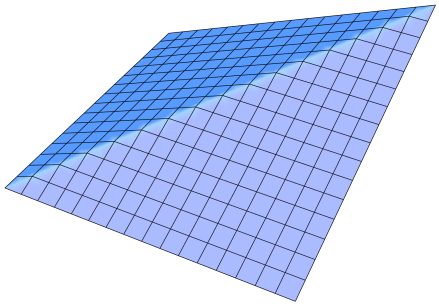

STA 4519H: Optimal Transport: Theory and Algorithms

Professor Leonard Wong • Fall 2019 • University of Toronto
Friday, 9:00 a.m.–12:00 p.m., Sidney Smith Hall 1085

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Friday, September 13, 2019

1. Foundations and Wasserstein Geometry

The Monge and Kantorovich Problems

Let X and Y be Polish spaces with Borel probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. In applications these measures are often empirical distributions,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \qquad \text{and} \qquad \nu = \frac{1}{M}\sum_{j=1}^{M} \delta_{y_j}.$$

A **cost function** $c\colon X \times Y \to \mathbb{R}$ specifies the expense of moving one unit of mass from x to y. In this sense, the cost equips the probability measures with a geometry inherited from the underlying state spaces.

The original transport problem asks for a deterministic reassignment of mass.

Definition 1.1 The **Monge problem** is

$$\inf_{T\colon T_{\#}\mu = \nu} \int_X c(x, T(x))\, \mathrm{d}\mu(x),$$

where $T_{\#}\mu$ is the **pushforward measure** defined by

$$T_{\#}\mu(B) = \mu(T^{-1}(B))$$

for every Borel set $B \subseteq Y$.

The constraint in Definition 1.1 is nonlinear, and a feasible map need not exist. For example, a map cannot split a point mass between two destinations. Kantorovich's relaxation replaces maps by joint distributions.

Definition 1.2 A **coupling** of μ and ν is a measure $\gamma \in \mathcal{P}(X \times Y)$ whose marginals are

μ and ν. The set of all such couplings is denoted by

$$\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(X \times Y) : (\pi_X)_{\#}\gamma = \mu, (\pi_Y)_{\#}\gamma = \nu\}.$$

The **Monge–Kantorovich problem** is

$$\mathcal{T}_c(\mu,\nu) \coloneqq \inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\,\mathrm{d}\gamma(x,y). \tag{1.1}$$

Disintegration writes a coupling as

$$\mathrm{d}\gamma(x,y) = \mathrm{d}\mu(x)\,\mathrm{d}\gamma(y \mid x).$$

The conditional distribution $\gamma(\cdot \mid x)$ makes explicit the possibility of splitting the mass initially located at x. When $\gamma = (\mathrm{Id}, T)_{\#}\mu$, this conditional distribution is the point mass $\delta_{T(x)}$, and (1.1) reduces to the Monge problem.

The distinction is illustrated in Figure 1. A **transport map** selects one destination for each source, whereas a general coupling may send fractions of the same source mass to several destinations.

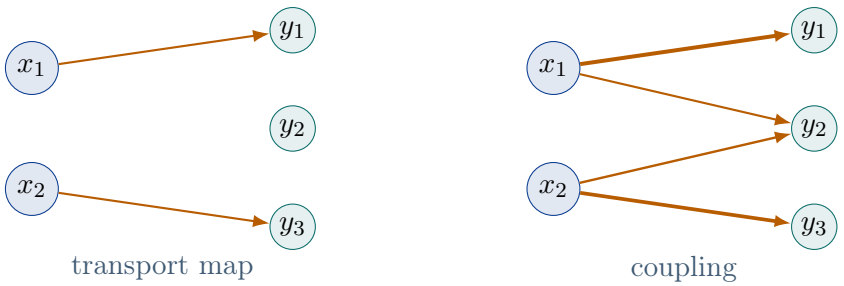

Figure 1

The set $\Pi(\mu,\nu)$ is convex: the marginal conditions are linear in γ. The objective in (1.1) is also linear. Thus the Kantorovich formulation is an infinite dimensional linear programming problem, and its duality theory will be central.

Important examples include

$$c(x,y) = |x-y|, \qquad c(x,y) = \frac{1}{2}|x-y|^2, \qquad \text{and} \qquad c(x,y) = \frac{1}{2}d(x,y)^2$$

on a metric space or a Riemannian manifold.

Existence of Optimal Couplings

Theorem 1.3 Suppose that $c\colon X \times Y \to \mathbb{R}$ is continuous and bounded below. Then, for every $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, the infimum in (1.1) is attained whenever its value is finite.

Proof. Choose a minimizing sequence $(\gamma_n)_{n\geq 1} \subseteq \Pi(\mu,\nu)$. The family $\Pi(\mu,\nu)$ is tight because its two marginals are fixed and tight. Prokhorov's theorem therefore gives a weakly convergent subsequence, which we continue to denote by (γ_n), with $\gamma_n \Rightarrow \gamma$.

The marginal maps are continuous under weak convergence. Hence, for every bounded continuous f on X and g on Y,

$$\int f(x)\,\mathrm{d}\gamma = \lim_{n\to\infty} \int f(x)\,\mathrm{d}\gamma_n = \int f\,\mathrm{d}\mu$$

and similarly $\int g(y)\,\mathrm{d}\gamma = \int g\,\mathrm{d}\nu$. Thus $\gamma \in \Pi(\mu,\nu)$.

After adding a constant, we may assume that $c \geq 0$. Set $c_k = c \wedge k$. Each c_k is bounded and continuous, so

$$\int c_k\,\mathrm{d}\gamma = \lim_{n\to\infty} \int c_k\,\mathrm{d}\gamma_n \leq \liminf_{n\to\infty} \int c\,\mathrm{d}\gamma_n.$$

Letting $k \to \infty$ and applying the monotone convergence theorem yields

$$\int c\,\mathrm{d}\gamma \leq \liminf_{n\to\infty} \int c\,\mathrm{d}\gamma_n = \mathcal{T}_c(\mu,\nu).$$

The reverse inequality follows from feasibility, so γ is optimal. □

Cyclical Monotonicity and c-Concavity

For the **quadratic cost** on $\mathbb{R}^n$, suppose that sending $\boldsymbol{x}_i$ to $\boldsymbol{y}_i$ is optimal. Any cyclic rearrangement of the destinations must be no cheaper, so

$$\sum_{i=1}^{N} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 \leq \sum_{i=1}^{N} \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|^2$$

for every cyclic permutation σ. Expanding the squares gives a monotonicity inequality. Conversely, if ϕ is convex and $\boldsymbol{y}_i \in \partial\phi(\boldsymbol{x}_i)$, then the subgradient inequalities telescope around every cycle and give precisely the same optimality condition. This motivates the following definition.

Definition 1.4 A set $\Gamma \subseteq X \times Y$ is ***c*-cyclically monotone** if, for every $N \geq 1$, every $(x_i, y_i)_{i=1}^N \subseteq \Gamma$, and every cyclic permutation σ of $\{1, \dots, N\}$,

$$\sum_{i=1}^N c(x_i, y_i) \leq \sum_{i=1}^N c(x_i, y_{\sigma(i)}). \tag{1.2}$$

Definition 1.5 For $\psi \colon Y \to \mathbb{R} \cup \{-\infty\}$, its ***c*-transform** is

$$\psi^c(x) = \inf_{y \in Y} \{c(x, y) - \psi(y)\}.$$

A function $\varphi \colon X \to \mathbb{R} \cup \{-\infty\}$ is ***c*-concave** if $\varphi = \psi^c$ for some ψ. Its ***c*-superdifferential** is

$$\partial^c \varphi = \{(x, y) \in X \times Y : \varphi(x) + \varphi^c(y) = c(x, y)\}.$$

When the fibre $\{y : (x, y) \in \partial^c \varphi\}$ is a singleton, its unique element is denoted by $\nabla^c \varphi(x)$.

The inequality $\varphi(x) + \varphi^c(y) \leq c(x, y)$ holds for all (x, y). Equality identifies the pairs that are compatible with the potential.

For the quadratic cost on $\mathbb{R}^n$ there is a direct link with ordinary convex analysis.

Proposition 1.6 Let $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2/2$. A function φ is c-concave if and only if

$$\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})$$

is lower semicontinuous and convex. Under this correspondence,

$$\partial^c \varphi = \partial \phi.$$

Proof. Expanding the cost gives

$$\varphi(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 - \psi(\boldsymbol{y}) \right\} = \frac{1}{2}\|\boldsymbol{x}\|^2 - \sup_{\boldsymbol{y}} \left\{ \boldsymbol{x} \cdot \boldsymbol{y} - \left(\frac{1}{2}\|\boldsymbol{y}\|^2 - \psi(\boldsymbol{y}) \right) \right\}.$$

Thus ϕ is a **convex conjugate**. Conversely, if ϕ is lower semicontinuous and convex, set

$$\psi(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2 - \phi^*(\boldsymbol{y}).$$

The identity $\phi = \phi^{**}$ then gives

$$\frac{1}{2}\|\boldsymbol{x}\|^2 - \phi(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \left\{ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 - \psi(\boldsymbol{y}) \right\},$$

which proves the converse. Moreover,

$$\varphi(\boldsymbol{x}) + \varphi^c(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$$

is equivalent to the **Fenchel equality**

$$\phi(\boldsymbol{x}) + \phi^*(\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y},$$

which holds if and only if $\boldsymbol{y} \in \partial\phi(\boldsymbol{x})$. □

The Fundamental Optimality Theorem

Theorem 1.7 Let $c\colon X \times Y \to \mathbb{R}$ be continuous and bounded below. Suppose that there exist $a \in L^1(\mu)$ and $b \in L^1(\nu)$ such that

$$c(x, y) \leq a(x) + b(y). \tag{1.3}$$

For $\gamma \in \Pi(\mu, \nu)$, the following statements are equivalent:

1. γ is optimal for the Monge–Kantorovich problem;
2. $\operatorname{supp}(\gamma)$ is c-cyclically monotone;
3. there exists a c-concave function φ such that $\max\{\varphi, 0\} \in L^1(\mu)$ and

$$\operatorname{supp}(\gamma) \subseteq \partial^c\varphi.$$

Proof. (1) *implies* (2). Suppose instead that $\operatorname{supp}(\gamma)$ is not c-cyclically monotone. Then there are $(x_i, y_i)_{i=1}^N \subseteq \operatorname{supp}(\gamma)$ and a cyclic permutation σ for which the inequality in (1.2) is strict in the opposite direction. Continuity and the Hausdorff property give neighbourhoods $U_i \times V_i$ on which the same strict improvement persists; they may be chosen disjoint for distinct pairs, while repeated occurrences use the same neighbourhood.

Let γ_i be the normalization of the restriction of γ to $U_i \times V_i$, and let μ_i and ν_i be its marginals.

Choose $\varepsilon > 0$ small enough that subtracting all occurrences of every restricted measure preserves nonnegativity, and define

$$\widetilde{\gamma} = \gamma - \varepsilon \sum_{i=1}^{N} \gamma_i + \varepsilon \sum_{i=1}^{N} \mu_i \otimes \nu_{\sigma(i)}.$$

The two signed perturbations have the same marginals, so $\widetilde{\gamma} \in \Pi(\mu, \nu)$. The strict local cost inequality makes $\widetilde{\gamma}$ cheaper than γ, contradicting optimality.

(2) *implies* (3). Set $\Gamma = \operatorname{supp}(\gamma)$ and fix $(\bar{x}, \bar{y}) \in \Gamma$. For $x \in X$, define

$$\varphi(x) = \inf \Bigg\{ c(x, y_N) - c(x_N, y_N) + \sum_{i=1}^{N-1} \big(c(x_{i+1}, y_i) - c(x_i, y_i) \big) : N \geq 1, (x_i, y_i) \in \Gamma, (x_1, y_1) = (\bar{x}, \bar{y}) \Bigg\}. \tag{1.4}$$

Cyclical monotonicity implies $\varphi(\bar{x}) = 0$, so the function is proper. Each expression inside the infimum has the form $c(x, y_N) - A$; hence φ is c-concave. Appending a point $(x, y) \in \Gamma$ to a chain in (1.4) yields

$$\varphi(x') \leq \varphi(x) + c(x', y) - c(x, y)$$

for every $x' \in X$. This is equivalent to $(x, y) \in \partial^c \varphi$, and therefore $\Gamma \subseteq \partial^c \varphi$. The majorant in (1.3), together with an additive normalization of φ, gives $\max\{\varphi, 0\} \in L^1(\mu)$.

(3) *implies* (1). For every coupling $\widetilde{\gamma} \in \Pi(\mu, \nu)$,

$$c(x, y) \geq \varphi(x) + \varphi^c(y).$$

Equality holds γ-almost everywhere because $\operatorname{supp}(\gamma) \subseteq \partial^c \varphi$. Consequently,

$$\begin{aligned} \int c \, \mathrm{d}\gamma &= \int \varphi \, \mathrm{d}\mu + \int \varphi^c \, \mathrm{d}\nu \\ &= \int \big(\varphi(x) + \varphi^c(y) \big) \, \mathrm{d}\widetilde{\gamma}(x, y) \leq \int c \, \mathrm{d}\widetilde{\gamma}. \end{aligned}$$

Thus γ is optimal. □

The theorem shows that optimality depends on the geometry of the support rather than on the particular amount of mass assigned to each supported pair.

Exercise 1.8 Let γ be optimal, and let $A \subseteq X \times Y$ satisfy $\gamma(A) > 0$. Show that the normalized restriction

$$\widetilde{\gamma} = \frac{1}{\gamma(A)} \gamma|_A$$

is optimal with respect to its own marginals.

Solution. Let $\widetilde{\mu} = (\pi_X)_{\#}\widetilde{\gamma}$ and $\widetilde{\nu} = (\pi_Y)_{\#}\widetilde{\gamma}$. For every Borel set $B \subseteq X$,

$$\widetilde{\mu}(B) = \frac{\gamma\big(A \cap (B \times Y)\big)}{\gamma(A)} \leq \frac{\mu(B)}{\gamma(A)},$$

and similarly $\widetilde{\nu}(C) \leq \nu(C)/\gamma(A)$ for every Borel set $C \subseteq Y$. Consequently, the functions in the integrable majorant (1.3) satisfy

$$\int_X |a| \, \mathrm{d}\widetilde{\mu} \leq \frac{1}{\gamma(A)} \int_X |a| \, \mathrm{d}\mu < \infty \qquad \text{and} \qquad \int_Y |b| \, \mathrm{d}\widetilde{\nu} \leq \frac{1}{\gamma(A)} \int_Y |b| \, \mathrm{d}\nu < \infty.$$

The implication from (1) to (2) in Theorem 1.7 shows that $\mathrm{supp}(\gamma)$ is c-cyclically monotone. Since

$$\mathrm{supp}(\widetilde{\gamma}) \subseteq \mathrm{supp}(\gamma),$$

the support of $\widetilde{\gamma}$ is also c-cyclically monotone. The implication from (2) to (1) in Theorem 1.7, applied to the marginals $\widetilde{\mu}$ and $\widetilde{\nu}$, now shows that $\widetilde{\gamma}$ is optimal. □

Exercise 1.9 Let $T \colon X \to Y$ be measurable and have a c-cyclically monotone graph. Given $\mu \in \mathcal{P}(X)$, set $\nu = T_{\#}\mu$. Show, under the integrability condition of Theorem 1.7, that T is an optimal transport map from μ to ν.

Solution. Define

$$\gamma = (\mathrm{Id}, T)_{\#}\mu.$$

Its marginals are μ and $T_{\#}\mu = \nu$, so $\gamma \in \Pi(\mu, \nu)$. Let G be the graph of T. Since the map $x \mapsto (x, T(x))$ takes values in G, $\gamma(X \times Y \setminus \overline{G}) = 0$, and hence

$$\mathrm{supp}(\gamma) \subseteq \overline{G}.$$

Moreover, $\overline{G}$ remains c-cyclically monotone. Indeed, for any finite collection $(x_i, y_i)_{i=1}^N \subseteq \overline{G}$, choose $(x_i^{(k)}, y_i^{(k)}) \in G$ converging to (x_i, y_i) for each i. The cyclical monotonicity inequality holds for the

points $(x_i^{(k)}, y_i^{(k)})$, and continuity of c allows us to pass to the limit. Thus $\operatorname{supp}(\gamma)$ is c-cyclically monotone.

Under the integrability hypothesis, the implication from (2) to (1) in Theorem 1.7 shows that γ is an optimal coupling. Since γ is induced by T, this says precisely that T is an optimal transport map from μ to ν. □

Kantorovich Duality

Theorem 1.10 (Kantorovich duality) Under the hypotheses of Theorem 1.7,

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int c \, \mathrm{d}\gamma = \sup_{\substack{\varphi \in L^1(\mu),\, \psi \in L^1(\nu) \\ \varphi(x)+\psi(y) \le c(x,y)}} \left\{ \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu \right\}. \tag{1.5}$$

The supremum is attained by a pair (φ, φ^c) in which φ is c-concave. Such a pair is called a pair of **Kantorovich potentials**.

Proof. For every feasible dual pair and every coupling γ,

$$\int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu = \int (\varphi(x) + \psi(y)) \, \mathrm{d}\gamma(x,y) \le \int c \, \mathrm{d}\gamma.$$

This is **weak duality**. By Theorem 1.3, there is an optimal coupling γ^*. The implication from (1) to (3) in Theorem 1.7 gives a c-concave φ with $\operatorname{supp}(\gamma^*) \subseteq \partial^c \varphi$. Equality therefore holds in the pointwise dual constraint γ^*-almost everywhere, and

$$\int c \, \mathrm{d}\gamma^* = \int \varphi \, \mathrm{d}\mu + \int \varphi^c \, \mathrm{d}\nu.$$

This proves equality of the primal and dual values and attainment of the dual value. □

The dual variables have the economic interpretation of prices at the source and destination. **Complementary slackness** is the equality

$$\varphi(x) + \varphi^c(y) = c(x,y)$$

on the support of an optimal coupling.

Lecture 2 — Friday, September 20, 2019

Quadratic Transport and Brenier's Theorem

The quadratic cost

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad X = Y = \mathbb{R}^n,$$

turns the convex analytic correspondence in Proposition 1.6 into an explicit transport map.

Theorem 2.1 (Brenier's theorem) Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^n)$, and suppose that μ is absolutely continuous with respect to Lebesgue measure. There is a lower semicontinuous convex function $\phi\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that

$$T(\boldsymbol{x}) = \nabla\phi(\boldsymbol{x})$$

is an optimal transport map from μ to ν. The optimal transport map is unique μ-almost everywhere.

Proof. We prove existence. Let γ be an optimal coupling. By Theorem 1.7, there is a c-concave function φ such that

$$\Gamma := \operatorname{supp}(\gamma) \subseteq \partial^c\varphi.$$

By Proposition 1.6,

$$\phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \varphi(\boldsymbol{x})$$

is lower semicontinuous and convex, and $\Gamma \subseteq \partial\phi$. The fibre $\partial\phi(\boldsymbol{x})$ is nonempty for μ-almost every $\boldsymbol{x}$, so these points lie in $\operatorname{Dom}(\phi)$. A lower semicontinuous convex function is locally Lipschitz, and hence differentiable almost everywhere, in the interior of its effective domain. Rademacher's theorem and the absolute continuity of μ therefore imply that $\nabla\phi$ is defined μ-almost everywhere. At every differentiability point,

$$\partial\phi(\boldsymbol{x}) = \{\nabla\phi(\boldsymbol{x})\}.$$

Thus γ is concentrated on the graph of $\nabla\phi$, proving that $T = \nabla\phi$ is an optimal map.

It remains to justify the two details used above. The boundary of the effective domain of a convex function has Lebesgue measure zero unless that domain is contained in a lower dimensional affine subspace; the latter alternative is impossible here because $\partial\phi(\boldsymbol{x})$ is nonempty for μ-almost every

$\boldsymbol{x}$ and μ is absolutely continuous. Thus μ assigns full mass to points in the interior of the effective domain at which ϕ is differentiable.

For uniqueness, let γ_1 and γ_2 be optimal couplings. Their average is also optimal. Applying the preceding argument to $(\gamma_1 + \gamma_2)/2$ shows that this average is concentrated on the graph of a single gradient. Hence both γ_1 and γ_2 are concentrated on that same graph and must agree. In particular, every optimal coupling is induced by the unique map $T = \nabla\phi$, up to a μ-null set. □

In one dimension, the gradient of a convex function is nondecreasing. Consequently, Theorem 2.1 reduces to the **monotone coupling**.

Theorem 2.2 Let $c(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$ on $\mathbb{R}^n$, where h is strictly convex and satisfies the standard coercivity and regularity conditions. If μ is absolutely continuous, then there is a unique optimal transport map,

$$T(\boldsymbol{x}) = \boldsymbol{x} - \nabla h^*(\nabla\varphi(\boldsymbol{x})),$$

where φ is a c-concave potential.

The same strategy extends the cost $d^2/2$ to Riemannian manifolds. These extensions are developed in Gangbo and McCann (1996) and McCann (2001).

The regularity step rests on a useful stability property of **moduli of continuity**.

Lemma 2.3 Suppose that every member of a family $(f_\alpha)_\alpha$ satisfies

$$\left|f_\alpha(x) - f_\alpha(x')\right| \le \omega(d(x, x'))$$

for the same modulus of continuity ω. Then $f = \inf_\alpha f_\alpha$ satisfies the same estimate.

Proof. For every $\varepsilon > 0$, choose α with $f_\alpha(x') \le f(x') + \varepsilon$. Then

$$f(x) - f(x') \le f_\alpha(x) - f_\alpha(x') + \varepsilon \le \omega(d(x, x')) + \varepsilon.$$

Let $\varepsilon \downarrow 0$ and interchange x and x'. □

If smooth positive densities f and g and a smooth orientation-preserving diffeomorphism T satisfy

$T_{\#}(f\,\mathrm{d}\boldsymbol{x}) = g\,\mathrm{d}\boldsymbol{y}$, then the change of variables formula gives

$$f(\boldsymbol{x}) = g(T(\boldsymbol{x}))|\det DT(\boldsymbol{x})|.$$

With $T = \nabla\phi$, this becomes the **Monge–Ampère equation**

$$f(\boldsymbol{x}) = g(\nabla\phi(\boldsymbol{x}))\det D^2\phi(\boldsymbol{x}). \tag{2.1}$$

This identity holds at points where the classical change of variables formula applies. Without this smoothness, the equation must be interpreted in a weak sense. Its regularity theory is nonlinear and delicate; for general costs, a central role is played by the **Ma–Trudinger–Wang condition**.

The Wasserstein Metric

Definition 2.4 Let (X, d) be a Polish metric space and let $p \geq 1$. The ***p*-Wasserstein space** is

$$\mathcal{P}_p(X) = \left\{\mu \in \mathcal{P}(X) : \int_X d(x_0, x)^p\,\mathrm{d}\mu(x) < \infty \text{ for some } x_0 \in X\right\}.$$

For $\mu, \nu \in \mathcal{P}_p(X)$, the ***p*-Wasserstein distance** is

$$W_p(\mu,\nu) = \left(\inf_{\gamma\in\Pi(\mu,\nu)}\int_{X\times X} d(x,y)^p\,\mathrm{d}\gamma(x,y)\right)^{1/p}. \tag{2.2}$$

For point masses, $W_p(\delta_x, \delta_y) = d(x, y)$. Hence $x \mapsto \delta_x$ is an isometric embedding of (X, d) into $(\mathcal{P}_p(X), W_p)$.

Theorem 2.5 The function W_p is a metric on $\mathcal{P}_p(X)$.

Proof. Only the triangle inequality requires a new argument. Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(X)$, and choose optimal couplings $\gamma_{12} \in \Pi(\mu_1, \mu_2)$ and $\gamma_{23} \in \Pi(\mu_2, \mu_3)$. Disintegrate them over their common marginal μ_2 and define

$$\mathrm{d}\gamma_{123}(x_1, x_2, x_3) = \mathrm{d}\mu_2(x_2)\ \mathrm{d}\gamma_{12}(x_1 \mid x_2)\ \mathrm{d}\gamma_{23}(x_3 \mid x_2).$$

This **gluing construction** has $(1, 2)$-marginal γ_{12} and $(2, 3)$-marginal γ_{23}. Its $(1, 3)$-marginal is a feasible coupling of μ_1 and μ_3. The triangle inequality for d followed by Minkowski's inequality

gives

$$\begin{aligned} W_p(\mu_1, \mu_3) &\le \left(\int d(x_1, x_3)^p \, \mathrm{d}\gamma_{123} \right)^{1/p} \\ &\le \left(\int d(x_1, x_2)^p \, \mathrm{d}\gamma_{123} \right)^{1/p} + \left(\int d(x_2, x_3)^p \, \mathrm{d}\gamma_{123} \right)^{1/p} \\ &= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3). \end{aligned}$$

□

For Gaussian measures the quadratic Wasserstein metric separates location and dispersion.

Proposition 2.6 Let $\mu = \mathcal{N}(\boldsymbol{m}_\alpha, \boldsymbol{\Sigma}_\alpha)$ and $\nu = \mathcal{N}(\boldsymbol{m}_\beta, \boldsymbol{\Sigma}_\beta)$. Then

$$W_2(\mu, \nu)^2 = \|\boldsymbol{m}_\alpha - \boldsymbol{m}_\beta\|^2 + B(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2,$$

where the **Bures–Wasserstein distance** is

$$B(\boldsymbol{\Sigma}_\alpha, \boldsymbol{\Sigma}_\beta)^2 = \operatorname{tr}\left(\boldsymbol{\Sigma}_\alpha + \boldsymbol{\Sigma}_\beta - 2(\boldsymbol{\Sigma}_\alpha^{1/2} \boldsymbol{\Sigma}_\beta \boldsymbol{\Sigma}_\alpha^{1/2})^{1/2} \right).$$

Dirichlet Transport and Divergences

The open simplex

$$\Delta_n = \left\{ \boldsymbol{p} = (p_1, \dots, p_n) \in (0,1)^n : \sum_{i=1}^n p_i = 1 \right\}$$

supports a multiplicative analogue of quadratic transport. Its cost is

$$c(\boldsymbol{p}, \boldsymbol{q}) = \log\left(\frac{1}{n} \sum_{i=1}^n \frac{q_i}{p_i} \right) - \frac{1}{n} \sum_{i=1}^n \log\left(\frac{q_i}{p_i} \right). \tag{2.3}$$

The expression is nonnegative by the arithmetic–geometric mean inequality. It is a **divergence** rather than a metric: symmetry and the triangle inequality need not hold.

One motivation comes from stochastic portfolio theory. If $m_i(t)$ is the market capitalization of stock i divided by the total market capitalization, then $\boldsymbol{m}(t) \in \Delta_n$. A positive concave function

Φ measures market diversity. Typical examples are

$$\Phi(\boldsymbol{p}) = \left(\sum_{i=1}^{n} p_i^{\lambda} \right)^{1/\lambda}, \quad 0 < \lambda < 1 \qquad \text{and} \qquad \Phi(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \log p_i.$$

The first is a **diversity index**; the second is **Shannon entropy**. Stability of the capital distribution can therefore be phrased as persistence of diversity along the market path.

A positive differentiable concave function Φ generates a portfolio map $\boldsymbol{\pi} \colon \Delta_n \to \overline{\Delta}_n$ by

$$\pi_i(\boldsymbol{p}) = p_i \left(1 + D_{\boldsymbol{e}_i - \boldsymbol{p}} \log \Phi(\boldsymbol{p})\right). \tag{2.4}$$

Geometrically, the supporting hyperplane to Φ at $\boldsymbol{p}$ determines the portfolio weights, as shown in Figure 2.

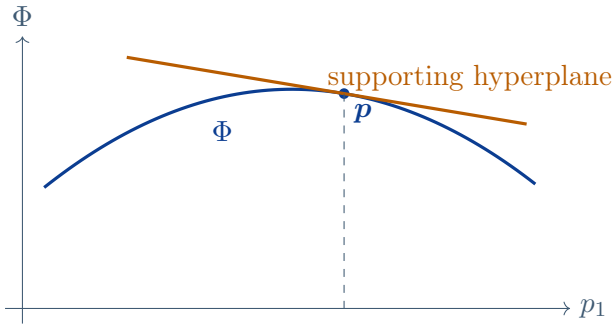

Figure 2

Two useful examples follow directly from (2.4). If $\boldsymbol{\pi} \in \overline{\Delta}_n$ and

$$\Phi(\boldsymbol{p}) = \prod_{i=1}^{n} p_i^{\pi_i},$$

then the generated portfolio is the constant weight vector $\boldsymbol{\pi}$. If

$$\Phi(\boldsymbol{p}) = \left(\sum_{i=1}^{n} p_i^{\lambda} \right)^{1/\lambda},$$

then $\pi_i(\boldsymbol{p})$ is proportional to p_i^{λ}.

Theorem 2.7 For a portfolio map $\boldsymbol{\pi} \colon \Delta_n \to \overline{\Delta}_n$, the following statements are equivalent:

1. The portfolio is functionally generated.
2. For every cycle $\boldsymbol{p}(0), \boldsymbol{p}(1), \ldots, \boldsymbol{p}(T) = \boldsymbol{p}(0)$,

$$\prod_{t=0}^{T-1} \left(\sum_{i=1}^{n} \pi_i(\boldsymbol{p}(t)) \frac{p_i(t+1)}{p_i(t)} \right) \geq 1.$$

Condition (2) says that the portfolio does not underperform the market over a closed path. After a change of coordinates, it is the c-cyclical monotonicity condition for the cost in (2.3).

The relevant multiplicative linear structure on Δ_n is defined by

$$(\boldsymbol{p} \oplus \boldsymbol{q})_i = \frac{p_i q_i}{\sum_{j=1}^{n} p_j q_j} \qquad \text{and} \qquad (\lambda \odot \boldsymbol{p})_i = \frac{p_i^{\lambda}}{\sum_{j=1}^{n} p_j^{\lambda}}.$$

The neutral element is $\boldsymbol{e} = (1/n, \ldots, 1/n)$, and the inverse $\boldsymbol{p}^{-1}$ is obtained by normalizing the reciprocals. This is **Aitchison's geometry** for compositional data.

Lemma 2.8 The cost in (2.3) satisfies

$$c(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{KL}(\boldsymbol{e} \mid \boldsymbol{q} \oplus \boldsymbol{p}^{-1}),$$

where the **relative entropy** is

$$\mathrm{KL}(\boldsymbol{r} \mid \boldsymbol{s}) = \sum_{i=1}^{n} r_i \log \left(\frac{r_i}{s_i} \right).$$

Thus the **Dirichlet cost** plays the role that squared Euclidean distance plays in quadratic transport.

Theorem 2.9 Suppose that $\mu \in \mathcal{P}(\Delta_n)$ is absolutely continuous and that μ and ν satisfy the required moment conditions. There is a concave function $\Phi \colon \Delta_n \to (0, \infty)$ for which the deterministic optimal transport map has components

$$q_i = \frac{p_i \pi_i(\boldsymbol{p}^{-1})}{\sum_{j=1}^{n} p_j \pi_j(\boldsymbol{p}^{-1})}.$$

Equivalently, $\boldsymbol{q} = \boldsymbol{p} \oplus \boldsymbol{\pi}(\boldsymbol{p}^{-1})$.

The increment $\boldsymbol{q} \oplus \boldsymbol{p}^{-1}$ is the multiplicative counterpart of the displacement $\boldsymbol{y} - \boldsymbol{x}$.

Definition 2.10 Let $\Omega \subseteq \mathbb{R}^n$ be convex, and let $\phi \colon \Omega \to \mathbb{R}$ be differentiable and strictly convex. The **Bregman divergence** generated by ϕ is

$$D_\phi[\boldsymbol{x} : \boldsymbol{x}'] = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \nabla\phi(\boldsymbol{x}') \cdot (\boldsymbol{x} - \boldsymbol{x}').$$

Convexity gives $D_\phi[\boldsymbol{x} : \boldsymbol{x}'] \geq 0$, with equality only when $\boldsymbol{x} = \boldsymbol{x}'$. The divergence need not be symmetric. If ϕ is twice differentiable at $\boldsymbol{x}$, then locally,

$$D_\phi[\boldsymbol{x} + \Delta\boldsymbol{x} : \boldsymbol{x}] = \frac{1}{2}\Delta\boldsymbol{x}^{\mathsf{T}} D^2\phi(\boldsymbol{x})\Delta\boldsymbol{x} + o(\|\Delta\boldsymbol{x}\|^2),$$

so the Hessian of ϕ determines a Riemannian metric. The quadratic generator $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2/2$ gives squared Euclidean distance, while negative Shannon entropy gives relative entropy.

Definition 2.11 Let T be an optimal transport map with graph $G = \{(\boldsymbol{p}, T(\boldsymbol{p}))\}$, and let (φ, ψ) be Kantorovich potentials satisfying

$$\varphi(\boldsymbol{p}) + \psi(\boldsymbol{q}') \leq c(\boldsymbol{p}, \boldsymbol{q}'),$$

with equality precisely when $\boldsymbol{q}' = T(\boldsymbol{p})$. For $x = (\boldsymbol{p}, \boldsymbol{q})$ and $x' = (\boldsymbol{p}', \boldsymbol{q}')$ in G, define the ***c*-divergence**

$$D[x : x'] = c(\boldsymbol{p}, \boldsymbol{q}') - \varphi(\boldsymbol{p}) - \psi(\boldsymbol{q}') \geq 0.$$

The crossed pair $(\boldsymbol{p}, \boldsymbol{q}')$ generally lies off the optimal graph, as illustrated in Figure 3; the c-divergence measures its **dual slack**.

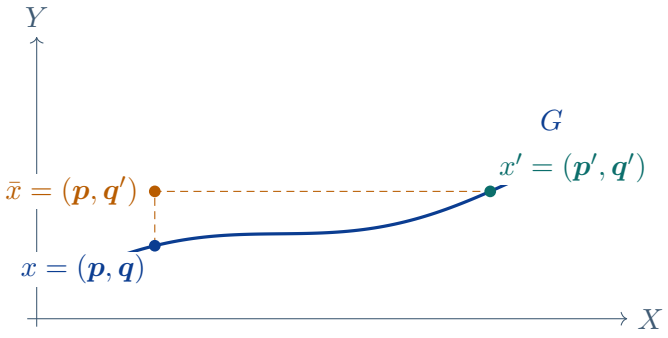

Figure 3

Theorem 2.12 For $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2/2$, the c-divergence is a Bregman divergence.

Proof. Apply the convex correspondence in Proposition 1.6. The equality set $\partial^c\varphi$ becomes the ordinary subdifferential of the convex function $\phi = \|\cdot\|^2/2 - \varphi$, and the dual slack becomes

$$\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \nabla\phi(\boldsymbol{x}') \cdot (\boldsymbol{x} - \boldsymbol{x}').$$

□

Definition 2.13 Let $\alpha > 0$. A differentiable function $\varphi\colon \Omega \to \mathbb{R}$ is **α-exponentially concave** if $\Phi = \mathrm{e}^{\alpha\varphi}$ is concave. Its **$L^{(\alpha)}$-divergence** is

$$D^{(\alpha)}[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \frac{1}{\alpha} \log\big(1 + \alpha\nabla\varphi(\boldsymbol{\xi}') \cdot (\boldsymbol{\xi} - \boldsymbol{\xi}')\big) - \big(\varphi(\boldsymbol{\xi}) - \varphi(\boldsymbol{\xi}')\big).$$

The first term uses a logarithmic approximation to the graph of φ, while a Bregman divergence uses a linear approximation. As $\alpha \downarrow 0$, the logarithmic approximation becomes linear and $D^{(\alpha)}$ converges to the corresponding Bregman divergence. The Dirichlet transport is the case $\Omega = \Delta_n$ and $\alpha = 1$.

Wasserstein Convergence and Displacement Interpolation

Theorem 2.14 Let $(\mu_k)_{k\geq 1} \subseteq \mathcal{P}_p(X)$ and $\mu \in \mathcal{P}_p(X)$. The following statements are equivalent:

1. $W_p(\mu_k, \mu) \to 0$;
2. $\mu_k \Rightarrow \mu$ and, for some, hence every, $x_0 \in X$,

$$\int_X d(x_0, x)^p \,\mathrm{d}\mu_k(x) \longrightarrow \int_X d(x_0, x)^p \,\mathrm{d}\mu(x).$$

Proof. We prove the result when X is compact; the general case replaces boundedness by uniform integrability.

Suppose first that (2) holds. The moment condition is automatic on a compact space. By the

Skorokhod representation theorem, there are random elements $X_k \sim \mu_k$ and $X \sim \mu$ on a common probability space with $X_k \to X$ almost surely. Their joint laws are couplings, and the bounded convergence theorem gives

$$W_p(\mu_k, \mu)^p \le \mathbb{E}[d(X_k, X)^p] \longrightarrow 0.$$

Conversely, suppose that (1) holds and let (X_k, X) be an optimal coupling of (μ_k, μ). For $\delta > 0$, Markov's inequality gives

$$\mathbb{P}(d(X_k, X) \ge \delta) \le \frac{W_p(\mu_k, \mu)^p}{\delta^p}.$$

If f is continuous, compactness makes it bounded by some M and uniformly continuous. Given $\varepsilon > 0$, choose $\delta > 0$ such that $d(x, x') < \delta$ implies $|f(x) - f(x')| < \varepsilon$. Then

$$\begin{aligned}\left| \int f \, \mathrm{d}\mu_k - \int f \, \mathrm{d}\mu \right| &\le \mathbb{E}[|f(X_k) - f(X)|] \\ &\le \varepsilon + 2M \frac{W_p(\mu_k, \mu)^p}{\delta^p}.\end{aligned}$$

Letting $k \to \infty$ and then $\varepsilon \downarrow 0$ proves weak convergence. The moment convergence again follows automatically from compactness. □

Definition 2.15 Let $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^n)$, with μ_0 absolutely continuous, and let $\nabla\phi$ be the **Brenier map** satisfying $\mu_1 = (\nabla\phi)_{\#}\mu_0$. Define

$$\phi_t(\boldsymbol{x}) = (1-t)\frac{1}{2}\|\boldsymbol{x}\|^2 + t\phi(\boldsymbol{x})$$

and the **displacement interpolation**

$$[\mu_0, \mu_1]_t = (\nabla\phi_t)_{\#}\mu_0 = \big((1-t)\,\mathrm{Id} + t\nabla\phi\big)_{\#}\mu_0. \tag{2.5}$$

Ordinary mixture interpolation fades one density out while fading the other in. Displacement interpolation instead moves individual particles along straight lines. Figure 4 shows how two separated masses travel rather than coexist at intermediate times.

Figure 4

Proposition 2.16 Let $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^n)$. Then the following hold:

1. $[\mu,\nu]_t \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^n)$ for $0 < t < 1$;
2. $[\mu,\nu]_t = [\nu,\mu]_{1-t}$;
3. $[[\mu,\nu]_t, [\mu,\nu]_{t'}]_s = [\mu,\nu]_{(1-s)t+st'}$;
4. For μ-almost every $\boldsymbol{x}$, the path $t \mapsto (1-t)\boldsymbol{x} + t\nabla\phi(\boldsymbol{x})$ is a constant velocity straight line.

Theorem 2.17 The displacement interpolation is a **constant speed geodesic**:

$$W_2([\mu,\nu]_s, [\mu,\nu]_t) = |s-t| W_2(\mu,\nu).$$

Proof. Write $\mu_t = [\mu_0,\mu_1]_t$. Coupling μ_s and μ_t through the same starting point gives

$$\begin{aligned} W_2(\mu_s,\mu_t)^2 &\le \int \|\nabla\phi_s(\boldsymbol{x}) - \nabla\phi_t(\boldsymbol{x})\|^2 \,\mathrm{d}\mu_0(\boldsymbol{x}) \\ &= (s-t)^2 \int \|\nabla\phi(\boldsymbol{x}) - \boldsymbol{x}\|^2 \,\mathrm{d}\mu_0(\boldsymbol{x}) \\ &= (s-t)^2 W_2(\mu_0,\mu_1)^2. \end{aligned}$$

Applying this estimate to the segments from 0 to s, from s to t, and from t to 1, the triangle inequality forces equality. □

Definition 2.18 A set $A \subseteq \mathcal{P}_2(\mathbb{R}^n)$ is **displacement convex** if it contains $[\mu,\nu]_t$ whenever it contains μ and ν and the Brenier interpolation in (2.5) is defined. A functional $F\colon A \to \mathbb{R}$ is **displacement convex** if

$$t \longmapsto F([\mu,\nu]_t)$$

is convex for every such pair $\mu, \nu \in A$.

This convexity was introduced to study interacting gases. If ρ is a density, consider

$$E(\rho) = \int A(\rho(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\iint V(\boldsymbol{x}-\boldsymbol{y})\rho(\boldsymbol{x})\rho(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.$$

The terms represent **internal compression energy** and **interaction energy**. Although E need not be convex under ordinary mixtures of densities, it is displacement convex under natural

physical conditions on A and V.

The same geodesic picture survives beyond Euclidean space.

Theorem 2.19 If (X, d) is Polish and geodesic, then $(\mathcal{P}_2(X), W_2)$ is geodesic. A curve $t \mapsto \mu_t$ is a constant speed geodesic if and only if there is a measure $\mathbb{P} \in \mathcal{P}_2(\mathrm{Geod}(X))$ whose marginal at time t is μ_t and whose marginal at every pair of times is an optimal coupling.

The measure $\mathbb{P}$ describes a random particle whose path is itself a geodesic in X.

Dynamic Optimal Transport

In the static problem, a pair (x, y) records only the endpoints. A **dynamic cost** assigns an **action** $C[(z_t)_{0\le t\le 1}]$ to a path, with compatibility condition

$$c(x, y) = \inf \left\{ C[(z_t)_{0\le t\le 1}] : z_0 = x,\ z_1 = y \right\}.$$

For the quadratic cost, the basic identity is

$$\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \inf_{\substack{\boldsymbol{x}_0 = \boldsymbol{x} \\ \boldsymbol{x}_1 = \boldsymbol{y} \\ \boldsymbol{x}_\cdot \text{ absolutely continuous}}} \int_0^1 \frac{1}{2}\|\dot{\boldsymbol{x}}_t\|^2 \, \mathrm{d}t. \tag{2.6}$$

The unique minimizing path is the constant velocity line segment.

In the **Lagrangian description**, one follows individual particles. If $\mathcal{D}_1(\mu, \nu)$ is the set of laws on path space with endpoint marginals μ and ν, then

$$\frac{1}{2} W_2(\mu, \nu)^2 = \inf_{\mathbb{P} \in \mathcal{D}_1(\mu,\nu)} \mathbb{E}_{\mathbb{P}} \left[\int_0^1 \frac{1}{2} \left\| \dot{\boldsymbol{X}}_t \right\|^2 \mathrm{d}t \right]. \tag{2.7}$$

An optimizer is obtained by first drawing an optimally coupled endpoint pair and then joining it by the line in (2.6).

The **Eulerian description** instead follows the collective density. Let $\boldsymbol{v}(t, \boldsymbol{x})$ be a **velocity field** and let μ_t be the distribution of particles solving

$$\dot{\boldsymbol{x}}_t = \boldsymbol{v}(t, \boldsymbol{x}_t).$$

Lemma 2.20 Let $(\boldsymbol{X}_t)_{0\le t\le 1}$ be an absolutely continuous random path satisfying

$$\dot{\boldsymbol{X}}_t = \boldsymbol{v}(t, \boldsymbol{X}_t)$$

for almost every t, and assume that $\mathbb{E}\int_0^1 \|\boldsymbol{v}(t,\boldsymbol{X}_t)\|\,\mathrm{d}t < \infty$. Then its marginal laws $\mu_t = \mathcal{L}(\boldsymbol{X}_t)$ satisfy the **continuity equation**

$$\partial_t \mu_t + \nabla\cdot(\mu_t \boldsymbol{v}) = 0 \tag{2.8}$$

in the weak sense.

Proof. For every smooth ζ with compact spatial support on $[0,1]\times\mathbb{R}^n$, the chain rule along almost every path gives

$$\zeta(1,\boldsymbol{X}_1) - \zeta(0,\boldsymbol{X}_0) = \int_0^1 \{\partial_t\zeta(t,\boldsymbol{X}_t) + \nabla\zeta(t,\boldsymbol{X}_t)\cdot\boldsymbol{v}(t,\boldsymbol{X}_t)\}\,\mathrm{d}t.$$

The integrability hypothesis permits expectation and time integration to be interchanged. Rewriting the expectations using the marginal laws yields

$$\int_0^1\int_{\mathbb{R}^n} (\partial_t\zeta + \nabla\zeta\cdot\boldsymbol{v})\,\mathrm{d}\mu_t\,\mathrm{d}t = \int_{\mathbb{R}^n}\zeta(1,\boldsymbol{x})\,\mathrm{d}\mu_1(\boldsymbol{x}) - \int_{\mathbb{R}^n}\zeta(0,\boldsymbol{x})\,\mathrm{d}\mu_0(\boldsymbol{x}),$$

which is the weak form of (2.8). □

Theorem 2.21 (Benamou–Brenier formula) For $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^n)$,

$$\frac{1}{2}W_2(\mu,\nu)^2 = \inf_{(\rho,\boldsymbol{v})}\int_0^1\int_{\mathbb{R}^n}\frac{1}{2}\|\boldsymbol{v}(t,\boldsymbol{x})\|^2\,\mathrm{d}\rho_t(\boldsymbol{x})\,\mathrm{d}t, \tag{2.9}$$

where the infimum is over solutions of (2.8) with $\rho_0=\mu$ and $\rho_1=\nu$.

Here the admissible curves are narrowly continuous, the velocity field is measurable, and the displayed action is finite; the continuity equation is interpreted distributionally.

The Lagrangian and Eulerian pictures are two views of the same motion: one resolves particle trajectories, while the other records only the density and velocity field.

Empirical Wasserstein Distances

Given independent samples $X_1, \dots, X_N \sim \mu$ and $Y_1, \dots, Y_N \sim \nu$, define the **empirical measures**

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i} \qquad \text{and} \qquad \nu_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{Y_i}.$$

On bounded d-dimensional domains in the high dimensional regime $d > 4$, a characteristic estimate is

$$\mathbb{E}\left[|W_2(\mu_N, \nu_N) - W_2(\mu, \nu)|\right] = O(N^{-1/d})$$

reveals the **curse of dimensionality**. The critical dimension $d = 4$ introduces a logarithmic correction, while below it the dimension-free sampling term dominates the generic moment bound; additional regularity can improve these rates.

More refined bounds use the moment

$$M_q(\mu) = \int_{\mathbb{R}^d} \|\boldsymbol{x}\|^q \, \mathrm{d}\mu(\boldsymbol{x}).$$

Theorem 2.22 Let $\mu \in \mathcal{P}(\mathbb{R}^d)$, let $p \geq 1$, and suppose that $M_q(\mu) < \infty$ for some $q > p$. There is a constant C, depending only on p, d, and q, such that

$$\mathbb{E}\big[W_p(\mu_N, \mu)^p\big] \leq C M_q(\mu)^{p/q} R_{p,d,q}(N),$$

where

$$R_{p,d,q}(N) = \begin{cases} N^{-1/2} + N^{-(q-p)/q}, & p > d/2 \text{ and } q \neq 2p, \\ N^{-1/2}\log(1+N) + N^{-(q-p)/q}, & p = d/2 \text{ and } q \neq 2p, \\ N^{-p/d} + N^{-(q-p)/q}, & 1 \leq p < d/2 \text{ and } q \neq dp/(d-p). \end{cases}$$

This result is due to Fournier and Guillin (2015). It separates the **geometric rate**, controlled by the dimension, from the **tail rate**, controlled by q.

Lecture 3 — Friday, September 27, 2019

2. Discrete Algorithms and Entropic Regularization

Discrete Transport as a Linear Program

Suppose that μ and ν are supported on finite sets and are represented by probability vectors

$$\boldsymbol{a} = (a_i)_{i=1}^n \qquad \text{and} \qquad \boldsymbol{b} = (b_j)_{j=1}^m.$$

The cost is represented by a matrix $\mathbf{C} = (C_{ij}) \in \mathbb{R}^{n \times m}$. A coupling is now a nonnegative matrix $\mathbf{P} = (P_{ij})$ in the **transport polytope**

$$\mathcal{U}(\boldsymbol{a}, \boldsymbol{b}) = \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P}\mathbf{1}_m = \boldsymbol{a},\ \mathbf{P}^\mathsf{T}\mathbf{1}_n = \boldsymbol{b} \right\}. \tag{3.1}$$

Thus the Monge–Kantorovich problem becomes the linear program

$$L_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\mathbf{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle = \min_{\mathbf{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij}. \tag{3.2}$$

In standard linear programming notation, a primal problem

$$\max_{\boldsymbol{x}} \ \boldsymbol{c}^\mathsf{T}\boldsymbol{x} \quad \text{subject to} \quad \mathbf{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \mathbf{0}$$

has the dual problem

$$\min_{\boldsymbol{y}} \ \boldsymbol{b}^\mathsf{T}\boldsymbol{y} \quad \text{subject to} \quad \mathbf{A}^\mathsf{T}\boldsymbol{y} \geq \boldsymbol{c}.$$

Applied to (3.2), these constraints reproduce the Kantorovich dual in Theorem 1.10.

The variables also form the bipartite network in Figure 5. A positive entry P_{ij} is a flow from source vertex i to target vertex j', and its width records the transported mass.

Proposition 3.1 If $\mathbf{P}$ is an **extreme point** of $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$, then $\mathbf{P}$ has at most $n + m - 1$ nonzero entries.

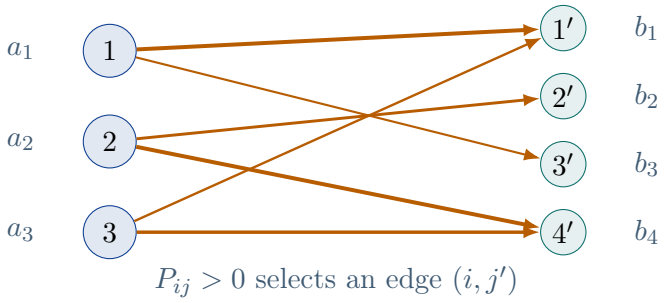

Figure 5

Proof. Consider the **support subgraph**

$$\widetilde{E} = \{(i, j') : P_{ij} > 0\}.$$

We claim that this graph has no cycle. Suppose that it contains the alternating cycle

$$(i_1, j_1'), (i_2, j_1'), (i_2, j_2'), \ldots, (i_k, j_k'), (i_1, j_k').$$

Construct a matrix $\mathbf{E}$ whose entries alternate between ε and $-\varepsilon$ around this cycle and vanish elsewhere, where

$$0 < \varepsilon < \min\{P_{ij} : (i, j') \in \widetilde{E}\}.$$

Every row sum and column sum of $\mathbf{E}$ is zero, so both $\mathbf{P} + \mathbf{E}$ and $\mathbf{P} - \mathbf{E}$ belong to $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$. Since

$$\mathbf{P} = \frac{1}{2}(\mathbf{P} + \mathbf{E}) + \frac{1}{2}(\mathbf{P} - \mathbf{E}),$$

this contradicts extremality. A forest on $n + m$ vertices has at most $n + m - 1$ edges, proving the result. □

In particular, when $n = m$, an extreme optimal coupling has order n nonzero entries rather than order n^2. For convex costs on a one dimensional ordered state space, this sparsity appears along the monotone diagonal. Concave costs may instead favour keeping some mass fixed and moving the remainder farther.

Classical solvers exploit the network structure. Scaling versions of the **auction algorithm** have polynomial pseudopolynomial bounds involving the range of the costs. The **network simplex algorithm** is often fast in practice, although common pivot rules can take exponentially many iterations in the worst case; specially designed scaling pivot rules have polynomial guarantees. **Interior point methods** solve more general linear programs but require repeated **Newton systems**. These costs motivate regularization methods whose basic operation is matrix–vector

multiplication.

Entropic Regularization and Sinkhorn Scaling

Throughout this subtopic, assume that every component of $\boldsymbol{a}$ and $\boldsymbol{b}$ is strictly positive. Zero-mass rows and columns can be removed before applying the scaling formulas.

Definition 3.2 The **Shannon entropy** of a coupling matrix is

$$H(\mathbf{P}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij}\log P_{ij}.$$

For $\varepsilon > 0$, the **entropically regularized transport value** is

$$L_{\mathbf{C}}^{\varepsilon}(\boldsymbol{a},\boldsymbol{b}) = \min_{\mathbf{P}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \big\{\langle \mathbf{P},\mathbf{C}\rangle - \varepsilon H(\mathbf{P})\big\}. \tag{3.3}$$

Since $x \mapsto x\log x$, with $0\log 0 = 0$, is strictly convex on $[0,\infty)$, (3.3) has a unique minimizer. The method was introduced computationally by Cuturi (2013).

For positive vectors of equal total mass, define

$$\mathrm{KL}(\boldsymbol{p} \mid \boldsymbol{q}) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right).$$

If $\boldsymbol{q}$ is uniform, then $\mathrm{KL}(\boldsymbol{p} \mid \boldsymbol{q})$ differs from $-H(\boldsymbol{p})$ by a constant. Thus regularization can be viewed as constraining the relative entropy from the uniform measure. For positive measures of possibly different masses, the extended definition is

$$\mathrm{KL}(\boldsymbol{a} \mid \boldsymbol{b}) = \sum_i \left(a_i \log\left(\frac{a_i}{b_i}\right) - a_i + b_i\right). \tag{3.4}$$

With the convention

$$D_f[\boldsymbol{a}:\boldsymbol{b}] = \sum_i a_i f\left(\frac{b_i}{a_i}\right),$$

this is the $\boldsymbol{f}$**-divergence** generated by $f(u) = -\log u + u - 1$.

Let the **Gibbs kernel** be

$$K_{ij} = \exp(-C_{ij}/\varepsilon). \tag{3.5}$$

A direct calculation shows that the objective in (3.3) differs from $\varepsilon\,\mathrm{KL}(\mathbf{P} \mid \mathbf{K})$ only by constants

that do not depend on $\mathbf{P}$. Therefore

$$\mathbf{P}^\varepsilon = \operatorname*{argmin}_{\mathbf{P}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \mathrm{KL}(\mathbf{P} \mid \mathbf{K}). \tag{3.6}$$

The solution is the **relative entropy projection** of the Gibbs kernel onto the transport polytope.

Theorem 3.3 Let $\mathbf{P}^\varepsilon$ be the unique minimizer in (3.3). Then

$$\lim_{\varepsilon\downarrow 0} \mathbf{P}^\varepsilon = \operatorname{argmax}\left\{H(\mathbf{P}) : \mathbf{P}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b}),\ \langle \mathbf{P},\mathbf{C}\rangle = L_{\mathbf{C}}(\boldsymbol{a},\boldsymbol{b})\right\},$$

whereas

$$\lim_{\varepsilon\to\infty} \mathbf{P}^\varepsilon = \boldsymbol{a}\boldsymbol{b}^{\mathsf{T}}.$$

Proof. Choose a sequence $\varepsilon_\ell \downarrow 0$. Compactness of the transport polytope gives a subsequence for which $\mathbf{P}^{\varepsilon_\ell} \to \mathbf{P}^*$. If $\mathbf{P}$ is any unregularized optimizer, then optimality of $\mathbf{P}^{\varepsilon_\ell}$ gives

$$0 \le \langle \mathbf{C}, \mathbf{P}^{\varepsilon_\ell}\rangle - \langle \mathbf{C}, \mathbf{P}\rangle \le \varepsilon_\ell\big(H(\mathbf{P}^{\varepsilon_\ell}) - H(\mathbf{P})\big).$$

Continuity of entropy on the finite transport polytope shows that $\mathbf{P}^*$ is optimal. Dividing the same inequality by ε_ℓ and passing to the limit yields $H(\mathbf{P}) \le H(\mathbf{P}^*)$. Strict concavity makes the **maximum entropy optimizer** unique, so the entire family converges.

For the second limit, if (X,Y) has joint law $\mathbf{P}$, then

$$I(X;Y) = \mathrm{KL}(\mathbf{P} \mid \boldsymbol{a}\boldsymbol{b}^{\mathsf{T}}) = \sum_{i,j} P_{ij} \log\left(\frac{P_{ij}}{a_i b_j}\right).$$

This **mutual information** is nonnegative and vanishes precisely when X and Y are independent. As the entropy term dominates, the optimizer therefore converges to $\boldsymbol{a}\boldsymbol{b}^{\mathsf{T}}$. □

Theorem 3.4 The minimizer of (3.3) has the form

$$P^\varepsilon_{ij} = u_i v_j K_{ij} \qquad \text{and} \qquad \mathbf{P}^\varepsilon = \operatorname{diag}(\boldsymbol{u})\mathbf{K}\operatorname{diag}(\boldsymbol{v}), \tag{3.7}$$

for positive vectors $\boldsymbol{u}$ and $\boldsymbol{v}$.

Proof. Introduce Lagrange multipliers f_i and g_j for the two marginal constraints. Differentiating

the Lagrangian with respect to P_{ij} gives

$$C_{ij} + \varepsilon(1 + \log P_{ij}) - f_i - g_j = 0.$$

The constant term can be absorbed into either multiplier. Hence

$$P_{ij} = \exp(f_i/\varepsilon) \exp(-C_{ij}/\varepsilon) \exp(g_j/\varepsilon),$$

which has the form (3.7). □

The **scaling vectors** are not individually unique: multiplying $\boldsymbol{u}$ by a positive constant and dividing $\boldsymbol{v}$ by the same constant leaves $\mathbf{P}$ unchanged. The marginal constraints become

$$\boldsymbol{u} \odot (\mathbf{K}\boldsymbol{v}) = \boldsymbol{a} \qquad \text{and} \qquad \boldsymbol{v} \odot (\mathbf{K}^{\mathsf{T}}\boldsymbol{u}) = \boldsymbol{b},$$

where $\odot$ denotes **componentwise multiplication**.

Algorithm 3.5 (Sinkhorn's algorithm) Choose a positive initial vector $\boldsymbol{v}^{(0)}$, such as $\mathbf{1}_m$. For $\ell \geq 0$, update componentwise by

$$\boldsymbol{u}^{(\ell+1)} = \frac{\boldsymbol{a}}{\mathbf{K}\boldsymbol{v}^{(\ell)}} \qquad \text{and} \qquad \boldsymbol{v}^{(\ell+1)} = \frac{\boldsymbol{b}}{\mathbf{K}^{\mathsf{T}}\boldsymbol{u}^{(\ell+1)}}.$$

Stop when both marginal residuals are within the required tolerance.

Each half step enforces one marginal exactly and perturbs the other. The iterates alternately rescale rows and columns until both constraints agree. The work is dominated by multiplication by $\mathbf{K}$ and $\mathbf{K}^{\mathsf{T}}$, which is well suited to parallel hardware and to simultaneous processing of several histograms.

Smaller ε produces a sharper, sparser coupling, while larger ε diffuses mass around the low cost region. Other regularizers can also promote sparsity, but entropy has a special probabilistic connection with the Schrödinger bridge.

For a target additive error τ, choosing $\varepsilon = \tau/(4 \log n)$ and applying a final rounding step yields a feasible coupling $\widehat{\mathbf{P}}$ with

$$\langle \widehat{\mathbf{P}}, \mathbf{C} \rangle \leq L_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) + \tau.$$

A representative bound is

$$O\big(\|\mathbf{C}\|_\infty^3 \log(n) \tau^{-3}\big)$$

Sinkhorn iterations and $O(n^2 \log(n) \tau^{-3})$ arithmetic operations when the cost normalization is

fixed.

Global convergence can be understood through the **Hilbert projective metric** on the **positive cone**:

$$d_H(\boldsymbol{u}, \boldsymbol{u}') = \left\|\log \boldsymbol{u} - \log \boldsymbol{u}'\right\|_{\mathrm{var}} = \log \max_{i,j} \frac{u_i u'_j}{u_j u'_i}, \tag{3.8}$$

where $\|\boldsymbol{z}\|_{\mathrm{var}} = \max_i z_i - \min_i z_i$. This is a metric after identifying positive multiples.

Theorem 3.6 (Birkhoff–Samelson contraction) If $\mathbf{K}$ has strictly positive entries, then

$$d_H(\mathbf{K}\boldsymbol{v}, \mathbf{K}\boldsymbol{v}') \le \lambda(\mathbf{K}) d_H(\boldsymbol{v}, \boldsymbol{v}'),$$

where $0 \le \lambda(\mathbf{K}) < 1$ is explicit.

Theorem 3.7 The scaling vectors in Algorithm 3.5 converge in the Hilbert projective metric to scaling vectors $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ for the unique regularized coupling.

Proof. Entrywise reciprocal scaling is an isometry in d_H. Combining this fact with Theorem 3.6 gives a strict contraction over a full row–column update. For example,

$$d_H(\boldsymbol{u}^{(\ell+1)}, \boldsymbol{u}^*) = d_H(\mathbf{K}\boldsymbol{v}^{(\ell)}, \mathbf{K}\boldsymbol{v}^*) \le \lambda(\mathbf{K}) d_H(\boldsymbol{v}^{(\ell)}, \boldsymbol{v}^*).$$

The analogous estimate for $\boldsymbol{v}$ completes the contraction argument. □

Sinkhorn Divergences and Deconvolution

Even when $C_{ii} = 0$ and $C_{ij} > 0$ for $i \neq j$, the regularized value need not vanish on the diagonal. A debiased version is

$$\widetilde{L}^{\varepsilon}_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) = L^{\varepsilon}_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{b}) - \frac{1}{2} L^{\varepsilon}_{\mathbf{C}}(\boldsymbol{a}, \boldsymbol{a}) - \frac{1}{2} L^{\varepsilon}_{\mathbf{C}}(\boldsymbol{b}, \boldsymbol{b}). \tag{3.9}$$

Under the symmetry, regularity, and positive-kernel hypotheses of the cited result, this **Sinkhorn divergence** is nonnegative and vanishes only when its arguments agree. As the regularization changes, it interpolates between transport geometry and kernel discrepancy geometry; see Feydy, Séjourné, Vialard, Amari, Trouvé, and Peyré (2019).

There is also a statistical interpretation. For quadratic cost, set

$$W_\sigma^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left\{ \frac{1}{2}\int \|\boldsymbol{x}-\boldsymbol{y}\|^2 \,\mathrm{d}\gamma + \sigma^2 I(\gamma) \right\}, \tag{3.10}$$

where $I(\gamma) = \mathrm{KL}(\gamma \mid \mu\otimes\nu)$.

Suppose that latent observations $X_1,\dots,X_n$ are independent with law $\mathbb{P}$, but only

$$Y_i = X_i + \varepsilon_i \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0,\sigma^2\mathbf{I})$$

are observed. Estimating $\mathbb{P}$ from the convolution of $\mathbb{P}$ with Gaussian noise is a **deconvolution problem**.

Theorem 3.8 Let $\mathcal{S}$ be a model class closed under domination: if $\mathbb{Q}\in\mathcal{S}$ and $\mathbb{M}\ll\mathbb{Q}$, then $\mathbb{M}\in\mathcal{S}$. The **nonparametric maximum likelihood estimator** for the deconvolution problem satisfies

$$\widehat{\mathbb{P}} = \operatorname*{argmin}_{\mathbb{Q}\in\mathcal{S}} W_\sigma^2\left(\mathbb{Q}, \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}\right).$$

Thus entropic transport is not merely a numerical smoothing device: its mutual information penalty encodes a **latent variable noise model**. The result is due to Rigollet and Weed (2018).

Schrödinger Bridges and Large Deviations

Consider a cloud of independent particles evolving according to a reference stochastic law $\mathbb{M}$ on a path space Ω. Given an initial distribution μ_0 and an observed terminal distribution μ_1, the **Schrödinger problem** asks for the most likely path law compatible with these endpoints.

For a single Brownian particle conditioned to travel from $\boldsymbol{a}$ to $\boldsymbol{b}$, the **Brownian bridge** fluctuates around the straight segment. Its continuous sample paths remain rough at every scale and have no classical velocity. The three realizations in Figure 6 illustrate this irregularity. As the volatility vanishes, the random path collapses to the constant velocity path from (2.6).

More formally, take $\Omega = C([0,1],\mathbb{R}^d)$ and let $X_t(\omega) = \omega(t)$ be the **canonical process**. Under **Wiener measure**, (X_t) is **Brownian motion**; under another probability on Ω, the same coordinate maps describe a different process.

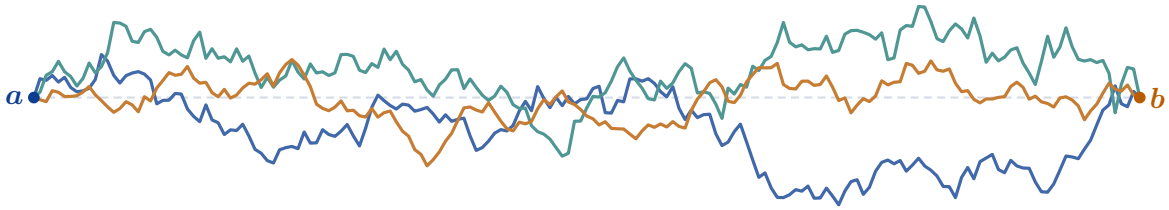

Figure 6

Definition 3.9 Let $\mathbb{M}$ be a reference probability on path space. The **dynamic Schrödinger problem** is

$$\inf_{\substack{\mathbb{P}:\mathbb{P}_0=\mu_0\\ \mathbb{P}_1=\mu_1}} \mathrm{KL}(\mathbb{P} \mid \mathbb{M}).$$

If $\mathbb{M}_{01}$ is the joint endpoint law, the **static Schrödinger problem** is

$$\inf_{\gamma\in\Pi(\mu_0,\mu_1)} \mathrm{KL}(\gamma \mid \mathbb{M}_{01}).$$

Let $\mathbb{M}^{xy}$ be the **reference bridge** from x to y, and disintegrate a candidate path law as $\mathbb{P}^{xy}\,\mathrm{d}\mathbb{P}_{01}(x,y)$. The **chain rule for relative entropy** gives

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{M}) = \mathrm{KL}(\mathbb{P}_{01} \mid \mathbb{M}_{01}) + \int \mathrm{KL}(\mathbb{P}^{xy} \mid \mathbb{M}^{xy})\,\mathrm{d}\mathbb{P}_{01}(x,y). \tag{3.11}$$

The second term is minimized by using the reference bridges themselves. Hence the static and dynamic problems are equivalent; the static optimizer chooses endpoint pairs, and the reference fills in the paths.

Suppose that, relative to a measure η on $X \times X$,

$$\mathrm{d}\mathbb{M}^{\varepsilon}_{01}(x,y) \asymp Z(\varepsilon)\exp\left(-\frac{c(x,y)}{\varepsilon}\right)\mathrm{d}\eta(x,y).$$

Writing $\mathrm{d}\gamma = \rho\,\mathrm{d}\eta$ gives

$$\varepsilon\,\mathrm{KL}(\gamma \mid \mathbb{M}^{\varepsilon}_{01}) = \int c\,\mathrm{d}\gamma + \varepsilon\int \rho\log\rho\,\mathrm{d}\eta + \text{a constant}.$$

Thus the static Schrödinger problem is an entropically regularized transport problem.

Definition 3.10 Let $(\mathbb{P}^\varepsilon)_{\varepsilon>0}$ be probabilities on X, and let $I\colon X \to [0,\infty]$ be lower semi-continuous. The family satisfies a **large deviation principle** with **rate function** I if, for regular enough Borel sets A,

$$\lim_{\varepsilon \downarrow 0} \varepsilon \log \mathbb{P}^\varepsilon(A) = -\inf_{x \in A} I(x).$$

Equivalently at the heuristic exponential scale,

$$\mathbb{P}^\varepsilon(A) \asymp \exp\left(-\frac{1}{\varepsilon} \inf_{x \in A} I(x)\right).$$

The rigorous definition uses separate lower and upper bounds for open and closed sets.

Theorem 3.11 (Sanov's theorem) Let X be finite and fix $\nu \in \mathcal{P}(X)$. If $\mathbb{P}_n$ is the law of the empirical measure

$$\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad X_i \text{ independent with law } \nu,$$

then $(\mathbb{P}_n)$ satisfies a large deviation principle with speed n and rate

$$I(\boldsymbol{p}) = \mathrm{KL}(\boldsymbol{p} \mid \boldsymbol{\nu}).$$

Theorem 3.12 (Schilder's theorem) Let $\boldsymbol{B}$ be standard Brownian motion in $\mathbb{R}^d$, and let $\mathbb{P}_k$ be the law of $\boldsymbol{B}/\sqrt{k}$ on $C([0,1],\mathbb{R}^d)$. Then $(\mathbb{P}_k)$ satisfies a large deviation principle with rate

$$I(\boldsymbol{\omega}) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^1 \|\dot{\boldsymbol{\omega}}_t\|^2 \,\mathrm{d}t, & \boldsymbol{\omega}_0 = \mathbf{0} \text{ and } \boldsymbol{\omega} \text{ is absolutely continuous,} \\ \infty, & \text{otherwise.} \end{cases}$$

The rate in Theorem 3.12 is exactly the **kinetic action** in (2.6). This identifies quadratic transport as the **zero noise limit** of the Brownian Schrödinger bridge.

Theorem 3.13 For each k, let $\rho_k(x,\cdot)$ be a transition kernel and set

$$\mathrm{d}\mathbb{M}_{01}^k(x,y) = \mathrm{d}\mu_0(x)\rho_k(x,\mathrm{d}y).$$

Suppose that, for each x, the kernels satisfy a large deviation principle with rate $c(x,\cdot)$ and speed k. Under the required tightness conditions,

$$\inf_{\gamma\in\Pi(\mu_0,\mu_1)} \frac{1}{k}\,\mathrm{KL}(\gamma \mid \mathbb{M}_{01}^{k}) \longrightarrow \mathcal{T}_c(\mu_0,\mu_1).$$

Every cluster point of the Schrödinger optimizers is optimal for the Monge–Kantorovich problem.

For Brownian kernels, the **entropic interpolation** converges to the displacement interpolation in Definition 2.15.

A Dirichlet Schrödinger Problem

The cost in (2.3) also arises as a large deviation rate. Recall that a random variable G has a gamma distribution with shape α and rate β if

$$f_G(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\mathrm{e}^{-\beta y}, \quad y>0.$$

If G_i are independent with $G_i \sim \mathrm{Gamma}(\alpha_i, 1)$ and

$$D_i = \frac{G_i}{\sum_{j=1}^n G_j},$$

then $\boldsymbol{D} \sim \mathrm{Dir}(\alpha_1,\dots,\alpha_n)$ has a **Dirichlet distribution**.

Use the multiplicative simplex operation to set $\boldsymbol{Q} = \boldsymbol{p} \oplus \boldsymbol{D}$.

Lemma 3.14 The conditional density of $\boldsymbol{Q}$ given $\boldsymbol{p}$ is

$$f(\boldsymbol{q} \mid \boldsymbol{p}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j p_j \Gamma(\alpha_j)} \prod_{i=1}^n \left(\frac{q_i}{p_i}\right)^{\alpha_i - 1} \left(\sum_{j=1}^n \frac{q_j}{p_j}\right)^{-\sum_j \alpha_j}.$$

Proposition 3.15 For each $\lambda > 0$, let $\alpha_i = \lambda/n$. Then

$$\lim_{\lambda\to\infty} -\frac{1}{\lambda}\log f_\lambda(\boldsymbol{q} \mid \boldsymbol{p}) = c(\boldsymbol{p},\boldsymbol{q})$$

up to terms independent of $\boldsymbol{q}$. Thus the **Dirichlet kernels** satisfy a large deviation principle with the Dirichlet transport cost as rate function.

The Dirichlet distribution therefore plays the role that the normal distribution plays in quadratic transport.

To obtain a particle interpretation, draw

$$\boldsymbol{p}^{(1)}, \dots, \boldsymbol{p}^{(N)} \sim \mu_0 \qquad \text{and} \qquad \boldsymbol{q}^{(1)}, \dots, \boldsymbol{q}^{(N)} \sim \mu_1.$$

Independently draw $\boldsymbol{D}^{(j)} \sim \mathrm{Dir}(\lambda/n, \dots, \lambda/n)$ and set

$$\widetilde{\boldsymbol{q}}^{(j)} = \boldsymbol{p}^{(j)} \oplus \boldsymbol{D}^{(j)}.$$

Conditioning the empirical law of the $\widetilde{\boldsymbol{q}}^{(j)}$ to equal the empirical target law produces a random coupling $\widetilde{\gamma}_N$.

Theorem 3.16 Under the technical regularity assumptions, if λ_N is of order $N^{2/n}$, then $\widetilde{\gamma}_N$ converges, with quantitative rates, to the solution of the Dirichlet Monge–Kantorovich problem.

The construction extends to other costs whenever an appropriate large deviation kernel is available. In this limit the c-divergence of Definition 2.11 appears as the local cost of a mismatched endpoint pair.

Finally, a Schrödinger bridge also admits an **Eulerian form**. If

$$\partial_t \rho_t + \nabla \cdot (\rho_t \boldsymbol{v}) = 0, \qquad \rho_0 = \mu, \qquad \text{and} \qquad \rho_1 = \nu,$$

then, under the Brownian reference with noise parameter ε, the bridge minimizes

$$\frac{1}{2} \int_0^1 \int_{\mathbb{R}^d} \left(\|\boldsymbol{v}(t, \boldsymbol{x})\|^2 + \frac{\varepsilon^2}{4} \|\nabla \log \rho_t(\boldsymbol{x})\|^2 \right) \mathrm{d}\rho_t(\boldsymbol{x})\, \mathrm{d}t. \tag{3.12}$$

The first term is **kinetic energy**. The second is the time integral of the Fisher information and penalizes sharply concentrated intermediate densities.

Lecture 4 — Friday, October 04, 2019

3. Riemannian and Information Geometry

Manifolds, Tangent Spaces, and Geodesics

A **smooth n-dimensional manifold** M is locally modelled on $\mathbb{R}^n$. A **chart** identifies a neighbourhood $U \subseteq M$ with coordinates in $\mathbb{R}^n$, and the **changes of coordinates** are diffeomorphisms. The same language can be used heuristically for spaces of probability measures. When the state space has $n+1$ points, the strictly positive distributions form the finite dimensional simplex

$$S_n = \left\{ \boldsymbol{p} = (p_0, \ldots, p_n) : p_i > 0,\ \sum_{i=0}^{n} p_i = 1 \right\}.$$

For a continuous state space, choosing a topology on a space of densities is part of the analysis rather than a formality.

The local geometry is illustrated in Figure 7 by two overlapping **coordinate charts**. Their overlap is encoded by a smooth change of coordinates.

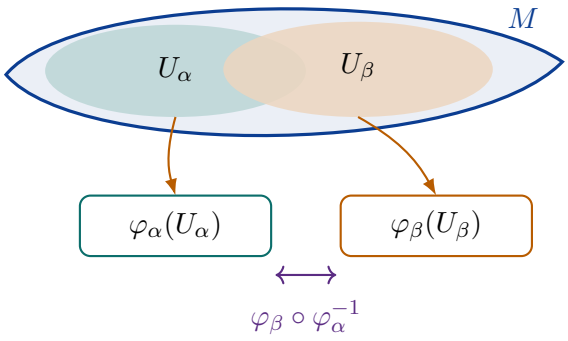

Figure 7

The **tangent space** T_pM consists of **velocity vectors** of smooth curves through p. If $\gamma(0) = p$, its velocity acts on a smooth function by

$$\dot{\gamma}(f) = \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t)\Big|_{t=0}.$$

Definition 4.1 A **Riemannian metric** is a smoothly varying family of inner products

$$g_p \colon T_pM \times T_pM \to \mathbb{R}.$$

In a basis $(\boldsymbol{e}_i)$ of T_pM,

$$g_p(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i,j} g_{ij}(p) u^i v^j.$$

The length of a smooth curve $\gamma \colon [0,1] \to M$ is

$$\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, \mathrm{d}t,$$

and the **Riemannian distance** is the infimum of this length over curves joining the endpoints. A length minimizing curve is a **geodesic**. Equivalently, we may minimize the energy

$$\int_0^1 g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))\, \mathrm{d}t.$$

In coordinates, an **affinely parameterized geodesic** solves

$$\ddot\gamma^k(t) + \sum_{i,j} \Gamma^k_{ij}(\gamma(t)) \dot\gamma^i(t) \dot\gamma^j(t) = 0.$$

The coefficients Γ^k_{ij} are the **Christoffel symbols** of the metric.

Otto's Wasserstein Geometry

The Benamou–Brenier formula in Theorem 2.21 suggests a **weak Riemannian structure** on the space of smooth positive densities. Let M be a compact Riemannian manifold and set

$$\mathcal{P}^\infty_+(M) = \left\{ \rho \in C^\infty(M) : \rho > 0,\ \int_M \rho \,\mathrm{d}\,\mathrm{vol} = 1 \right\}.$$

A **tangent vector** at ρ is a smooth function σ of zero integral. If

$$\sigma = -\operatorname{div}(\rho \nabla \varphi),$$

define **Otto's metric** by

$$g^W_\rho(\sigma, \sigma) = \int_M g^M_{\boldsymbol{x}}(\nabla\varphi, \nabla\varphi) \rho(\boldsymbol{x}) \,\mathrm{d}\,\mathrm{vol}(\boldsymbol{x}). \tag{4.1}$$

The continuity equation explains the definition: if $t \mapsto \rho_t$ has velocity field $\nabla\varphi$ at a given time, then its tangent is $\dot\rho_t = -\operatorname{div}(\rho_t \nabla\varphi)$.

In this notation, Theorem 2.21 becomes the **formal Riemannian energy representation**

$$W_2(\mu_0,\mu_1)^2 = \inf_{\substack{\rho_0=\mu_0\\ \rho_1=\mu_1}} \int_0^1 g^W_{\rho_t}(\dot\rho_t,\dot\rho_t)\,\mathrm{d}t. \tag{4.2}$$

This interpretation, associated with Otto (2001), is formal but predicts the correct differential equations and gradient flows.

Fisher–Rao Geometry and Sufficiency

Consider a smooth parametric family of densities $p(\boldsymbol{x};\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, and denote the corresponding probability measures by $\mathbb{P}_{\boldsymbol{\theta}}$.

Definition 4.2 The **Fisher information matrix** is

$$G_{ij}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\theta_i}\log p(X;\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log p(X;\boldsymbol{\theta})\right]. \tag{4.3}$$

The Cramér–Rao inequality states, under the usual fixed-support differentiation conditions and when $\mathbf{G}(\boldsymbol{\theta})$ is nonsingular, that an unbiased estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\operatorname{Cov}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) \succeq \mathbf{G}(\boldsymbol{\theta})^{-1}.$$

Rao's geometric insight was to regard $\mathbf{G}$ as a Riemannian metric rather than only an information bound.

The same metric is the second-order part of relative entropy.

Proposition 4.3 Suppose that the log density is three times differentiable in a neighbourhood of $\boldsymbol{\theta}$, with derivatives dominated sufficiently to interchange differentiation and integration. If $\boldsymbol{\theta}' = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, then

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}'} \mid \mathbb{P}_{\boldsymbol{\theta}}) = \frac{1}{2}\Delta\boldsymbol{\theta}^{\mathsf{T}}\mathbf{G}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} + O(\|\Delta\boldsymbol{\theta}\|^3).$$

The reverse relative entropy has the same quadratic term.

To see why the metric is intrinsic, let μ be a reference measure and consider the **formal tangent**

space at a positive density ρ. For a tangent σ with integral zero, set

$$g_\rho^F(\sigma,\sigma) = \int_X \frac{|\sigma|^2}{\rho}\,\mathrm{d}\mu. \tag{4.4}$$

Write $p_{\boldsymbol{\theta}} = \mathrm{d}\mathbb{P}_{\boldsymbol{\theta}}/\,\mathrm{d}\mu$. If $\boldsymbol{a} \in T_{\boldsymbol{\theta}}\Theta$, its **pushforward tangent** under $\boldsymbol{\theta} \mapsto p_{\boldsymbol{\theta}}$ is

$$\sigma = \sum_i a_i \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i}.$$

Substitution into (4.4) gives

$$g_{p_{\boldsymbol{\theta}}}^F(\sigma,\sigma) = \sum_{i,j} a_i a_j \int_X \frac{1}{p_{\boldsymbol{\theta}}} \frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_i}\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_j}\,\mathrm{d}\mu = \boldsymbol{a}^\mathsf{T}\mathbf{G}(\boldsymbol{\theta})\boldsymbol{a}.$$

Thus the Fisher information metric is the **pullback** of an **intrinsic metric** on densities.

On the **cone of positive measures**, its squared distance is represented dynamically by

$$d_{FR}(\rho_0,\rho_1)^2 = \inf_{\partial_t \rho = \zeta} \int_0^1 \int_X \frac{|\zeta(t,x)|^2}{\rho(t,x)}\,\mathrm{d}\mu(x)\,\mathrm{d}t. \tag{4.5}$$

The change of variable $r = 2\sqrt{\rho}$ makes the metric Euclidean, so a **positive measure geodesic** is

$$\rho_t(x) = \left((1-t)\sqrt{\rho_0(x)} + t\sqrt{\rho_1(x)}\right)^2.$$

If total mass is constrained to equal one, the **square-root densities** lie on a sphere, and geodesics are **great circle arcs** on that sphere.

For a Riemannian manifold, the gradient of F is determined by

$$g_p(\operatorname{grad} F(p), \boldsymbol{X}) = \boldsymbol{X}F(p).$$

On the simplex with the Fisher metric,

$$(\operatorname{grad} F(\boldsymbol{x}))_i = x_i\left(\frac{\partial F}{\partial x_i} - \sum_j x_j \frac{\partial F}{\partial x_j}\right). \tag{4.6}$$

The negative gradient flow is the **replicator equation** from evolutionary game theory.

The principal reason to privilege the Fisher metric is its invariance under sufficient statistics.

Definition 4.4 For a family $(\mathbb{P}_{\boldsymbol{\theta}})$ and an observation X, a statistic $T(X)$ is a **sufficient statistic** if the conditional distribution

$$\mathbb{P}_{\boldsymbol{\theta}}(X \in \cdot \mid T(X))$$

does not depend on $\boldsymbol{\theta}$.

On a finite state space $I = \{0, \dots, n\}$, a statistic determines a partition

$$T \colon I \to \{A_0, \dots, A_m\}.$$

The induced **coarse graining map** $f \colon S_n \to S_m$ is

$$f(\boldsymbol{p}) = \boldsymbol{q} \qquad \text{and} \qquad q_j = \sum_{i \in A_j} p_i.$$

Sufficiency for a model $M \subseteq S_n$ is equivalent to the factorization

$$p_i = r_{ij} q_j, \quad i \in A_j \qquad \text{and} \qquad \sum_{i \in A_j} r_{ij} = 1,$$

where the conditional probabilities r_{ij} are fixed over the model. This defines the embedding $h \colon S_m \to S_n$ in Figure 8, with $f \circ h = \mathrm{Id}$.

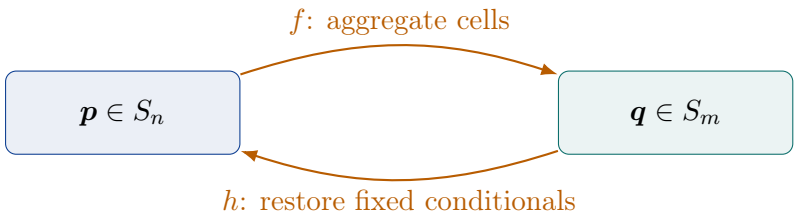

Figure 8

Theorem 4.5 Suppose that each simplex S_n, for $n \geq 2$, is equipped with a Riemannian metric $g^{(n)}$. Assume that every sufficient statistic embedding $h \colon S_m \to S_n$ of the preceding form is an isometry onto its image. Then

$$g^{(n)} = c\, g_F^{(n)}$$

for a constant $c > 0$ independent of n, where

$$g^{(n)}_{F,\boldsymbol{p}}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=0}^{n} \frac{u_i v_i}{p_i}.$$

Proof. At the uniform distribution

$$\bar{\boldsymbol{p}} = \left(\frac{1}{n+1}, \ldots, \frac{1}{n+1} \right),$$

every permutation of the coordinates is an isometry. Permutation invariance forces the ambient coordinate representation of the metric to have the form

$$g^{(n)}_{ij}(\bar{\boldsymbol{p}}) = C_n \delta_{ij} + D_n.$$

Tangent vectors satisfy $\sum_i u_i = \sum_i v_i = 0$, so the term involving D_n vanishes. We may therefore write

$$g^{(n)}_{\bar{\boldsymbol{p}}}(\boldsymbol{u}, \boldsymbol{v}) = C_n \sum_i u_i v_i.$$

Fix m and a rational point $\boldsymbol{q} \in S_m$ of the form

$$q_j = \frac{k_j}{n+1}, \quad k_j \in \mathbb{N} \qquad \text{and} \qquad \sum_{j=0}^{m} k_j = n+1.$$

Choose a partition with $|A_j| = k_j$ and the uniform conditional probabilities $r_{ij} = 1/k_j$ for $i \in A_j$. Then $h(\boldsymbol{q}) = \bar{\boldsymbol{p}}$, and the differential satisfies $(\mathrm{d}h(\boldsymbol{u}))_i = u_j/k_j$ for $i \in A_j$. The isometry assumption gives

$$\begin{aligned} g^{(m)}_{\boldsymbol{q}}(\boldsymbol{u}, \boldsymbol{v}) &= C_n \sum_{j=0}^{m} \sum_{i \in A_j} \frac{u_j}{k_j} \frac{v_j}{k_j} \\ &= \frac{C_n}{n+1} \sum_{j=0}^{m} \frac{u_j v_j}{q_j}. \end{aligned}$$

The left side is independent of the particular denominator $n+1$, so $C_n/(n+1) = c$ for a constant $c > 0$. Rational points are dense in S_m; continuity therefore extends the formula to every point. □

Relative entropy is itself invariant under sufficient statistic transformations. More generally this

holds for every f-divergence. Relative entropy is distinguished further as the only divergence on a finite simplex that is simultaneously an f-divergence and a Bregman divergence.

For one dimensional Gaussian distributions, the Fisher metric produces a familiar geometry. In the coordinates (m, σ), its line element is

$$\mathrm{d}s_F^2 = \frac{(\mathrm{d}m)^2 + 2(\mathrm{d}\sigma)^2}{\sigma^2}.$$

After setting $s = \sqrt{2}\sigma$, this becomes

$$\mathrm{d}s_F^2 = 2\,\frac{(\mathrm{d}m)^2 + (\mathrm{d}s)^2}{s^2}.$$

Thus the rescaled parameter half-plane (m, s) carries a constant multiple of the **Poincaré metric**, whose geodesics are the vertical lines and boundary-orthogonal semicircles shown in Figure 9.

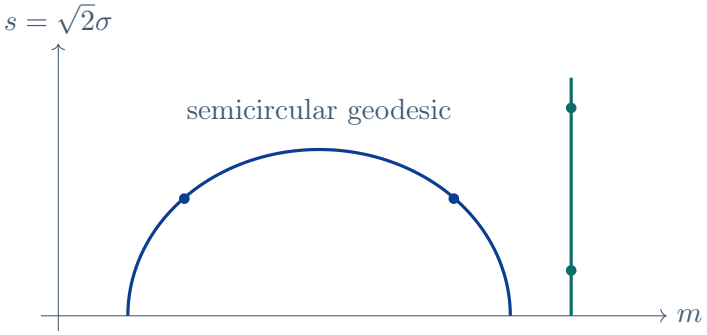

Figure 9

Wasserstein–Fisher–Rao Interpolation

The **Wasserstein action** transports mass using **momentum** $\boldsymbol{\omega} = \rho \boldsymbol{v}$, while the **Fisher–Rao action** creates or destroys mass through a **source** ζ:

$$\inf_{\partial_t \rho + \nabla \cdot \boldsymbol{\omega} = 0} \int_0^1 \int_{\mathbb{R}^n} \frac{\|\boldsymbol{\omega}\|^2}{\rho}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t$$

and

$$\inf_{\partial_t \rho = \zeta} \int_0^1 \int_{\mathbb{R}^n} \frac{|\zeta|^2}{\rho}\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t.$$

Combining the continuity laws gives **unbalanced transport**.

Definition 4.6 For $\kappa > 0$ and nonnegative finite measures ρ_0 and ρ_1, the **Wasserstein–Fisher–Rao distance** is defined by

$$WF_\kappa(\rho_0,\rho_1)^2 = \inf_{\rho,\boldsymbol{\omega},\zeta} \int_0^1 \left[\int_{\mathbb{R}^n} \frac{\|\boldsymbol{\omega}(t,\boldsymbol{x})\|^2}{\rho(t,\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} + \kappa^2 \int_{\mathbb{R}^n} \frac{|\zeta(t,\boldsymbol{x})|^2}{\rho(t,\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \right] \mathrm{d}t, \tag{4.7}$$

subject to

$$\partial_t \rho + \nabla\cdot\boldsymbol{\omega} = \zeta, \qquad \rho(0)=\rho_0, \qquad \text{and} \qquad \rho(1)=\rho_1.$$

The action uses the lower-semicontinuous convention that a quotient with $\rho = 0$ is zero when its numerator is zero and is $+\infty$ otherwise. For general measures, (4.7) denotes the corresponding measure-theoretic relaxation.

The parameter κ controls the cost of changing mass relative to moving it. Large κ favours classical transport; small κ permits mass to disappear at one location and reappear elsewhere.

Partial transport provides a related **static model**. If $\Pi_{\le}(\rho_0,\rho_1)$ denotes measures whose marginals are dominated by ρ_0 and ρ_1, one minimizes

$$\inf_{\substack{\gamma\in\Pi_{\le}(\rho_0,\rho_1)\\ \gamma(X\times Y)=m}} \int c\,\mathrm{d}\gamma.$$

Its dynamic formulation replaces the quadratic source cost in (4.7) by an $\boldsymbol{L^1}$ **source penalty**. More generally, for $p,q>1$ one considers

$$\inf \int_0^1 \left[\frac{1}{p}\int \frac{\|\boldsymbol{\omega}\|^p}{\rho^{p-1}} + \frac{\kappa^q}{q}\int \frac{|\zeta|^q}{\rho^{q-1}} \right] \mathrm{d}t.$$

The existence proof for (4.7) is based on convex duality.

Theorem 4.7 (Fenchel–Rockafellar duality) Let E and F be locally convex Hausdorff spaces, let $A\colon E\to F$ be continuous and linear, and let f and g be proper lower semicontinuous convex functions. If f is finite at some $x\in E$ at which g is continuous at Ax, and if

the primal value below is finite, then

$$\inf_{x\in E}\{f(x)+g(Ax)\} = \max_{y^*\in F^*}\{-f^*(-A^*y^*)-g^*(y^*)\}.$$

The dual maximum is attained. Primal and dual optimizers satisfy the corresponding subgradient conditions.

Theorem 4.8 The action in (4.7) admits an optimizer. Moreover:

1. WF_κ is a metric on the space of finite nonnegative measures;
2. its minimizing interpolation is a constant speed geodesic,
$$WF_\kappa(\rho_s,\rho_t) = |t-s|WF_\kappa(\rho_0,\rho_1);$$
3. for $\alpha>0$,
$$WF_\kappa(\alpha\rho_0,\alpha\rho_1) = \sqrt{\alpha}\,WF_\kappa(\rho_0,\rho_1).$$

Numerically, the interpolation can combine translation, deformation, disappearance, and creation. Two nearby modes primarily move; sufficiently distant modes may shrink at the source and grow at the target, as in Figure 10. Shape interpolation displays the same distinction: Wasserstein motion deforms the outline continuously, Fisher–Rao motion adjusts its intensity, and the combined metric selects the cheaper mechanism.

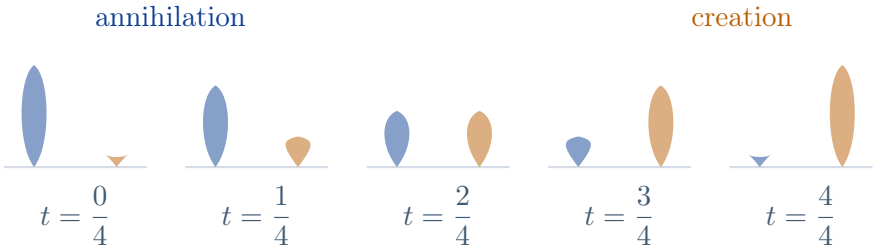

Figure 10

These results and numerical schemes are developed by Chizat, Peyré, Schmitzer, and Vialard (2018).

Natural Gradients and Wasserstein Statistical Manifolds

Euclidean gradient descent iterates

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta \nabla F(\boldsymbol{\theta}^{(k)}).$$

It can be viewed as minimizing the **local model**

$$F(\boldsymbol{\theta}) + \frac{1}{2\eta}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\right\|^2.$$

The Euclidean penalty is coordinate dependent. On a Riemannian model with metric matrix $\mathbf{G}(\boldsymbol{\theta})$, the gradient satisfies

$$\operatorname{grad} F(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})^{-1} \nabla F(\boldsymbol{\theta}).$$

This gives the **natural gradient update**

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta \mathbf{G}(\boldsymbol{\theta}^{(k)})^{-1} \nabla F(\boldsymbol{\theta}^{(k)}). \tag{4.8}$$

With the Fisher metric, the update searches inside a **local relative entropy ball**. Empirical neural network examples show that this **geometric preconditioning** can reduce the number of epochs, though computing and inverting $\mathbf{G}$ introduces its own cost.

The same construction can use the Wasserstein metric. Let

$$\mathcal{M} = \{\rho(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{P}_2(\mathbb{R}^d).$$

Pulling the Wasserstein metric back to Θ gives a **Wasserstein statistical manifold**. Its **intrinsic geodesics** are constrained to remain in $\mathcal{M}$ and therefore need not equal the ambient displacement interpolation. The two agree when $\mathcal{M}$ is **geodesically closed**, as happens for Gaussian families.

In one state space dimension the pullback is explicit.

Proposition 4.9 Let

$$F(x, \boldsymbol{\theta}) = \int_{-\infty}^{x} \rho(t, \boldsymbol{\theta})\, \mathrm{d}t.$$

The Wasserstein metric matrix is

$$\mathbf{G}_W(\boldsymbol{\theta}) = \int_{\mathbb{R}} \frac{1}{\rho(x, \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} F(x, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} F(x, \boldsymbol{\theta})^{\mathsf{T}}\, \mathrm{d}x.$$

For comparison, the Fisher metric is

$$\mathbf{G}_F(\boldsymbol{\theta}) = \int_{\mathbb{R}} \frac{1}{\rho(x,\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \rho(x,\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \rho(x,\boldsymbol{\theta})^{\mathsf{T}} \, \mathrm{d}x.$$

For a mixture of two normal densities, Wasserstein geodesics within the model move the mixture components smoothly, whereas a straight line in the density coordinates may create intermediate multimodal shapes. In parameter estimation, we may minimize

$$\min_{\boldsymbol{\theta}} W_2 \left(\rho(\cdot,\boldsymbol{\theta}), \frac{1}{N} \sum_{j=1}^{N} \delta_{x_j} \right)^2$$

using the **Wasserstein natural gradient**. The one dimensional monotone coupling makes the loss explicit, but scalability remains the main computational difficulty.

Entropy Relaxations and Pseudo-Riemannian Connections

For a finite state space, define the **Cuturi function**

$$C_\lambda(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{1+\lambda} \left(\langle \mathbf{M}, \mathbf{P}^*_\lambda \rangle - \lambda H(\mathbf{P}^*_\lambda) \right), \tag{4.9}$$

where $\mathbf{P}^*_\lambda$ is the entropically regularized optimal coupling with marginals $\boldsymbol{p}$ and $\boldsymbol{q}$. By Theorem 3.4,

$$P^*_{ij} = u_i v_j \exp(-M_{ij}/\lambda).$$

Equivalently, after absorbing normalization into potentials,

$$P^*_{ij} \propto \exp\left(\frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - \frac{1}{\lambda} M_{ij} \right).$$

Consequently, the family of all regularized optimal couplings obtained by varying the marginals is an exponential family, and the convex conjugate of its cumulant generating function is C_λ.

For a fixed reference distribution $\boldsymbol{r}$, define

$$D_{\boldsymbol{r},\lambda}[\boldsymbol{p} : \boldsymbol{q}] = \frac{\lambda}{1+\lambda} \, \mathrm{KL}\left(\mathbf{P}^*_\lambda(\boldsymbol{r},\boldsymbol{p}) \mid \mathbf{P}^*_\lambda(\boldsymbol{r},\boldsymbol{q}) \right).$$

When $\boldsymbol{r} = \boldsymbol{p}$, this has the following **representation resembling a Bregman divergence**:

$$D_{\boldsymbol{p},\lambda}[\boldsymbol{p} : \boldsymbol{q}] = C_\lambda(\boldsymbol{p},\boldsymbol{p}) - C_\lambda(\boldsymbol{p},\boldsymbol{q}) - \nabla_{\boldsymbol{q}} C_\lambda(\boldsymbol{p},\boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}),$$

and

$$\lim_{\lambda\to\infty} D_{\boldsymbol{r},\lambda}[\boldsymbol{p}:\boldsymbol{q}] = \mathrm{KL}(\boldsymbol{p}\mid\boldsymbol{q}).$$

This construction provides one interpolation between optimal transport and information geometry, but it does not by itself give a single geometry encompassing every such interpolation.

A different connection begins with the **cross-difference** of a cost $c\colon M\times M'\to\mathbb{R}$:

$$\delta(p,q',p',q'') = c(p,q'') + c(p',q') - c(p,q') - c(p',q''). \tag{4.10}$$

If (p,q') and (p',q'') lie on an optimal graph, then $\delta\geq 0$ by cyclical monotonicity. Taylor expansion of the cross-difference yields a **pseudo-Riemannian metric** on $M\times M'$. In **local coordinates** $(\boldsymbol{\xi},\boldsymbol{\eta}')$ it has block form

$$\mathbf{h} = \frac{1}{2}\begin{pmatrix} \mathbf{0} & -\overline{D}Dc \\ -D\overline{D}c & \mathbf{0}\end{pmatrix}. \tag{4.11}$$

The **Ma–Trudinger–Wang tensor** is an unnormalized **cross-sectional curvature** of this metric.

Restrict now to the optimal graph G and the c-divergence from Definition 2.11. For $x=(p,q)$ and $x'=(p',q')$ in G, write

$$\delta(x,x') = \delta(p,q,p',q').$$

Theorem 4.10 For $x,x'\in G$,

$$\delta(x,x') = D[x:x'] + D[x':x].$$

Thus the cross-difference is the **symmetrization** of the c-divergence. This identity connects optimality, **divergence geometry**, and the **pseudo-Riemannian curvature framework**.

Lecture 5 — Friday, October 11, 2019

4. Stochastic Control and Extensions

Hamilton–Jacobi Duality

For quadratic transport, the static and dynamic primal problems are

$$\inf_{\gamma\in\Pi(\mu,\nu)} \frac{1}{2}\int \|\boldsymbol{x}-\boldsymbol{y}\|^2 \,\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})$$

and, by Theorem 2.21,

$$\inf_{\rho,\boldsymbol{v}} \int_0^1 \int_{\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{v}_t(\boldsymbol{x})\|^2 \,\mathrm{d}\rho_t(\boldsymbol{x})\,\mathrm{d}t.$$

The dynamic formulation leads to a **Hamilton–Jacobi description** of the Kantorovich potentials.

If $(\rho,\boldsymbol{v})$ satisfies the continuity equation, then every smooth compactly supported test function φ satisfies

$$\int_0^1 \int_{\mathbb{R}^n} (\partial_t\varphi + \nabla\varphi\cdot\boldsymbol{v})\,\mathrm{d}\rho_t\,\mathrm{d}t = \int_{\mathbb{R}^n} \varphi(1,\boldsymbol{y})\,\mathrm{d}\nu(\boldsymbol{y}) - \int_{\mathbb{R}^n} \varphi(0,\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}). \tag{5.1}$$

Introduce the momentum $\boldsymbol{\omega} = \rho\boldsymbol{v}$ and enforce (5.1) by a Lagrange multiplier. Formally interchanging the infimum and supremum produces the pointwise minimization

$$\inf_{\rho\geq 0,\boldsymbol{\omega}} \left\{ \frac{\|\boldsymbol{\omega}\|^2}{2\rho} - (\partial_t\varphi)\rho - \nabla\varphi\cdot\boldsymbol{\omega} \right\}.$$

For fixed ρ, the minimizer is $\boldsymbol{\omega} = \rho\nabla\varphi$. The pointwise infimum over $\rho \geq 0$ is finite only for the dual subsolutions

$$\partial_t\varphi + \frac{1}{2}\|\nabla\varphi\|^2 \leq 0.$$

On the support of an optimal density, complementary slackness gives equality:

$$\partial_t\varphi + \frac{1}{2}\|\nabla\varphi\|^2 = 0. \tag{5.2}$$

At optimality, $-\varphi(0,\cdot)$ and $\varphi(1,\cdot)$ are Kantorovich potentials. Thus the static dual pair can be joined by a **Hamilton–Jacobi evolution**.

Stochastic Control and Girsanov's Theorem

Let $\varepsilon > 0$ be a noise parameter. Consider continuous semimartingales of the form

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{\beta}_{\boldsymbol{X}}(s,\boldsymbol{X})\,\mathrm{d}s + \sqrt{\varepsilon}\,\boldsymbol{W}_{\boldsymbol{X}}(t), \tag{5.3}$$

where $\boldsymbol{W}_{\boldsymbol{X}}$ is Brownian motion adapted to the **filtration** of $\boldsymbol{X}$. Given **endpoint laws** $\mathbb{P}_0$ and $\mathbb{P}_1$, define

$$V_\varepsilon(\mathbb{P}_0, \mathbb{P}_1) = \inf \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\boldsymbol{\beta}_{\boldsymbol{X}}(t, \boldsymbol{X})\|^2 \,\mathrm{d}t\right], \tag{5.4}$$

where the infimum is over processes satisfying (5.3), $\boldsymbol{X}_0 \sim \mathbb{P}_0$, and $\boldsymbol{X}_1 \sim \mathbb{P}_1$.

Unlike a **standard control problem**, both endpoint distributions are constrained. A more customary **controlled diffusion** begins from a fixed state,

$$\boldsymbol{X}_t^{\boldsymbol{\alpha}} = \boldsymbol{x}_0 + \int_0^t \boldsymbol{b}(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_s)\,\mathrm{d}s + \int_0^t \boldsymbol{\sigma}(s, \boldsymbol{X}_s^{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_s)\,\mathrm{d}\boldsymbol{W}_s,$$

and minimizes

$$J(\boldsymbol{\alpha}) = \mathbb{E}\left[\int_0^T f(t, \boldsymbol{X}_t^{\boldsymbol{\alpha}})\,\mathrm{d}t + g(\boldsymbol{X}_T^{\boldsymbol{\alpha}})\right].$$

The endpoint law constraint is what ties (5.4) to transport and Schrödinger bridges.

Suppose that $X \sim \mathcal{N}(0, \sigma^2)$ under $\mathbb{P}$. The following **change of measure** displays the basic mechanism:

$$Z = \exp\left(-\frac{\theta}{\sigma^2}X - \frac{\theta^2}{2\sigma^2}\right).$$

The density Z defines a probability $\mathbb{Q}$ by $\mathrm{d}\mathbb{Q} = Z\,\mathrm{d}\mathbb{P}$ under which $X \sim \mathcal{N}(-\theta, \sigma^2)$. Girsanov's theorem applies the same infinitesimal shift to Brownian increments.

Theorem 5.1 (Girsanov's theorem) Let $(\boldsymbol{\theta}_t)_{0\le t\le T}$ be adapted.

Suppose that the **exponential martingale**

$$Z_T = \exp\left(-\int_0^T \boldsymbol{\theta}_t \cdot \mathrm{d}\boldsymbol{W}_t - \frac{1}{2}\int_0^T \|\boldsymbol{\theta}_t\|^2\,\mathrm{d}t\right)$$

has expectation one. Define $\mathbb{Q}$ by $\mathrm{d}\mathbb{Q} = Z_T\,\mathrm{d}\mathbb{P}$. Then

$$\widetilde{\boldsymbol{W}}_t = \boldsymbol{W}_t + \int_0^t \boldsymbol{\theta}_s\,\mathrm{d}s$$

is Brownian motion under $\mathbb{Q}$.

Apply Theorem 5.1 on **canonical Wiener space** to

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \sqrt{\varepsilon}\boldsymbol{W}_t.$$

Choosing $\boldsymbol{\theta} = -\boldsymbol{\beta}/\sqrt{\varepsilon}$ turns the canonical process under the new measure into

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{\beta}_s \, \mathrm{d}s + \sqrt{\varepsilon}\widetilde{\boldsymbol{W}}_t.$$

The cost of this change of measure is relative entropy. Here $\mathbb{P}$ is the reference law and $\mathbb{Q}$ is the controlled law. The stochastic integral has mean zero under the controlled law, and

$$\mathrm{KL}(\mathbb{Q} \mid \mathbb{P}) = \frac{1}{2\varepsilon}\mathbb{E}_{\mathbb{Q}}\left[\int_0^1 \|\boldsymbol{\beta}_t\|^2 \, \mathrm{d}t\right]. \tag{5.5}$$

Equation (5.5) makes the equivalence between the **control energy** and the Schrödinger relative entropy transparent.

Noise adds diffusion to the **Hamilton–Jacobi equation**. The resulting **viscous Hamilton–Jacobi–Bellman equation** is

$$\partial_t \varphi + \frac{\varepsilon}{2}\Delta\varphi + \frac{1}{2}\|\nabla\varphi\|^2 = 0. \tag{5.6}$$

Theorem 5.2 Under the regularity assumptions for the **control problem with fixed endpoints**,

$$V_\varepsilon(\mathbb{P}_0, \mathbb{P}_1) = \sup_{\varphi} \left\{ \int \varphi(1, \boldsymbol{y}) \, \mathrm{d}\mathbb{P}_1(\boldsymbol{y}) - \int \varphi(0, \boldsymbol{x}) \, \mathrm{d}\mathbb{P}_0(\boldsymbol{x}) \right\},$$

where the supremum is over smooth subsolutions of

$$\partial_t \varphi + \frac{\varepsilon}{2}\Delta\varphi + \frac{1}{2}\|\nabla\varphi\|^2 \le 0,$$

or over suitably interpreted viscosity subsolutions. Under conditions that permit solving the backward equation from arbitrary terminal data, the supremum may equivalently be parametrized by solutions of (5.6).

As $\varepsilon \downarrow 0$, (5.6) approaches (5.2), and the **stochastic dual** converges to Kantorovich duality. This formulation is due to Mikami and Thieullen (2006) and Mikami and Thieullen (2008).

Transport with Prior Dynamics

Quadratic transport can be generalized by imposing a **preferred linear dynamics**:

$$\mathrm{d}\boldsymbol{x}(t) = \mathbf{A}(t)\boldsymbol{x}(t) \, \mathrm{d}t + \mathbf{B}(t)\boldsymbol{u}(t) \, \mathrm{d}t + \sqrt{\varepsilon}\mathbf{B}(t) \, \mathrm{d}\boldsymbol{w}(t). \tag{5.7}$$

The state satisfies $\boldsymbol{x}(0) \sim \mu_0$ and $\boldsymbol{x}(1) \sim \mu_1$, and the objective is

$$\mathbb{E}\left[\int_0^1 \frac{1}{2}\|\boldsymbol{u}(t)\|^2 \,\mathrm{d}t\right].$$

When $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \mathbf{I}$, and $\varepsilon = 0$, this is the Benamou–Brenier problem. A nonzero $\mathbf{A}$ models an existing flow or force that the control should exploit rather than cancel unnecessarily.

Let $\boldsymbol{\Phi}(t, s)$ be the transition matrix satisfying

$$\partial_t \boldsymbol{\Phi}(t, s) = \mathbf{A}(t)\boldsymbol{\Phi}(t, s) \qquad \text{and} \qquad \boldsymbol{\Phi}(s, s) = \mathbf{I},$$

and define the **controllability Gramian**

$$\mathbf{M}(t_1, t_0) = \int_{t_0}^{t_1} \boldsymbol{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{B}(\tau)^{\mathsf{T}}\boldsymbol{\Phi}(t_1, \tau)^{\mathsf{T}} \,\mathrm{d}\tau. \tag{5.8}$$

The system is **controllable** when this matrix is invertible. Probabilistically, control can then steer the process from any initial point into any nonempty open target set with positive probability.

Set $\boldsymbol{\Phi}_{10} = \boldsymbol{\Phi}(1, 0)$ and $\mathbf{M}_{10} = \mathbf{M}(1, 0)$. In the deterministic case, the minimum energy required to travel from $\boldsymbol{x}$ to $\boldsymbol{y}$ is

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\Phi}_{10}\boldsymbol{x})^{\mathsf{T}}\mathbf{M}_{10}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}_{10}\boldsymbol{x}). \tag{5.9}$$

The optimal control is

$$\boldsymbol{u}^*(t) = \mathbf{B}(t)^{\mathsf{T}}\boldsymbol{\Phi}(1, t)^{\mathsf{T}}\mathbf{M}_{10}^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}_{10}\boldsymbol{x}),$$

and the associated state trajectory is

$$\boldsymbol{x}^*(t) = \boldsymbol{\Phi}(t, 1)\mathbf{M}(1, t)\mathbf{M}_{10}^{-1}\boldsymbol{\Phi}_{10}\boldsymbol{x} + \mathbf{M}(t, 0)\boldsymbol{\Phi}(1, t)^{\mathsf{T}}\mathbf{M}_{10}^{-1}\boldsymbol{y}.$$

The change of variables

$$\widehat{\boldsymbol{x}} = \mathbf{M}_{10}^{-1/2}\boldsymbol{\Phi}_{10}\boldsymbol{x} \qquad \text{and} \qquad \widehat{\boldsymbol{y}} = \mathbf{M}_{10}^{-1/2}\boldsymbol{y}$$

turns (5.9) into $\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|^2/2$. Therefore the problem reduces to quadratic transport between the transformed endpoint laws. If $\widehat{T} = \nabla\phi$ is the corresponding Brenier map, then

$$T(\boldsymbol{x}) = \mathbf{M}_{10}^{1/2}\widehat{T}(\mathbf{M}_{10}^{-1/2}\boldsymbol{\Phi}_{10}\boldsymbol{x}). \tag{5.10}$$

The interpolating map is

$$T_t(\boldsymbol{x}) = \boldsymbol{\Phi}(t,1)\mathbf{M}(1,t)\mathbf{M}_{10}^{-1}\boldsymbol{\Phi}_{10}\boldsymbol{x} + \mathbf{M}(t,0)\boldsymbol{\Phi}(1,t)^{\mathsf{T}}\mathbf{M}_{10}^{-1}T(\boldsymbol{x}),$$

which is generally not a straight line displacement interpolation.

The difference is visible even in one dimension. Under the prior drift $\mathrm{d}x = -2x\,\mathrm{d}t + u\,\mathrm{d}t$, uncontrolled particles contract toward the origin. The optimal controlled paths in Figure 11 curve so that they cooperate with this drift, whereas **McCann interpolation** ignores it.

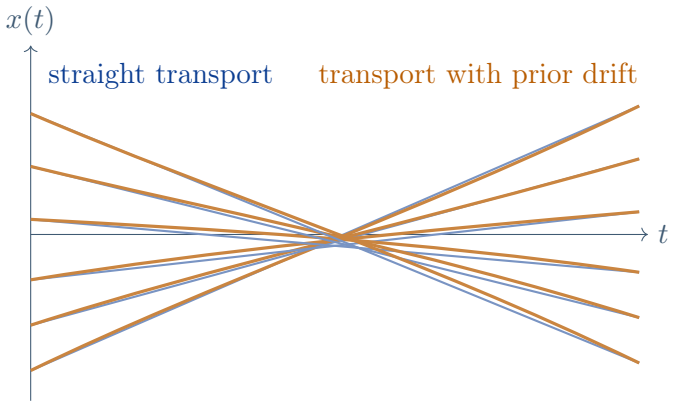

Figure 11

The same transport can be recovered from dynamic programming. For a terminal cost G, define

$$V(t,\boldsymbol{x}) = \inf_{\boldsymbol{u}} \left\{ \int_t^1 \frac{1}{2}\|\boldsymbol{u}(s)\|^2\,\mathrm{d}s + G(\boldsymbol{x}^{\boldsymbol{u}}(1)) \right\}.$$

Bellman's dynamic programming principle yields

$$0 = \inf_{\boldsymbol{u}} \left\{ \frac{1}{2}\|\boldsymbol{u}\|^2 + \partial_t V + \nabla V \cdot (\mathbf{A}(t)\boldsymbol{x} + \mathbf{B}(t)\boldsymbol{u}) \right\}.$$

Optimizing explicitly gives

$$\partial_t V + \nabla V \cdot \mathbf{A}(t)\boldsymbol{x} - \frac{1}{2}\nabla V^{\mathsf{T}}\mathbf{B}(t)\mathbf{B}(t)^{\mathsf{T}}\nabla V = 0, \tag{5.11}$$

with $V(1,\boldsymbol{y}) = G(\boldsymbol{y})$, and

$$\boldsymbol{u}^*(t,\boldsymbol{x}) = -\mathbf{B}(t)^{\mathsf{T}}\nabla V(t,\boldsymbol{x}).$$

An explicit potential is obtained from the **generalized Hopf–Lax formula**

$$\psi(t,\boldsymbol{x}) = \inf_{\boldsymbol{y}} \left\{ \psi(0,\boldsymbol{y}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\Phi}(t,0)\boldsymbol{y})^{\mathsf{T}}\mathbf{M}(t,0)^{-1}(\boldsymbol{x} - \boldsymbol{\Phi}(t,0)\boldsymbol{y}) \right\},$$

with $V = -\psi$ and an initial potential determined by the Brenier potential of the transformed problem.

Schrödinger Bridges with a Markov Reference

Let $q(s, \boldsymbol{x}, t, \boldsymbol{y})$ be a positive Markov transition density. It satisfies the Chapman–Kolmogorov identity

$$q(t_1, \boldsymbol{x}_1, t_3, \boldsymbol{x}_3) = \int q(t_1, \boldsymbol{x}_1, t_2, \boldsymbol{x}_2) q(t_2, \boldsymbol{x}_2, t_3, \boldsymbol{x}_3) \, \mathrm{d}\boldsymbol{x}_2. \tag{5.12}$$

Let $\mathbb{Q}$ be the corresponding Markov path law. Its endpoint distribution has density $q(0, \boldsymbol{x}, 1, \boldsymbol{y})$ relative to the chosen initial reference.

Theorem 5.3 The static Schrödinger bridge with positive Markov reference has endpoint law

$$\mathrm{d}\mathbb{P}_{01}(\boldsymbol{x}, \boldsymbol{y}) = q(0, \boldsymbol{x}, 1, \boldsymbol{y}) \widehat{\varphi}_0(\boldsymbol{x}) \varphi_1(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \tag{5.13}$$

for positive functions $\widehat{\varphi}_0$ and φ_1. The time-t marginal is

$$\rho_t(\boldsymbol{x}) = \varphi(t, \boldsymbol{x}) \widehat{\varphi}(t, \boldsymbol{x}), \tag{5.14}$$

where

$$\varphi(t, \boldsymbol{x}) = \int q(t, \boldsymbol{x}, 1, \boldsymbol{y}) \varphi_1(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$$

and

$$\widehat{\varphi}(t, \boldsymbol{x}) = \int q(0, \boldsymbol{y}, t, \boldsymbol{x}) \widehat{\varphi}_0(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.$$

The two functions propagate backward and forward through the **Markov semigroup**. Equation (5.12) then turns the endpoint factorization (5.13) into the intermediate factorization (5.14). In a finite state space, this is the continuous time counterpart of the two Sinkhorn scaling vectors.

Return to the stochastic linear system (5.7). The **Eulerian control problem** is

$$\inf_{\rho, \boldsymbol{u}} \int_0^1 \int \frac{1}{2} \|\boldsymbol{u}(t, \boldsymbol{x})\|^2 \rho(t, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t$$

subject to the **Fokker–Planck equation**

$$\partial_t \rho + \nabla \cdot \big((\mathbf{A}(t)\boldsymbol{x} + \mathbf{B}(t)\boldsymbol{u})\rho \big) - \frac{\varepsilon}{2} \sum_{i,j} \frac{\partial^2 (a_{ij}(t)\rho)}{\partial x_i \partial x_j} = 0, \tag{5.15}$$

$$\rho(0, \cdot) = \rho_0 \qquad \text{and} \qquad \rho(1, \cdot) = \rho_1,$$

where $\mathbf{a}(t) = \mathbf{B}(t)\mathbf{B}(t)^{\mathsf{T}}$.

Theorem 5.4 Under suitable regularity conditions, the optimal control is

$$\boldsymbol{u}^*(t, \boldsymbol{x}) = \varepsilon \mathbf{B}(t)^{\mathsf{T}} \nabla \log \varphi(t, \boldsymbol{x}),$$

where φ is the backward factor from Theorem 5.3 for the uncontrolled prior dynamics.

The drift is a **Doob h-transform**: the logarithmic gradient of the **backward likelihood** tilts the reference process just enough to attain the terminal marginal. When the endpoint laws are Gaussian and the prior is linear, every interpolating law remains Gaussian and the potentials are explicit. A second-order example,

$$\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{v}(t)\,\mathrm{d}t \qquad \text{and} \qquad \mathrm{d}\boldsymbol{v}(t) = \sqrt{\varepsilon}\,\mathrm{d}\boldsymbol{w}(t) + \boldsymbol{u}(t)\,\mathrm{d}t,$$

controls acceleration rather than position directly; the evolving distribution lives on **position–velocity phase space**.

Multimarginal, Unbalanced, and Martingale Transport

The **multimarginal extension** replaces a coupling of two measures by a joint law with prescribed marginals $\alpha_1, \ldots, \alpha_S$:

$$\inf_{\gamma \in \Pi(\alpha_1, \ldots, \alpha_S)} \int c(x_1, \ldots, x_S)\,\mathrm{d}\gamma. \tag{5.16}$$

An **entropic penalty** may be added exactly as in (3.3).

An economic example matches one consumer of type x_1 with contractors of types $x_2, \ldots, x_S$. If a team produces project y, its negative welfare is

$$-p_1(x_1, y) + \sum_{i=2}^{S} c_i(x_i, y).$$

Optimizing the project within each team defines the **multimarginal cost**

$$c(x_1, \ldots, x_S) = \inf_y \left\{ -p_1(x_1, y) + \sum_{i=2}^{S} c_i(x_i, y) \right\}.$$

Problem (5.16) then maximizes aggregate welfare by choosing the assignment of types.

Unbalanced transport relaxes the marginal constraints and penalizes their violation. For a non-

negative transport measure γ with marginals α and β, one minimizes

$$\inf_{\gamma \geq 0} \left\{ \int c \, \mathrm{d}\gamma + \tau D_f(\alpha \mid \mu) + \tau D_f(\beta \mid \nu) \right\}. \tag{5.17}$$

This is valuable when total signal strength matters, normalization is not meaningful, or the observed marginals contain noise. In such settings, enforcing exact mass conservation can amplify error.

Martingale transport adds a conditional expectation constraint. Let $\mu_1, \ldots, \mu_N$ be probability measures on $\mathbb{R}$ in **increasing convex order**, meaning that

$$t \longmapsto \int \phi \, \mathrm{d}\mu_t$$

is nondecreasing for every convex ϕ. This is necessary and sufficient for them to be the marginals of a martingale (X_t) satisfying

$$\mathbb{E}[X_t \mid X_1, \ldots, X_s] = X_s, \quad s \leq t.$$

The **martingale optimal transport problem** minimizes $\int c \, \mathrm{d}\gamma$ over all such joint laws. In finance, c may be the payoff of a **path-dependent option**. The infimum and supremum then give bounds on arbitrage-free prices without specifying a model, while the dual variables describe hedging strategies.

Transport can also be defined for positive matrices, possibly of unit trace. Unlike a scalar histogram, a matrix records both eigenvalues and eigenvectors. Relevant costs include the Bures–Wasserstein metric and **matrix relative entropy**

$$\mathrm{KL}(\mathbf{A} \mid \mathbf{B}) = \mathrm{tr}\left(\mathbf{A} \log \mathbf{A} - \mathbf{A} \log \mathbf{B} - \mathbf{A} + \mathbf{B}\right).$$

This points toward quantum and matrix analogues of the Dirichlet transport and $L^{(\alpha)}$-divergence.

Gromov–Wasserstein Geometry

When two measures live on different spaces, a pointwise cost between their locations may not be defined. **Gromov-type distances** compare the internal geometries instead.

For nonempty compact subsets A, B of a metric space (Z, d_Z), the **Hausdorff distance** is

$$H_Z(A, B) = \max\left\{ \sup_{a \in A} \inf_{b \in B} d_Z(a, b), \sup_{b \in B} \inf_{a \in A} d_Z(a, b) \right\}.$$

The **Gromov–Hausdorff distance** between compact metric spaces (X, d_X) and (Y, d_Y) is

$$GH(X, Y) = \inf_{Z, f, g} H_Z(f(X), g(Y)), \tag{5.18}$$

where $f\colon X \to Z$ and $g\colon Y \to Z$ are isometric embeddings.

An equivalent formulation uses a correspondence.

Definition 5.5 A **correspondence** between X and Y is a subset $R \subseteq X \times Y$ whose projections onto both factors are surjective.

Then

$$GH(X, Y) = \frac{1}{2} \inf_{R} \sup_{(x,y),(x',y') \in R} \left| d_X(x, x') - d_Y(y, y') \right|. \tag{5.19}$$

The supremum is the **distortion** of the correspondence.

Figure 12 compares two configurations by their internal distances alone.

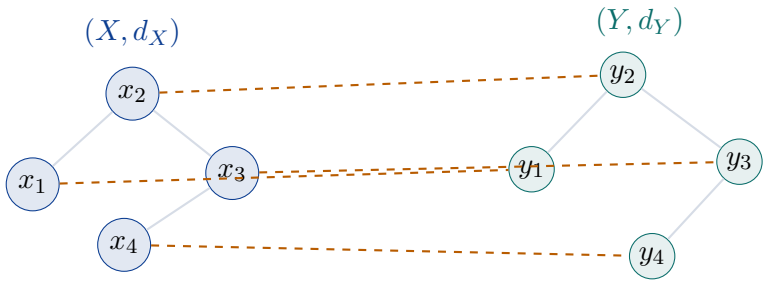

Figure 12

For metric-measure spaces (X, d_X, μ_X) and (Y, d_Y, μ_Y), replace a set correspondence by a probabilistic coupling.

Definition 5.6 The squared **Gromov–Wasserstein distance** is

$$GW(X, Y)^2 = \inf_{\gamma \in \Pi(\mu_X, \mu_Y)} \iint \left| d_X(x, x') - d_Y(y, y') \right|^2 \, \mathrm{d}\gamma(x, y) \, \mathrm{d}\gamma(x', y'). \tag{5.20}$$

This distance is invariant under measure-preserving isometries. It is a smooth probabilistic analogue of (5.19) and is useful in shape matching, where objects may have no shared coordinate system.

Lecture 6 — Friday, October 18, 2019

5. Variational Problems and Statistical Learning

Barycentres and Exponential Families

Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ and weights $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$, the **weighted Euclidean mean**

$$\bar{\boldsymbol{x}}_\lambda = \sum_{i=1}^N \lambda_i \boldsymbol{x}_i$$

is the unique minimizer of

$$\boldsymbol{x} \longmapsto \sum_{i=1}^N \lambda_i \|\boldsymbol{x} - \boldsymbol{x}_i\|^2.$$

A **barycentre** replaces the squared Euclidean distance by another divergence or metric.

Theorem 6.1 Let D_ϕ be the Bregman divergence from Definition 2.10. Its **right barycentre** is the arithmetic mean:

$$\operatorname*{argmin}_{\boldsymbol{s}} \sum_{i=1}^N \lambda_i D_\phi[\boldsymbol{x}_i : \boldsymbol{s}] = \sum_{i=1}^N \lambda_i \boldsymbol{x}_i.$$

Proof. Set $\bar{\boldsymbol{x}}_\lambda = \sum_i \lambda_i \boldsymbol{x}_i$. Expanding the definition of the Bregman divergence gives

$$\begin{aligned}\sum_{i=1}^N \lambda_i D_\phi[\boldsymbol{x}_i : \boldsymbol{s}] &= \sum_{i=1}^N \lambda_i \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{s}) - \nabla\phi(\boldsymbol{s}) \cdot (\bar{\boldsymbol{x}}_\lambda - \boldsymbol{s}) \\ &= \sum_{i=1}^N \lambda_i \phi(\boldsymbol{x}_i) - \phi(\bar{\boldsymbol{x}}_\lambda) + D_\phi[\bar{\boldsymbol{x}}_\lambda : \boldsymbol{s}].\end{aligned}$$

The first two terms do not depend on $\boldsymbol{s}$. The final term is nonnegative and, by strict convexity, vanishes only when $\boldsymbol{s} = \bar{\boldsymbol{x}}_\lambda$. Thus the weighted mean is the unique minimizer. □

This observation underlies **Bregman clustering**; see Banerjee, Merugu, Dhillon, and Ghosh (2005).

Now consider an exponential family

$$p(x;\boldsymbol{\theta}) = \exp\big(\boldsymbol{\theta}\cdot\boldsymbol{h}(x) - \phi(\boldsymbol{\theta})\big) \tag{6.1}$$

relative to a fixed base measure. The log-partition function ϕ is convex, and

$$D_\phi[\boldsymbol{\theta} : \boldsymbol{\theta}'] = \mathrm{KL}(p(\cdot;\boldsymbol{\theta}') \mid p(\cdot;\boldsymbol{\theta})).$$

Its convex conjugate is **negative Shannon entropy**,

$$\psi(\boldsymbol{\eta}) = -H(p(\cdot;\boldsymbol{\theta})) = \int p(x;\boldsymbol{\theta}) \log p(x;\boldsymbol{\theta})\, \mathrm{d}\mu(x),$$

where

$$\boldsymbol{\eta} = \nabla\phi(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{h}(X)]$$

is the expectation coordinate. If $\boldsymbol{\eta}' = \nabla\phi(\boldsymbol{\theta}')$, then

$$D_\phi[\boldsymbol{\theta} : \boldsymbol{\theta}'] = D_\psi[\boldsymbol{\eta}' : \boldsymbol{\eta}] = \phi(\boldsymbol{\theta}) + \psi(\boldsymbol{\eta}') - \boldsymbol{\theta}\cdot\boldsymbol{\eta}'. \tag{6.2}$$

For a data point x for which $\boldsymbol{h}(x)$ lies in the effective domain of ψ, take $\boldsymbol{\eta}' = \boldsymbol{h}(x)$. Equations (6.1) and (6.2) give

$$\log p(x;\boldsymbol{\theta}) = -D_\psi[\boldsymbol{h}(x) : \boldsymbol{\eta}] + \psi(\boldsymbol{h}(x)).$$

Therefore maximum likelihood for observations $x_1, \dots, x_N$ is equivalent to

$$\operatorname*{argmin}_{\boldsymbol{\eta}} \sum_{i=1}^{N} D_\psi[\boldsymbol{h}(x_i) : \boldsymbol{\eta}],$$

and Theorem 6.1 gives

$$\widehat{\boldsymbol{\eta}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}(x_i).$$

Thus the maximum likelihood estimator is a right Bregman barycentre in expectation coordinates.

Wasserstein Barycentres

Definition 6.2 Let $\nu_1, \dots, \nu_N \in \mathcal{P}_2(\mathbb{R}^d)$ and let $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$. A **Wasserstein barycentre** is a minimizer of

$$\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{N} \lambda_i W_2(\nu_i, \nu)^2. \tag{6.3}$$

Theorem 6.3 A Wasserstein barycentre exists. It is unique if at least one positive-weight measure ν_i vanishes on every Borel set of Hausdorff dimension at most $d-1$; in particular, absolute continuity of one positive-weight ν_i is sufficient.

This result is due to Agueh and Carlier (2011).

Zero-weight inputs may be discarded, so assume in the following dual formulas that every $\lambda_i > 0$. The dual problem uses a **weighted quadratic infimal convolution**. Define

$$S_\lambda f(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in \mathbb{R}^d} \left\{ \frac{\lambda}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 - f(\boldsymbol{y}) \right\}.$$

With a harmless factor of $1/2$ in the primal objective, the dual is

$$\sup_{\substack{f_1, \dots, f_N \\ \sum_i f_i = 0}} \sum_{i=1}^{N} \int S_{\lambda_i} f_i \, \mathrm{d}\nu_i. \tag{6.4}$$

Indeed,

$$S_{\lambda_i} f_i(\boldsymbol{x}) + f_i(\boldsymbol{y}) \leq \frac{\lambda_i}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Integrating this inequality against couplings of ν_i and a candidate barycentre ν, then summing over i, eliminates the functions f_i because $\sum_i f_i = 0$. Convex duality turns the resulting weak inequality into equality.

There is also a **multimarginal formulation**. For $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, set

$$T(\boldsymbol{x}) = \sum_{i=1}^{N} \lambda_i \boldsymbol{x}_i.$$

Consider

$$\inf_{\gamma\in\Pi(\nu_1,\dots,\nu_N)} \int \sum_{i=1}^{N} \frac{\lambda_i}{2} \|\boldsymbol{x}_i - T(\boldsymbol{x})\|^2 \,\mathrm{d}\gamma(\boldsymbol{x}). \tag{6.5}$$

Theorem 6.4 If γ solves (6.5), then

$$\nu = T_{\#}\gamma$$

solves (6.3).

Proof. Let $\pi_i(\boldsymbol{x}) = \boldsymbol{x}_i$. Then

$$\gamma_i = (\pi_i, T)_{\#}\gamma \in \Pi(\nu_i, \nu),$$

and hence

$$W_2(\nu_i, \nu)^2 \le \int \|\boldsymbol{x}_i - T(\boldsymbol{x})\|^2 \,\mathrm{d}\gamma(\boldsymbol{x}).$$

Summing with weights proves one inequality between the two problems. Conversely, glue optimal couplings of a candidate ν with each ν_i over their common ν-coordinate, just as in the proof of Theorem 2.5. Conditional on a barycentre point $\boldsymbol{y}$, the weighted mean minimizes

$$\boldsymbol{z} \longmapsto \sum_i \lambda_i \|\boldsymbol{x}_i - \boldsymbol{z}\|^2.$$

This gives the reverse inequality and shows that a multimarginal optimizer pushes forward to a barycentre. □

In one dimension, **quantiles** give an explicit solution. If F_i and F are the **distribution functions** of ν_i and ν, then

$$W_2(\nu_i, \nu)^2 = \int_0^1 \left|F_i^{-1}(u) - F^{-1}(u)\right|^2 \mathrm{d}u.$$

Pointwise minimization yields

$$F^{-1}(u) = \sum_{i=1}^{N} \lambda_i F_i^{-1}(u). \tag{6.6}$$

If ν_1 is atomless, this common-quantile coupling can equivalently be written as $\nu = T_{\#}\nu_1$, with

$$T = \sum_{i=1}^{N} \lambda_i T_{1i} \qquad \text{and} \qquad (T_{1i})_{\#}\nu_1 = \nu_i.$$

For equally weighted empirical measures with the same number of atoms, (6.6) averages atoms of **equal rank**, as illustrated in Figure 13.

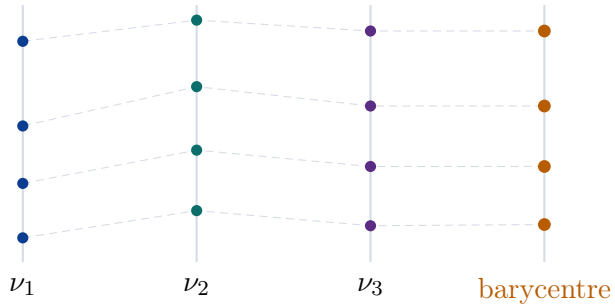

Figure 13

When $N = 2$, write the inputs as ν_0 and ν_1. If ν_0 is absolutely continuous, the barycentre with weights $(1-t, t)$ is precisely the displacement interpolation

$$\nu_t = ((1-t)\,\mathrm{Id} + tT)_{\#}\nu_0.$$

For centered Gaussian inputs with positive-definite covariance matrices $\nu_i = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, the barycentre is Gaussian, $\nu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\sum_{i=1}^{N} \lambda_i (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}^{1/2})^{1/2} = \boldsymbol{\Sigma}. \tag{6.7}$$

This matrix equation is the covariance analogue of averaging aligned quantiles.

In the discrete setting, an **entropic barycentre** solves

$$\min_{\boldsymbol{a}} \sum_{i=1}^{N} \lambda_i L^{\varepsilon}(\boldsymbol{a}, \boldsymbol{b}_i). \tag{6.8}$$

For $\varepsilon > 0$, L^{ε} is convex and differentiable on the interior of the product of simplices. Its gradient with respect to the marginals is given by the **optimal dual potentials** $(\boldsymbol{f}, \boldsymbol{g})$, where

$$L^{\varepsilon}(\boldsymbol{a}, \boldsymbol{b}) = \max_{\boldsymbol{f}, \boldsymbol{g}} \left\{ \langle \boldsymbol{f}, \boldsymbol{a} \rangle + \langle \boldsymbol{g}, \boldsymbol{b} \rangle - \varepsilon \left\langle \exp(\boldsymbol{f}/\varepsilon), \mathbf{K} \exp(\boldsymbol{g}/\varepsilon) \right\rangle + \varepsilon \right\}.$$

The **Sinkhorn scalings** satisfy

$$\boldsymbol{u} = \exp(\boldsymbol{f}/\varepsilon) \qquad \text{and} \qquad \boldsymbol{v} = \exp(\boldsymbol{g}/\varepsilon).$$

At a solution of (6.8), the **first order condition** is $\sum_i \lambda_i \boldsymbol{f}_i = \mathbf{0}$ up to the common **additive gauge**. This leads to both gradient methods and **generalized Sinkhorn iterations**. In imaging applications, the barycentre blends corresponding shapes and colours rather than averaging pixels at fixed locations.

The population version also has a law of large numbers. Let $\widetilde{\mathbb{P}}$ be a probability distribution on $\mathcal{P}_2(X)$ and suppose the **population barycentre**

$$\nu^* = \operatorname*{argmin}_{\nu} \int W_2(\mu, \nu)^2 \, \mathrm{d}\widetilde{\mathbb{P}}(\mu)$$

is unique.

Theorem 6.5 If $\mu_1, \dots, \mu_N$ are independent samples from $\widetilde{\mathbb{P}}$ and ν_N is their **empirical Wasserstein barycentre**, then, under the required moment and compactness assumptions,

$$W_2(\nu_N, \nu^*) \longrightarrow 0$$

almost surely as $N \to \infty$.

Metric Gradient Flows

For a differentiable function F on Euclidean space, the **gradient flow**

$$\dot{\boldsymbol{x}}(t) = -\nabla F(\boldsymbol{x}(t))$$

is the continuous limit of the **proximal recursion**

$$\boldsymbol{x}^{(\ell+1)} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{x}^{(\ell)} \right\|^2 + \tau F(\boldsymbol{x}) \right\}. \tag{6.9}$$

On a Riemannian manifold, the squared norm is replaced by squared Riemannian distance. This observation makes sense even in a metric space (X, d): define a candidate flow as the small-τ limit of

$$x^{(\ell+1)} = \operatorname*{argmin}_{x \in X} \left\{ \frac{1}{2} d(x, x^{(\ell)})^2 + \tau F(x) \right\}. \tag{6.10}$$

This is the **minimizing movement construction**.

The heat equation gives two complementary examples. First consider the **Dirichlet energy**

$$F(u) = \frac{1}{2}\int_{\mathbb{R}^n} \|\nabla u\|^2 \,\mathrm{d}\boldsymbol{x}.$$

For a smooth compactly supported variation v,

$$\begin{aligned} F(u+hv) &= \frac{1}{2}\int \|\nabla u + h\nabla v\|^2 \,\mathrm{d}\boldsymbol{x} \\ &= F(u) + h\int \nabla u \cdot \nabla v \,\mathrm{d}\boldsymbol{x} + O(h^2) \\ &= F(u) - h\int (\Delta u) v \,\mathrm{d}\boldsymbol{x} + O(h^2). \end{aligned}$$

With the L^2 inner product, $\operatorname{grad} F(u) = -\Delta u$. Hence

$$\partial_t u = \Delta u$$

is the $\boldsymbol{L^2}$ **gradient flow** of Dirichlet energy.

The same heat equation is the **Wasserstein gradient flow** of entropy. Let F be a smooth functional of densities, and denote its **first variation** by $\delta F/\delta\rho$. A tangent in Otto's metric has the form

$$\sigma = -\nabla\cdot(\rho\nabla\varphi).$$

Along a curve with this tangent,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}F(\rho_t)\right|_{t=0} = \int \frac{\delta F}{\delta\rho}\sigma \,\mathrm{d}\boldsymbol{x} = \int \nabla\left(\frac{\delta F}{\delta\rho}\right)\cdot\nabla\varphi \,\mathrm{d}\rho.$$

Comparing with (4.1) identifies

$$\operatorname{grad}_W F(\rho) = -\nabla\cdot\left(\rho\nabla\frac{\delta F}{\delta\rho}\right). \tag{6.11}$$

Therefore the negative gradient flow is

$$\partial_t\rho = \nabla\cdot\left(\rho\nabla\frac{\delta F}{\delta\rho}\right). \tag{6.12}$$

For negative entropy,

$$F(\rho) = \int \rho\log\rho \,\mathrm{d}\boldsymbol{x},$$

the first variation is $\log \rho + 1$. The additive constant disappears after taking a spatial gradient, so

$$\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right) = \nabla \cdot (\rho \nabla \log \rho) = \Delta \rho.$$

Equation (6.12) is again the heat equation.

More generally, the Fokker–Planck equation

$$\partial_t \rho = \nabla \cdot (\rho \nabla \Psi) + \beta^{-1} \Delta \rho \tag{6.13}$$

is the Wasserstein gradient flow of the **free energy**

$$F(\rho) = \int \Psi \rho \, \mathrm{d}\boldsymbol{x} + \beta^{-1} \int \rho \log \rho \, \mathrm{d}\boldsymbol{x}.$$

It is also the forward equation of the **Langevin diffusion**

$$\mathrm{d}\boldsymbol{X}_t = -\nabla \Psi(\boldsymbol{X}_t) \, \mathrm{d}t + \sqrt{2\beta^{-1}} \, \mathrm{d}\boldsymbol{W}_t.$$

When the normalizing integral is finite, the **stationary density** is proportional to $\mathrm{e}^{-\beta \Psi}$.

For $m > 1$, another example is the **porous medium equation with drift**,

$$\partial_t \rho - \Delta \rho^m - \nabla \cdot (\rho \nabla V) = 0,$$

whose energy is

$$F(\rho) = \frac{1}{m-1} \int \rho^m \, \mathrm{d}\boldsymbol{x} + \int V(\boldsymbol{x}) \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

The rigorous theory is developed by Ambrosio, Gigli, and Savaré (2008). The entropy interpretation of the Fokker–Planck equation originates with Jordan, Kinderlehrer, and Otto (1998).

Particle Discretizations

The minimizing movement scheme (6.10) for the Wasserstein metric becomes

$$\rho^{(\ell+1)} = \operatorname*{argmin}_{\rho} \left\{ \frac{1}{2} W_2(\rho, \rho^{(\ell)})^2 + \tau F(\rho) \right\}.$$

One numerical approximation represents the density by a **moving cloud**

$$\rho_t \approx \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{x}_i(t)}.$$

For sufficiently small time steps, the identity matching remains optimal, and

$$W_2(\rho_{t+\Delta t}, \rho_t)^2 = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i(t+\Delta t) - \boldsymbol{x}_i(t)\|^2.$$

Thus the measure-valued flow becomes an ordinary Euclidean gradient flow for the particle vector

$$\boldsymbol{X}(t) = (\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_n(t)) \qquad \text{and} \qquad \dot{\boldsymbol{X}}(t) = -\nabla \widetilde{F}(\boldsymbol{X}(t)).$$

For the Langevin example, the particles initially spread and then settle into the **Gibbs density**. The approximation converts diffusion into a deterministic interaction among particle locations through the **discretized entropy**.

The energy landscape in Figure 14 shows how each proximal step balances a short Wasserstein move against a decrease in free energy.

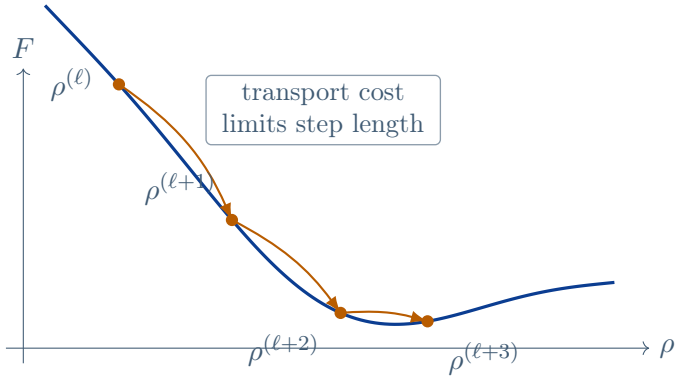

Figure 14

Minimum Kantorovich Estimation and Generative Models

Let $x_1, \dots, x_n$ be independent observations from an unknown distribution, and let

$$\beta_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

be the empirical measure. A **parametric model** is a map $\boldsymbol{\theta} \mapsto \alpha_{\boldsymbol{\theta}} \in \mathcal{P}(X)$. A general **minimum-distance estimator** solves

$$\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\alpha_{\boldsymbol{\theta}}, \beta_n). \tag{6.14}$$

If $\alpha_{\boldsymbol{\theta}}$ has density $p_{\boldsymbol{\theta}}$ and

$$\mathcal{L}(\alpha_{\boldsymbol{\theta}}, \beta_n) = -\int \log p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\beta_n(x) = -\frac{1}{n}\sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(x_i),$$

then (6.14) is maximum likelihood estimation. At the population level, if the true law β and the model law have densities p and $p_{\boldsymbol{\theta}}$ with respect to a common dominating measure, then

$$\mathrm{KL}(\beta \mid \alpha_{\boldsymbol{\theta}}) = \int p(x) \log\left(\frac{p(x)}{p_{\boldsymbol{\theta}}(x)}\right) \mathrm{d}x.$$

The maximum likelihood loss differs from this relative entropy by a term independent of $\boldsymbol{\theta}$.

Relative entropy becomes infinite unless the data law is absolutely continuous with respect to the model law. This is a serious problem when both distributions are concentrated on low dimensional sets. In high dimensional data analysis, a distribution may lie on an unknown manifold M with $\dim M \ll d$; two nearby model manifolds can then have disjoint supports even when they are geometrically close. Moreover, the model density may be computationally inaccessible.

A generative model avoids explicit densities. Let ζ be a simple distribution on a low dimensional latent space Z, and let $h_{\boldsymbol{\theta}} \colon Z \to X$ be a parameterized map (often a neural network). The model is

$$\alpha_{\boldsymbol{\theta}} = (h_{\boldsymbol{\theta}})_{\#}\zeta. \tag{6.15}$$

Samples are easy to generate even if a density for (6.15) is unavailable. A **weak transport discrepancy** is therefore a natural loss.

The standard comparison uses **total variation**, relative entropy, **Jensen–Shannon divergence**, and Wasserstein distance:

$$\begin{aligned}
\delta(\mathbb{P}_r, \mathbb{P}_g) &= \sup_A |\mathbb{P}_r(A) - \mathbb{P}_g(A)|, \\
\mathrm{KL}(\mathbb{P}_r \mid \mathbb{P}_g) &= \int \log\left(\frac{\mathrm{d}\mathbb{P}_r}{\mathrm{d}\mathbb{P}_g}\right) \mathrm{d}\mathbb{P}_r, \\
JS(\mathbb{P}_r, \mathbb{P}_g) &= \frac{1}{2}\mathrm{KL}(\mathbb{P}_r \mid \mathbb{P}_m) + \frac{1}{2}\mathrm{KL}(\mathbb{P}_g \mid \mathbb{P}_m) \qquad \text{and} \qquad \mathbb{P}_m = \frac{1}{2}(\mathbb{P}_r + \mathbb{P}_g).
\end{aligned}$$

Example 6.6 Let Z be uniform on $[0,1]$. Let $\mathbb{P}_0$ be the law of $(0, Z) \in \mathbb{R}^2$, and let $\mathbb{P}_\theta$ be the law of (θ, Z). Then

$$W_1(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|,$$

whereas

$$JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2, & \theta \neq 0, \\ 0, & \theta = 0, \end{cases} \qquad \text{and} \qquad \mathrm{KL}(\mathbb{P}_0 \mid \mathbb{P}_\theta) = \begin{cases} \infty, & \theta \neq 0, \\ 0, & \theta = 0, \end{cases}$$

and

$$\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1, & \theta \neq 0, \\ 0, & \theta = 0. \end{cases}$$

The geometry in Figure 15 is simple: Wasserstein distance records the horizontal displacement, while divergences that are sensitive to the support jump as soon as the lines separate.

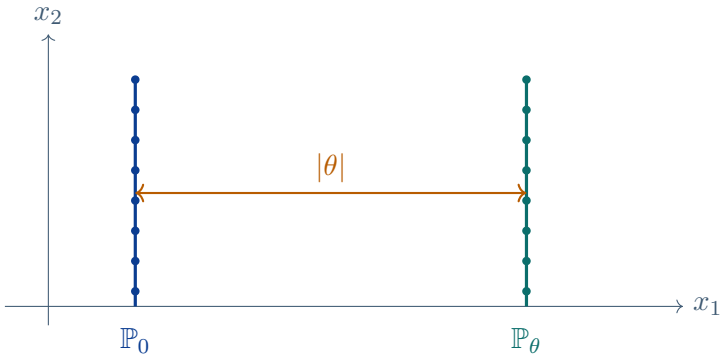

Figure 15

Theorem 6.7 Let $\mathbb{P}_r$ be fixed, let Z have a fixed latent distribution, and let $g\colon Z \times \mathbb{R}^d \to X$ define $\mathbb{P}_\theta$ as the law of $g(Z, \boldsymbol{\theta})$.

1. If g is continuous in $\boldsymbol{\theta}$ and the required moments are locally dominated, then $W_p(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous in $\boldsymbol{\theta}$.
2. If g is locally Lipschitz and satisfies the corresponding regularity condition, then $W_p(\mathbb{P}_r, \mathbb{P}_\theta)$ is continuous everywhere and differentiable almost everywhere.
3. The analogous assertions fail in general for Jensen–Shannon divergence and both orientations of relative entropy.

The theorem, together with Example 6.6, explains why **Wasserstein loss** can support **training by gradient methods** even when model and data distributions occupy different low dimensional sets.

For the cost $c(x, y) = d(x, y)$, c-concave potentials are exactly 1-Lipschitz functions, and $f^c = -f$. Kantorovich duality therefore becomes the Kantorovich–Rubinstein formula

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{ \int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu \right\}. \tag{6.16}$$

For a generative model $\mathbb{P}_{\boldsymbol{\theta}} = (g_{\boldsymbol{\theta}})_{\#}\zeta$,

$$W_1(\mathbb{P}_r, \mathbb{P}_{\boldsymbol{\theta}}) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{ \mathbb{E}_{\mathbb{P}_r}[f(X)] - \mathbb{E}_{\zeta}[f(g_{\boldsymbol{\theta}}(Z))] \right\}.$$

The original **Wasserstein generative adversarial network** parameterizes f by a **critic network** $f_{\boldsymbol{w}}$ and solves the saddle problem

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{w}} \left\{ \mathbb{E}_{\mathbb{P}_r}[f_{\boldsymbol{w}}(X)] - \mathbb{E}_{\zeta}[f_{\boldsymbol{w}}(g_{\boldsymbol{\theta}}(Z))] \right\}, \tag{6.17}$$

subject to an approximate Lipschitz constraint on the critic. This formulation was introduced by Arjovsky, Chintala, and Bottou (2017).

Appendix: Lecture and Tutorial Index