to 0 \), and, hence, by exponential equivalence \( H(\tilde{\mu}_h | \tilde{\pi}_h) \to 0 \). Thus

\[
K_h - \frac{1}{h} \mathbb{W}_g(\rho_0, \rho_1) \leq - \frac{1}{2} \log \det \nabla^2 g(0) + H(\rho_1 | \mu_g) + o(1).
\]

Combining with the lower bound in (33), we have the statement for \( K_h \).

Now, for \( K'_h \) use the relationship from (10)

\[
K'_h = K_h + \text{Ent}(\rho_0) - \int \log \Lambda_h(x)\rho_0(x)dx.
\]

By Lemma 3 (i),

\[
\lim_{h \to 0^+} \int \log \Lambda_h(x)\rho_0(x)dx - \frac{d}{2} \log(2\pi h) = - \frac{1}{2} \log \det \nabla^2 g(0).
\]

Thus, asymptotically,

\[
K'_h = \text{Ent}(\rho_0) - \frac{d}{2} \log(2\pi h) + \frac{1}{h} \mathbb{W}_g(\rho_0, \rho_1) + H(\rho_1 | \mu_g) + o(1).
\]

This gives us the statement regarding \( K'_h \).

The rest of the statement follows from Lemma 3 (iii). \( \square \)

3. Dirichlet Transport and the Proof of Theorem 2

Recall the cost function in (15):

\[
c(p, q) = \log \left( \frac{1}{n} \sum_{i=1}^{n} \frac{q_i}{p_i} \right) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{q_i}{p_i}, \quad p, q \in \Delta_n.
\]

The so-called exponential coordinate system of the unit simplex refers to the following map from \( \Delta_n \to \mathbb{R}^{n-1} \):

\[
\Theta(p) = \log \left( p_i/p_n \right), \quad i = 1, 2, \ldots, n - 1.
\]
The inverse of this map takes \((\theta_1, \ldots, \theta_{n-1}) \in \mathbb{R}^{n-1}\) to an element \(p \in \Delta_n\), where
\[
p_i = \frac{e^{\theta_i}}{1 + \sum_{j=1}^{n-1} e^{\theta_j}}, \quad i = 1, 2, \ldots, n-1, \quad p_n = \frac{1}{1 + \sum_{j=1}^{n-1} e^{\theta_j}}.
\]

If \(\Theta(p) = \theta = (\theta_1, \ldots, \theta_{n-1})\) and \(\Theta(q) = (\rho_1, \ldots, \rho_{n-1})\) are the respective exponential coordinates of \(p\) and \(q\), one can express \(c(p, q)\) as
\[
\overline{c}(\theta, \rho) = \log \left( \frac{1}{n} + \frac{1}{n} \sum_{i=1}^{n-1} e^{\rho_i - \theta_i} \right) - \frac{1}{n} \sum_{i=1}^{n-1} (\rho_i - \theta_i) = g(\theta - \rho),
\]
where \(g\) is indeed a convex function. For a detailed analysis using this coordinate system see [23].

However, the exponential coordinate system is not the most natural coordinate system for this transport problem. The multiplicative group \(\odot\) appears to give it more natural properties, especially related to the behavior of entropy. Hence, the global coordinate system we use on \(\Delta_n\) is given by the first \((n - 1)\) coordinates of \(p \in \Delta_n\). It will be convenient to have a new notation:

\[
\tilde{p} = (p_1, \ldots, p_{n-1}), \quad p_i = \tilde{p}_i, \quad \text{for} \quad i = 1, 2, \ldots, n-1 \quad \text{and} \quad p_n = 1 - \sum_{i=1}^{n-1} \tilde{p}_i.
\]

In the calculations below, all derivatives, Taylor expansions, and integrals will be performed with respect to the Lebesgue measure on the set
\[
\left\{ (p_1, \ldots, p_{n-1}) \mid p_i > 0, \forall i, \sum_{i=1}^{n-1} p_i < 1 \right\}.
\]

We now describe the structure of the Monge solution.

A function \(\varphi : \Delta_n \to [-\infty, \infty)\) is called exponentially concave if \(e^\varphi\) is a (nonnegative) concave function on \(\Delta_n\). Hence \(\varphi\) is differentiable Lebesgue almost everywhere. In particular, its gradient \(\nabla \varphi\) is a.e. defined. In case \(\varphi\) is twice continuously differentiable as a function of \(\tilde{p}\), one can define the following nonpositive definite matrix
\[
(36) \quad \nabla^2 \varphi(p) = \nabla^2 \varphi(p) + \nabla \varphi(p) \nabla^T \varphi(p) \leq 0.
\]

**Definition 4 (Portfolio map).** Let \(\varphi\) be exponentially concave on \(\Delta_n\). When \(\varphi\) is differentiable at \(r \in \Delta_n\), we define \(\pi(r) \in \Delta_n\) by
\[
(37) \quad (\pi(r))_i = r_i (1 + \nabla_{e_i - r} \varphi(r)), \quad i = 1, \ldots, n,
\]
where \(\{e_1, \ldots, e_n\}\) is the standard basis of \(\mathbb{R}^n\) and \(\nabla_{e_i - r}\) is the directional derivative in the direction \(e_i - r\). We call \(\pi\) the portfolio map generated by \(\varphi\).

The fact that the range of \(\pi\) is contained in \(\Delta_n\) has been first observed by Fernholz in the context of stochastic portfolio theory (hence the name). See [13,23]. The map \(\pi(r)\) is the multiplicative equivalent of the additive map \(x \mapsto x - \nabla \phi(x)\), for a convex function \(\phi\), for the quadratic Wasserstein transport problem.

Consider two probability densities \(\rho_0, \rho_1\) that are compactly supported in \(\Delta_n\). Consider the optimal transport from \(\rho_0\) to \(\rho_1\) with cost \(c\). There exists a unique solution to the Monge problem whose solution can be described in terms of the portfolio map of an exponentially concave function. The following is a restatement of [24] Theorem 4 and eqn. (38)].
There exists an exponentially concave function \( \varphi : \Delta_n \to \mathbb{R} \) such that the following statements hold.

(i) If \( \pi \) is the portfolio map generated by \( \varphi \), the mapping \( q \mapsto T(q) = p \) where
\[
P^{-1} := q^{-1} \circ \pi(q),
\]
is \( \rho_1 \) a.e. defined and pushforwards \( \rho_1 \) to \( \rho_0 \).

(ii) The deterministic coupling \( (T(q), q) \) is the optimal Monge solution and is \( \rho_1 \)-a.e. unique.

To drive home the similarity with the Wasserstein case we will denote the pair \( (T(q), q) \) by \( (p, p^\ast) \) which is well-defined for almost every \( p \) in the support of \( \rho_0 \).

The divergence for an exponentially concave function is called the \( L \)-divergence and was originally defined in \cite{tv1}. More detailed studies on the geometry of \( L \)-divergence can be found in \cite{ty2, ty1, tv2}. The following definition is equivalent to Definition \( \text{[5]} \) by passing to exponential coordinates (see \cite{tv2}) but avoids developing a duality theory for exponentially concave functions.

**Definition 5** (**L-divergence**). Let \( \varphi \) be a differentiable exponentially concave function on \( \Delta_n \). The \( L \)-divergence of \( \varphi \) is defined by
\[
L[r | r'] = \log (1 + \nabla \varphi(r') \cdot (r - r')) - (\varphi(r) - \varphi(r')), \quad r, r' \in \Delta_n,
\]
where \( a \cdot b \) is the Euclidean dot product.

By the exponential concavity of \( \varphi \), it can be shown that \( L[r | r'] \geq 0 \) and \( L[r | r] = 0 \). If \( e^\varphi \) is strictly concave, then \( L[r | r'] > 0 \) for all \( r \neq r' \). Using the definition of \( \pi \) (see \( \text{[37]} \)), we can write
\[
L[r | r'] = \log \left( \sum_{i=1}^n \left( \pi(r') \right)_i \frac{r_i}{r'_i} \right) - (\varphi(r) - \varphi(r')).
\]

As \( r - r' \to 0 \), we can now consider a Taylor approximation to \( L \)-divergence. More details about the following calculations can be found in \cite{tv2} Section 4.2]. Let \( L(r) := -\nabla^2 \varphi(r) \) from \( \text{[36]} \) denote the positive semidefinite matrix of the \( L \)-divergence under the coordinate system \( r \mapsto \tilde{r} \). We make the following assumptions that mirror Assumptions \( \text{[4]} \) and \( \text{[2]} \).

**Assumption 3.** Assume that \( \rho_0, \rho_1 \) are continuous and compactly supported on \( \Delta_n \). Let \( S := \text{spt}(\rho_1) \) denote the support of \( \rho_1 \). Assume \( \rho_1 \) is smooth and \( \inf_{y \in S} \rho_1(y) > 0 \). Moreover, \( S \) has a nice boundary in the sense that, if we consider its \( \epsilon \)-expansion \( S^\epsilon := \{ y : \inf_{z \in S} \| y - z \| \leq \epsilon \} \), then \( \text{Vol} \left( S^\epsilon \right) - \text{Vol}(S) = O(\epsilon) \), as \( \epsilon \to 0+ \).

**Assumption 4.** Assume that for every \( r' \) on the support of \( \rho_1 \) the following quadratic approximation to the divergence holds uniformly on compact sets in \( \Delta_n \).
\[
R[r | r'] := L[r | r'] - \frac{1}{2}(\bar{r} - r')^T L(r')(\bar{r} - r') = o \left( |\bar{r} - r'|^2 \right), \quad \text{as } r \to r',
\]
where \( L(r') = -\nabla^2 \varphi(r'), \quad r' \in \text{spt}(\rho_1) \) is assumed to be continuous. The family \( L(\cdot) \) is also assumed to be smooth and uniformly elliptic in the sense that there exists an \( \epsilon > 0 \) such that the eigenvalues of \( L(r') \) lie in \( (\epsilon, 1/\epsilon) \) for all \( r' \in \text{spt}(\rho_1) \).
3.1. The Dirichlet transport. Recall the symmetric Dirichlet distribution \(\text{Diri}(\lambda)\) from the Introduction [10] and the density \(\nu_0\) of \(\text{Diri}(0)\). Fix \(p \in \Delta_n\), generate \(G \sim \text{Diri}(\lambda)\) and consider the random variable \(p \circ G\). The density of the random variable (see [24]) at any \(q \in \Delta_n\) is

\[
F_\lambda(p,q) = \frac{\Gamma(\lambda)}{(\Gamma(\lambda/n))^n} \prod_{i=1}^n \frac{q_i}{p_i} \left(\sum_{i=1}^n q_i/p_i\right)^{-\lambda} = \frac{\Gamma(\lambda)n^{-\lambda}}{(\Gamma(\lambda/n))^n} \nu_0(q) \exp(-\lambda c(p,q)).
\]

We start by carefully analyzing the normalizing constant in the above density as \(\lambda \to \infty\). By Stirling approximation to the gamma function

\[
\log \frac{\Gamma(\lambda)n^{-\lambda}}{(\Gamma(\lambda/n))^n} = \log(\lambda) - \lambda \log n - n \log \Gamma(\lambda/n)
\]

\[
\sim \left(\lambda - \frac{1}{2}\right) \log \lambda - \lambda + \frac{1}{2} \log(2\pi) - \lambda \log n - \left(\lambda - \frac{n}{2}\right) \log \left(\frac{\lambda}{n}\right) + \frac{n}{2} \log(2\pi)
\]

\[
= \frac{(n-1)}{2} \log \lambda - \lambda + \frac{n}{2} \log n - \frac{(n-1)}{2} \log(2\pi)
\]

\[
= \frac{(n-1)}{2} \log \left(\frac{\lambda}{2\pi}\right) - \frac{n}{2} \log n = O(\log \lambda).
\]

This gives us a transition probability on \(\Delta_n\) such that \(\lim_{\lambda \to \infty} -\frac{1}{\lambda} \log F_\lambda(q \mid p) = c(p,q)\). Thus \(h := \frac{1}{\lambda}\) is a measure of noise, and as \(\lambda \to 0^+\), we recover the optimal transport with cost \(c\).

As \(\lambda \to 0^+\) (or, equivalently \(h \to \infty\)), the symmetric Dirichlet distribution vaguely converges to \(\text{Diri}(0)\). In particular, from (42) notice that

\[
F_\lambda(p,q) = \frac{\Gamma(\lambda)n^{-\lambda}}{(\Gamma(\lambda/n))^n} \nu_0(q)e^{-\lambda c(p,q)} \sim n^{-n/2} \left(\frac{\lambda}{2\pi}\right)^{(n-1)/2} \nu_0(q)e^{-\lambda c(p,q)}, \quad \text{for large \(\lambda\),}
\]

where \(\sim\) means that the leading terms in the log of both sides are the same asymptotically as \(\lambda \to \infty\).

The proof of Theorem 2 is very similar to that Theorem 1. We will reuse similar notations to stress this point and skip similar steps in the proof. First we change the parameter \(\lambda\) to \(h = 1/\lambda\) in the transition density \(F_\lambda\) by defining

\[
f_h(p,q) = \frac{1}{(1/h)^{n^{-1/h}}} F_\lambda(p,q) = \frac{1}{(1/h)^{n^{-1/h}}} \nu_0(q)e^{-\frac{1}{h} c(p,q)} \sim \frac{n^{-n/2}}{(2\pi h)^{(n-1)/2}} \nu_0(q)e^{-\frac{1}{h} c(p,q)}
\]

This gives us a joint density \(\mu_h(p,q) = \rho_0(p)f_h(p,q)\).

We will now define an exponential tilting of the transition probability \(f_h\) that penalizes based on \(L\)-divergence from the Monge map. Consider the map \(p = T(q)\) from (35). Let \(p^\ast = T^{-1}(p)\) denote the inverse map, which is well-defined on the support of \(\rho_0\). Then \(p \mapsto p^\ast\) is the Monge map transporting \(\rho_0\) to \(\rho_1\) with cost \(c\).
Let $\Phi(p, q) := c(p, q) - L[q \mid p^*]$. More explicitly, $\Phi(p, q)$ can be found from (41) by replacing $\hat{q}$ by $p^*$ and $\tilde{q}$ by $q$.

$$c(p, q) = \log \left( \frac{1}{n} \sum_{i=1}^{n} q_i \right) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{q_i}{p_i}$$

$$= L[q \mid p^*] + \frac{1}{n} \sum_{i=1}^{n} \frac{p_i^2}{p_i} + \varphi(q) - \varphi(p^*) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{q_i}{p_i}$$

$$= L[q \mid p^*] + c(p, p^*) + \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_i^2}{p_i} + \varphi(q) - \varphi(p^*) - \frac{1}{n} \sum_{i=1}^{n} \log \frac{q_i}{p_i}$$

$$= L[q \mid p^*] + c(p, p^*) + \frac{1}{n} \sum_{i=1}^{n} \log p_i^* - \frac{1}{n} \sum_{i=1}^{n} \log q_i + \varphi(q) - \varphi(p^*).$$

Thus

$$(44) \quad \Phi(p, q) = c(p, p^*) + \frac{1}{n} \sum_{i=1}^{n} \log p_i^* - \frac{1}{n} \sum_{i=1}^{n} \log q_i + \varphi(q) - \varphi(p^*).$$

As before, because $\rho_0$ and $\rho_1$ are compactly supported on $\Delta_n$, we will assume without loss of generality that $c(p, q) = \infty$ outside a compact subset of $\Delta_n$. Hence, all integrals below are finite.

The normalizing function

$$Z_h(p) := \frac{1}{\nu_0(p^*)} \int_{\Delta_n} \exp \left( -\frac{1}{h} L[q \mid p^*] \right) \nu_0(q) dq,$$

defines an exponential tilting of the transition density $f_h$ on $\Delta_n$:

$$\bar{f}_h(p, q) = \frac{1}{Z_h(p) \nu_0(p^*)} \exp \left[ \frac{1}{h} \Phi(p, q) \right] f_h(p, q)$$

$$= \frac{\Gamma(1/h)n^{-1/h}}{(\Gamma(1/nh))^n} \frac{\nu_0(q)}{Z_h(p) \nu_0(p^*)} \exp \left[ -\frac{1}{h} c(p, q) + \frac{1}{h} \Phi(p, q) \right]$$

$$= \frac{\Gamma(1/h)n^{-1/h}}{(\Gamma(1/nh))^n} \frac{\nu_0(q)}{Z_h(p) \nu_0(p^*)} \exp \left( -\frac{1}{h} L[q \mid p^*] \right).$$

Observe that, for $h \approx 0$, $\exp \left( -\frac{1}{h} L[q \mid p^*] \right)$ is very small, unless $q \approx p^*$. In the latter case, we have the quadratic approximation from (41)

$$\exp \left[ -\frac{1}{h} L[q \mid p^*] \right] \approx \exp \left( -\frac{1}{2h}(\tilde{q} - p^*)^T L(p^*) (\tilde{q} - p^*) \right).$$

Suppose we forget that we are on the simplex, and consider the measure with the unnormalized density as above, it gives us a Gaussian distribution with mean $\tilde{p}^*$ and a covariance matrix $A$ such that $A^{-1} = hL(p^*) = -h\nabla^2 \varphi(p^*)$. Because $\rho_0$ is compactly supported in $\Delta_n$, there is some $\delta > 0$ such that $B(p, \delta) \subset \Delta_n$. Hence, for small values of $h$, $p_i^* / q_i \approx 1$ since $q \in B(p, \delta)$ with exponentially high probability.

Thus, $\bar{f}_h$ is approximately the Gaussian distribution

$$q_h(p, q) = \frac{|J|^{1/2}(p^*)}{(2\pi h)^{\alpha-1/2}} \exp \left( -\frac{1}{2h}(\tilde{q} - p^*)^T L(p^*) (\tilde{q} - p^*) \right).$$
where $|J|(p^*)$ is the determinant of $L(p^*)$. Hence, using the Stirling approximation from (43), we get that

$$Z_h(p) \approx \frac{n^{-n/2}}{|J|^{1/2}(p^*)}, \text{ as } h \to 0^+.$$ 

Hence the proof of the following lemma follows from that of Lemma 3.

**Lemma 4.** Under Assumptions 3 and 4, the following statements hold.

(i) The normalizing constant $Z_h(p)$ has the following limit locally uniformly in $p$ in the support of $\rho_0$.

$$\lim_{h \to 0^+} Z_h(p) \sqrt{|J|(p^*)} = n^{-n/2}.$$ 

(ii) Moreover,

$$\lim_{h \to 0^+} -\int \log Z_h(p) \rho_0(p) dp = \frac{1}{2} (\text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0))$$ 

$$- \frac{n}{2} C(\rho_0, \rho_1) + \int \log \nu_0(q) \rho_1(q) dq.$$ 

(iii) Finally, let $\rho_1^h$ denote the law of $Y$ under the joint distribution $\tilde{\mu}_h = \rho_0(p) \tilde{f}_h(p,q)$, then, $\rho_1^h$ converges weakly to $\rho_1$ as $h \to 0^+$, and

$$\lim_{h \to 0^+} \text{Ent}(\rho_1^h) = \text{Ent}(\rho_1).$$ 

**Proof.** The proofs of (i) and (iii) are almost identical to that of Lemma 3(ii) and (iv) and have already been outlined above.

For the proof of (ii) we use [24, Section 4.2, 4.3], especially the proof of Theorem 16. The rest of the argument alludes to the notation used in that reference. However, in order to use this in our case, we will need to unpack the notations.

Suppose that $P_0$ and $P_1$ are the two probability measures on $\Delta_n$ with compactly supported densities. Consider the problem of transporting $P_0$ to $P_1$ with cost $c$. As shown in [24, Theorem 4], there exists an exponentially concave function $\varphi^*$ such that $p \mapsto q = T(p) = p \odot \pi(p^{-1})$. This is the dual map to [38] that maps $q$ to $p$. However, $L$ is still the matrix of $L$-divergence for this exponentially concave function $\varphi^*$. Then (see [24, Section 4.3, in Proof of Theorem 16 put $t = 1$)

$$\text{Ent}_0(P_1) = \text{Ent}_0(P_0) + n \log \frac{1}{n} - n C(P_0, P_1)$$

$$- \int \log \det (L(p^{-1})) dP_0 + 2 \int \log \nu_0 \rho_1^{-1} dP_0. \tag{45}$$

(Recall that inversion in $\odot$ is akin to sign inversion in the quadratic Wasserstein.)

We now invert this duality to translate (45) in the current set-up. Let $\rho_0^-$ (respectively, $\rho_1^-$) denote the density of the push-forward of $\rho_0$ (respectively, $\rho_1$) under the map $p \mapsto p^{-1}$. Let $P_0$ be the measure with density $\rho_0^-$ and $P_1$ be the measure with density $\rho_1^-$.

First, note from the paragraph following (45), our cost function satisfies the following multiplicative symmetry $c(p, q) = c(q^{-1}, p^{-1})$. Hence $C(P_0, P_1) = C(\rho_0, \rho_1)$. This is expected since inversions are multiplicative sign changes. We now rewrite the relation (45) in terms of $\rho_0$ and $\rho_1$. 


The Jacobian of the transformation $\tilde{p} \mapsto \tilde{r} : p^{-1}$ has been computed in [24 eqn. (72)]:

$$
\begin{aligned}
\frac{\partial (r_1, \ldots, r_{n-1})}{\partial (p_1, \ldots, p_{n-1})} &= \frac{r_1 \cdots r_n}{p_1 \cdots p_n} = \frac{\nu_0(p)}{\nu_0(r)}.
\end{aligned}
$$

Hence, by the change of variable formula

$$
\begin{aligned}
\text{Ent}_0(P_1) &= \text{Ent}(P_1) - \int \log \nu_0(p) dP_1 \\
&= \text{Ent}(\rho_0) - \rho_0 \left( \log \left( \frac{\partial (r_1, \ldots, r_{n-1})}{\partial (p_1, \ldots, p_{n-1})} \right) \right) - \rho_0 \left( \log \nu_0(r) \right) \\
&= \text{Ent}(\rho_0) - \rho_0 \left( \log \nu_0(p) \right) = \text{Ent}_0(\rho_0).
\end{aligned}
$$

Similarly $\text{Ent}_0(P_0) = \text{Ent}_0(\rho_1)$. Again, these relationships are expected since inversions with respect to $\odot$ are multiplicative analogs of sign-changes and $\nu_0$ is the multiplicative analog of the Lebesgue measure.

Finally, by a change of variable,

$$
\begin{aligned}
- \int \log |L(p^{-1})| dP_0 &= -\rho_1 \left( \log |J| (q) \right) = -\rho_0 \left( \log |J| (p^*) \right), \quad \text{and} \\
2 \int \log \nu_0(p^{-1}) dP_0 &= 2\rho_1 \left( \log \nu_0(q) \right).
\end{aligned}
$$

Hence, if we translate all the terms from [45], we get

$$
\begin{aligned}
\text{Ent}_0(\rho_0) &= \text{Ent}_0(\rho_1) + n \log \frac{1}{n} - nC(\rho_0, \rho_1) - \rho_0 \left( \log |J| (p^*) \right) + 2\rho_1 \left( \log \nu_0(q) \right).
\end{aligned}
$$

Rearranging terms gives us

$$
\rho_0 \left( \log |J| (p^*) \right) = \text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0) - n \log n - nC(\rho_0, \rho_1) + 2\rho_1 \left( \log \nu_0(q) \right).
$$

Hence, by part (i) of this lemma,

$$
\begin{aligned}
\lim_{h \to 0^+} - \int \log Z_h(p) \rho_0(p) dp &= \frac{1}{2} \rho_0 \left( \log |J| (p^*) \right) + \frac{n}{2} \log n \\
&= \frac{1}{2} \left( \text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0) \right) - \frac{n}{2} C(\rho_0, \rho_1) + \rho_1 \left( \log \nu_0(q) \right).
\end{aligned}
$$

This completes the proof of the lemma. \hfill \Box

**Proof of Theorem 2.** We now complete the proof exactly as the proof of Theorem 1 Define $\tilde{\mu}_h(p, q) = \rho_0(p) \tilde{f}_h(p, q)$. Then, for any $\nu \in \Pi(\rho_0, \rho_1)$,

$$
\begin{aligned}
H(\nu \mid \mu_h) &= \nu \left( \log \frac{d\tilde{\mu}_h}{d\mu_h} \right) + H(\nu \mid \tilde{\mu}_h) \\
&= \frac{1}{h} \nu(\Phi(p, q)) - \nu \left( \log Z_h(p) \right) - \nu \left( \log \nu_0(p^*) \right) + H(\nu \mid \tilde{\mu}_h) \\
&= \frac{1}{h} C(\rho_0, \rho_1) - \int \log Z_h(p) \rho_0(p) dp - \rho_0 \left( \log \nu_0(p^*) \right) + H(\nu \mid \tilde{\mu}_h).
\end{aligned}
$$

The last line above comes from [44], in particular, from the facts that

$$
\nu \left( c(p, p^*) \right) = \rho_0 \left( c(p, p^*) \right) = C(\rho_0, \rho_1).
$$

$$
\rho_0 \left( \frac{1}{n} \sum_{i=1}^{n} \log p_i^* \right) = \rho_1 \left( \frac{1}{n} \sum_{i=1}^{n} \log q_i \right), \quad \rho_0 \left( \varphi(p^*) \right) = \rho_1 \left( \varphi(q) \right).
$$
As in proof of Theorem 1 as $h \to 0+$, we get
\[
K_h(\rho_0, \rho_1) = \frac{1}{h} C(\rho_0, \rho_1) - \int \log Z_h(p) \rho_0(p) dp - \rho_0 \left( \log \nu_0(p^*) \right) + o(1)
\]
\[
= \frac{1}{h} C(\rho_0, \rho_1) - \int \log Z_h(p) \rho_0(p) dp - \rho_1 \left( \log \nu_0(q) \right) + o(1)
\]
\[
= \left( \frac{1}{h} - \frac{n}{2} \right) C(\rho_0, \rho_1) + \frac{1}{2} \left( \text{Ent}_0(\rho_1) - \text{Ent}_0(\rho_0) \right) + o(1),
\]
from Lemma 4(ii). This gives us Theorem 2. \qed

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. *Communications in Mathematical Physics*, 307(3):791–815, 2011.

[2] S.-i. Amari. *Information Geometry and Its Applications*. Springer, 2016.

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: In Metric Spaces and in the Space of Probability Measures, 2nd. Edition*. Lectures in Mathematics. Birkhäuser, Basel, 2008.

[4] G. Conforti. A second order equation for Schrödinger bridges with applications to the hot gas experiment and entropic transportation cost. *Probability Theory and Related Fields*, 174(1–2):1–47, 2019.

[5] M. Cuturi. Sinkhorn distance: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. MIT Press, Cambridge, MA, 2013.

[6] H. Dietert. Characterisation of gradient flows on finite state Markov chains. *Electron. Commun. Probab.*, 20(29):8 pp., 2015.

[7] M. H. Duong, V. Laschos, and M. Renger. Wasserstein gradient flows from large deviations of many-particle limits. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(4):1166–1188, 2013. Erratum at [www.wias-berlin.de/people/renger/Erratum/DLR2015ErratumFinal.pdf](http://www.wias-berlin.de/people/renger/Erratum/DLR2015ErratumFinal.pdf).

[8] M. Erbar. Gradient flows of the entropy for jump processes. *Annales de l’Institut Henri Poincaré*, 50(3):920–945, 2014.

[9] M. Erbar, M. Fathi, V. Laschos, and A. Schlichting. Gradient flow structure for McKean-Vlasov equations on discrete spaces. *Discrete & Continuous Dynamical Systems - A*, 36:6799–6833, 2016.

[10] M. Erbar and J. Maas. Ricci curvature of finite Markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, 2012.

[11] M. Erbar and J. Maas. Gradient flow structures for discrete porous medium equations. *Discrete & Continuous Dynamical Systems - A*, 34:1355–1374, 2014.

[12] M. Erbar, J. Maas, and D. R. M. Renger. From large deviations to Wasserstein gradient flows in multiple dimensions. *Electronic Communications in Probability*, 20(89):1–12, 2015.

[13] E. R. Fernholz. *Stochastic Portfolio Theory*. Applications of Mathematics. Springer, 2002.

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

[15] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[16] I. Karatzas, W. Schachermayer, and B. Tschiderer. Pathwise Otto calculus. Available at math arxiv 1811.08686v2, 2019.

[17] G. Khan and J. Zhang. The Kähler geometry of certain optimal transport problems. Available at math arxiv:1812.00032v3, 2019.

[18] C. Léonard. From the Schrödinger problem to the Monge-Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[19] 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.

[20] J. Maas. Gradient flows of entropy for finite Markov chains. *Journal of Functional Analysis*, 261:2250–2292, 2011.
[21] S. Pal. Embedding optimal transports in statistical manifolds. *Indian Journal of Pure and Applied Mathematics*, 48(4):541–550, 2017.
[22] S. Pal and T.-K. L. Wong. The geometry of relative arbitrage. *Mathematics and Financial Economics*, 10:263–293, 2016.
[23] S. Pal and T.-K. L. Wong. Exponentially concave functions and a new information geometry. *The Annals of Probability*, 46(2):1070–1113, 2018.
[24] S. Pal and T.-K. L. Wong. Multiplicative Schrödinger problem and the Dirichlet transport. arXiv preprint 1806.05649, 2018.
[25] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
[26] T.-K. L. Wong. Optimization of relative arbitrage. *Annals of Finance*, 11(3-4):345–382, 2015.
[27] T.-K. L. Wong. Logarithmic divergences from optimal transport and Rényi geometry. *Information Geometry*, 1(1):39–78, 2018.

University of Washington, Seattle

E-mail address: soumikpal@gmail.com