November 11, 2022

Lecture 3: Optimal Transport with the Monge Cost

Minerva mini-course on Convexity in High dimensions by Bo'az Klartag

Last week we discussed convex localization which is based on hyperplane bisections. The basic property was that if we bisect a convex body in $\mathbb{R}^n$ using hyperplane, we obtain two convex pieces. We can repeat this procedure in the sphere S^n or in the hyperbolic plane $\mathbb{H}^n$, because if we consider a convex set in one of these constant-curvature spaces, and bisect it with respect to a hyperplane, then we obtain two convex pieces. The proof of the Gaussian waist inequality from last week can be adapted to yield the waist of the sphere theorem:

Theorem 1 (Gromov '02)**.** *Let $1 \leq k \leq n$ and let $f : S^n \to \mathbb{R}^k$ be a continuous function. Then there exists $t \in \mathbb{R}^k$ such that the fiber $L = f^{-1}(t)$ satisfies*

$$\sigma_n(L + r) \geq \gamma_n(S^{n-k} + r) \qquad \textit{for all } r > 0.$$

where σ is the surface area measure on S^n and $L + r$ is the r-neighborhood of L.

The convex localization process begins with a convex body $K \subseteq \mathbb{R}^n$, we bisect again and again, and after N steps we obtain a partition of K into 2^N convex bodies $K_1, \ldots, K_{2^N}$ which are approximate 1-dimensional needles or approximate ℓ-dimensional pancakes. There is much freedom in constructing the partition, and one uses these degrees of freedom in order to impose conditions such as

$$\mathcal{I}(K_1) = \ldots = \mathcal{I}(K_{2^N})$$

where $\mathcal{I} : \{\text{c}\mathit{onvex\ bodies}\} \to \mathbb{R}^\ell$ is a continuous functional of our choice.

Let us have another look at the convex localization process in a particular case, the *additive* one-dimensional case of convex localization, where the topological part of the argument is much easier, and the recursive nature of the process is evident. Given a convex body $K \subseteq \mathbb{R}^n$, a log-concave measure μ on K, and a function $f : K \to \mathbb{R}$ of integral zero, and we recursively obtain a partition of K into 2^ℓ convex bodies $K_1, \ldots, K_{2^\ell}$ which are approximate needles with

$$\int_{K_i} f d\mu = 0 \qquad \text{for } i = 1, \ldots, 2^\ell.$$

This is a sequence of more and more refined partitions. Is there a limiting object when N tends to infinity? Yes. There is a little bit of measure-theory involved (see Alesker '98), but it is possible to take the limit $N \to \infty$ and obtain the following:

Limit Object (Needle Decomposition):

1. A partition $\{K_\omega\}_{\omega \in \Omega}$ of K into segments (a.k.a "needles". Some of them may be singletons, or full lines, or rays).

2. A measurable partition induces **disintegration of measure** or conditional probabilities. These are measures $\{\mu_\omega\}_{\omega\in\Omega}$ on K, and ν on Ω, with
$$\mu = \int_\Omega \mu_\omega d\nu(\omega)$$

3. ν-almost every μ_ω is supported on K_ω with $\displaystyle\int_{K_\omega} f d\mu_\omega = 0$.

4. ν-almost every μ_ω is log-concave (by Brunn-Minkowski or Prékopa-Leindler).

We recall that while in general it is impossible to condition a probability measure on a *single* zero-measure set, it is typically possible to condition with respect to a *partition* into zero-measure sets. Here are some examples for such needle decomposition. The first example is when $K = [0,1]^2 \subseteq \mathbb{R}^2$ and $f(x,y) = f(x)$ with $\int_K f = 0$. Here there is a partition into segments parallel to the x-axis, and the needles μ_ω are Lebesgue measures,
$$d\mu_\omega(x) = dx.$$
The second example is where $K = B(0,1) \subseteq \mathbb{R}^2$ is the unit disc and $f(x,y) = f(\sqrt{x^2+y^2})$ with $\int_K f = 0$. Here the partition is into radii, and the needles μ_ω satisfy
$$d\mu_\omega(r) = rdr$$
by polar coordinates. This is a log-concave function in the interval $[0,1]$, of course. In general, needle decomposition may be viewed as a generalization of polar coordinates. This story has a close connection to Optimal Transport with the Monge cost.

. . . .

A review of Optimal Transport theory with the Monge cost

Let μ_1 and μ_2 be two measures in $\mathbb{R}^n$, say compactly-supported and absolutely continuous, with the same total mass. We would like to push-forward the measure μ_1 to the measure μ_2 in the most efficient way, that minimizes the average distance that points have to travel. That is, we look at the optimization problem
$$\inf_{S_*(\mu)=\nu} \int_{\mathbb{R}^n} |Sx - x| d\mu_1(x).$$
This is the problem of Optimal Transport with the Monge cost or the L^1 cost, considered by Monge in 1781. Here is a heuristics from Monge's paper that explains why this problem induces a partition into segments.

Monge heuristics: For the optimal transport map T, the segments $(x, T(x))$ $(x \in Supp(\mu_1))$ do not intersect, unless they overlap.

Explanation. Suppose that the segments (x, Tx) and (y, Ty) intersect at a point z, and apply the Triangle Inequality. □

This is related to the following elementary riddle: given 50 red points and 50 blue points in the plane, in general position, find a matching so that the corresponding segments do not intersect.

Since the above argument relies only on the triangle inequality, you would expect that the Optimal Transport problem would induce a partition into geodesics also for Riemannian manifolds, or Finslerian manifolds, or measure metric spaces of some type – basically wherever the triangle inequality holds true (under some regularity assumptions). This is in contrast to the hyperplane bisection method, that applies only in highly symmetric spaces such as $\mathbb{R}^n$, S^n or $\mathbb{H}^n$.

. . . .

Linear programming relaxation and the dual problem (Kantorovich 1940s)

In Monge's problem we minimize over all maps S that push-forward μ_1 to μ_2. There is a relaxation of this problem, that looks at all possible *couplings*, or transport plans, of the two distributions. That is, instead of mapping a point x to a single point Tx, we are allowed to spread the mass across a region. We look at all probability measures γ on $\mathbb{R}^n \times \mathbb{R}^n$ with

$$(\pi_1)_* \gamma = \mu_1 \qquad \text{and} \qquad (\pi_2)_* \gamma = \mu_2.$$

where $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$. Such a probability measure is called a *coupling* of μ and ν. In other words, we now look at *transport plans* rather than *transport maps*. The nice thing is that the space of all couplings is a convex set. The relaxed problem invovles minimizing the average distance that points travel, namely we look at

$$\inf_{(\pi_1)_*\gamma=\mu, (\pi_2)_*\gamma=\nu} \int_{\mathbb{R}^n \times \mathbb{R}^n} |x - y| d\gamma(x, y).$$

Hence we minimize a linear function on a convex set, this is Linear Programming or Functional Analysis.

Theorem 2. *(The dual problem) Let μ_1, μ_2 be two absolutely-continuous probability measures in $\mathbb{R}^n$ (or better, in some geodesically-convex Riemannian manifold M). Assume that for some $x_0 \in \mathbb{R}^n$,*

$$\int_{\mathbb{R}^n} d(x, x_0) d\mu_1(x) < \infty \qquad \textit{and} \qquad \int_{\mathbb{R}^n} d(x, x_0) d\mu_2(x) < \infty.$$

Denote $\mu = \mu_1 - \mu_2$. Then the following quantities are equal:

1. *The minimum over all couplings γ of μ_1 and μ_2 of the integral*

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} d(x, y) d\gamma(x, y).$$

2. *The maximum over all* 1*-Lipschitz functions* $u : \mathbb{R}^n \to \mathbb{R}$ *of*

$$\int_{\mathbb{R}^n} u d\mu$$

3. *The minimum over all maps* T *with* $T_*\mu_1 = \mu_2$ *of*

$$\int_{\mathbb{R}^n} d(x, Tx) d\mu_1(x).$$

Proof sketch. We refer to Ambrosio's lecture notes on Optimal Transport Problems '03 for full details. For the easy direction of the linear programming duality, pick a 1-Lipschitz map u and a coupling γ. For any points $x, y \in \mathbb{R}^n$,

$$u(x) - u(y) \leq d(x, y).$$

Integrating with respect to γ, we get

$$\int_{\mathbb{R}^n} u d\mu = \int_{\mathbb{R}^n \times \mathbb{R}^n} [u(x) - u(y)] d\gamma(x, y) \leq \int_{\mathbb{R}^n \times \mathbb{R}^n} d(x, y) d\gamma(x, y). \tag{1}$$

Hence we need to find u and γ so that equality is attained in (1). The argument goes roughly as follows. A compactness argument shows that the infimum over all couplings is attained. Similarly to the Monge heuristics, one may show that the optimality implies that the support of γ must be cyclically monotone: If $(x_i, y_i) \in Supp(\gamma) \subseteq \mathbb{R}^n \times \mathbb{R}^n$ for $i = 1, \ldots, N$ then for any permutation $\sigma \in S_N$,

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

By an elementary argument similar to Rockafellar's theorem from convex geometry, this condition implies that there exists a 1-Lipschitz function $u : \mathbb{R}^n \to \mathbb{R}$ with

$$(x, y) \in Supp(\gamma) \qquad \Longrightarrow \qquad u(y) - u(x) = d(x, y).$$

[Hint: Fix $u(x_0) = 0$ and define $u(x)$ as the infimum over all constraints]. This way we find u and γ so that equality is attained in (1). The proof that γ can also be replaced by a transport map is due to Evans and Gangbo '98. This relies rom the analysis of the structure of u that will be explain next. □

Remark. The minimizers γ or T are highly not unique, it is actually the 1-Lipschitz function u which is more-or-less determined. More precisely, the gradient ∇u is determined μ-almost everywhere.

A few words about the structure of the optimal 1-Lipschitz function u, and about 1-Lipschitz functions in general. When a 1-Lipschitz function u satisfies $|u(x) - u(y)| = d(x, y)$, it necessarily grows in speed one along the segment from x to y.

Fact: After eliminating a set of measure zero form $\mathbb{R}^n$, the relation

$$x \sim y \qquad \Longleftrightarrow \qquad |u(x) - u(y)| = d(x,y)$$

is an equivalence relation, and the equivalence classes are line segments (or geodesics in the case of a Riemannian manifold) called *transport rays*. Moreover, it is guaranteed that transport rays of positive length cover the entire support of the measure μ, up to a set of measure zero.

Exercises: what are the transport rays of $u(x) = x_1$? and of $u(x) = |x|$?

Understanding the measure disintegration induced by the partition into transport rays requires regularity analysis of the function u, basically one needs to show that the 1-Lipschitz function u is in fact almost C^2 in some sense. This follows from the seminal work of Evans-Gangbo, which was continued by Caffarelli, Feldman and McCann and by myself. Let u be a minimizer as above, with $\mu = \mu_1 - \mu_2$ with the two measures being absolutely-continuous in $\mathbb{R}^n$ and satisfying the above mild integrability condition. Write

$$f = \frac{d\mu}{d\lambda}$$

where λ is the Lebesgue measure on $\mathbb{R}^n$ (or better, we may work with any log-concave reference measure in $\mathbb{R}^n$, not just the Lebesgue measure). Then we require that

$$\int_{\mathbb{R}^n} f d\lambda = 0.$$

The following theorem from K. '17 is analogous to integration in polar coordinates, yet with respect to a general 1-Lipschitz guiding function, rather than just $u(x) = |x|$.

Theorem 3. *There is a partition $\{\mathcal{I}_\omega\}_{\omega \in \Omega}$ of $\mathbb{R}^n$ and measures ν on Ω, and $\{\mu_\omega\}_{\omega \in \Omega}$ on $\mathbb{R}^n$ such that*

1. *Disintegration of measure*

$$\lambda = \int_\Omega \mu_\omega d\nu(\omega).$$

2. *For any $\omega \in \Omega$ the measure μ_ω is supported on the segment*

$$\mathcal{I}_\omega = \{\gamma_\omega(t)\}_{t \in (a_\omega, b_\omega)} \qquad \textit{(arclength parametrization)}$$

with C^∞-smooth, positive density $\rho = \rho_\omega : (a_\omega, b_\omega) \to \mathbb{R}$. The segments are transport rays of the 1-Lipschitz function u.

3. *For any $\omega \in \Omega$,*

$$\int_{\mathcal{I}_\omega} f d\mu_\omega = 0.$$

4. *For any $\omega \in \Omega$, the function ρ is log-concave.*
(In fact, in the case where λ is the Lebesgue measure, it is a polynomial of degree $n-1$ with real roots, which does not vanish in the interval in which ρ is defined).

This recovers the convex localization technique in the one-dimensional linear case, but even in this case there is some advantage here, which is the 1-Lipschitz function u that grows with speed one along the needles.

Remark. Another advantage of this theorem is that it works in any Riemannian manifold with non-negative Ricci curvature. More generally, if we set $\kappa(t) = Ricci(\dot\gamma(t), \dot\gamma(t)), n = \dim(M)$, then we have

$$\left(\rho^{\frac{1}{n-1}}\right)'' + \frac{\kappa}{n-1} \cdot \rho^{\frac{1}{n-1}} \leq 0.$$

The Riemannian version may be used to prove isoperimetric inequalities under lower bounds on the Ricci curvature, Poincaré inequalities, log-Sobolev inequalities, Brunn-Minkowski, and in general it is a rather strong technique for proving geometric inequalities on manifolds.

As an application, let us prove the reverse Cheeger inequality of Buser and Ledoux, and in fact the following refinement due to E. Milman:

Proposition 4. *Let μ be a log-concave probability measure on $\mathbb{R}^n$ and $R > 0$. Assume that for any 1-Lipschitz function $u : \mathbb{R}^n \to \mathbb{R}$ there exists $\alpha \in \mathbb{R}$ with*

$$\int_{\mathbb{R}^n} |u(x) - \alpha| d\mu(x) \leq R. \tag{2}$$

(this is a weaker condition than having a spectral gap of $1/R^2$). Then for any measurable set $S \subseteq \mathbb{R}^n$ and $0 < \varepsilon < R$,

$$\mu(S_\varepsilon \setminus S) \geq c \cdot \frac{\varepsilon}{R} \cdot \mu(S) \cdot (1 - \mu(S)), \tag{3}$$

where $c > 0$ is a universal constant, and where S_ε is the ε-neighborhood of S.

Proof. See K. '17 for more details. Denote $t = \mu(S) \in [0,1]$ and set $f(x) = 1_S(x) - t$ for $x \in \mathbb{R}^n$. Then $\int f d\mu = 0$. Let u be a 1-Lipschitz function maximizing

$$\int_{\mathbb{R}^n} uf d\mu.$$

After adding a constant to u, we may assume that

$$\int_{\mathbb{R}^n} |u| d\mu \leq R.$$

By the theorem, we obtain a needle decomposition: measures $\{\mu_{\mathcal{I}}\}_{\mathcal{I}\in\Omega}$ on $\mathbb{R}^n$, and a measure ν on the space Ω of transport rays which yield a disintegration of measure. We may normalize and assume that all of these measures are probability measures. Hence,

$$\int_\Omega \left(\int_{\mathcal{I}} |u| d\mu_{\mathcal{I}} \right) d\nu(\mathcal{I}) = \int_{\mathbb{R}^n} |u| d\mu \leq R.$$

Denote

$$B = \left\{ \mathcal{I} \in \Omega \, ; \, \int_{\mathcal{I}} |u| d\mu_{\mathcal{I}} \leq 2R \right\}.$$

By the Markov-Chebyshev inequality,

$$\nu(B) \geq 1/2. \tag{4}$$

For all intervals $\mathcal{I} \in \Omega$ we know that $\int_{\mathcal{I}} f d\mu_{\mathcal{I}} = 0$, hence

$$\mu_{\mathcal{I}}(S) = t \cdot \mu_{\mathcal{I}}(\mathbb{R}^n) = t.$$

We would like to prove that for any $\mathcal{I} \in B$ and any $0 < \varepsilon < R$,

$$\mu_{\mathcal{I}}(S_\varepsilon \setminus S) \geq c \cdot \frac{\varepsilon}{R} \cdot t(1-t), \tag{5}$$

for a universal constant $c > 0$. Once (5) is proven, the bound (3) follows by integrating (5) with respect to ν and using (4).

What remains to be proven is a one-dimensional statement about log-concave measures: If $\nu = \mu_{\mathcal{I}}$ is a log-concave probability measure on $\mathbb{R}$ with $\int_{\mathbb{R}} |t| d\nu(t) \leq R$, then (5) holds true. It suffices to prove this under the additional assumption that S is connected (Bobkov '96), and in fact a half line. This one-dimensional statement is proven in Bobkov '96 (see K. '17). $\square$

The same proof applies for any complete Riemannian manifold with non-negative Riemannian curvature. No need for completeness, the weaker geodesic-convexity assumption suffices.