Lecture 6: Concentration via Optimal Transport

Tianpei Xie

Jan. 24th., 2023

Contents

1 Optimal Transport Basis

1.1 Optimal Transport Problem and its Dual Problem

- **Definition** (***Pushforward Measure***) [Peyr and Cuturi, 2019]
  Let $(\mathcal{X}, \mathscr{B}_X)$ and $(\mathcal{Y}, \mathscr{B}_Y)$ be two topological measurable spaces. Denote the spaces of *general (Radon) measures* on $\mathcal{X}, \mathcal{Y}$ as $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathcal{Y})$. Also let $\mathcal{C}(\mathcal{X})$ be space of continuous functions on $\mathcal{X}$. For a *continous* map $T : \mathcal{X} \to \mathcal{Y}$, the ***push-forward operator*** is defined as $T_{\#} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ that satisfies

$$\forall\, h \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{Y}} h(y)\; d\left(T_{\#}\alpha\right)(y) = \int_{\mathcal{X}} h(T(x))\; d\alpha(x). \tag{1}$$

$$\text{or equivalently,} \quad \left(T_{\#}\alpha\right)(B) := \alpha\left(\{x : T(x) \in B \subset \mathcal{Y}\}\right) = \alpha(T^{-1}(B)) \tag{2}$$

  where the ***push-forward measure*** $\beta := T_{\#}\alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$, $T^{-1}(\cdot)$ is the pre-image of T.

- **Remark** (***Density Function of Pushforward Measure***)
  Assume that (α, β) have densities $(\rho_\alpha, \rho_\beta)$ with respect to a fixed measure, and $\beta = T_{\#}\alpha$. We see that $T_{\#}$ acts on a density ρ_α linearly to a density ρ_β as a change of variable, i.e.

$$\rho_\alpha(\boldsymbol{x}) = \left|\det(T'(\boldsymbol{x}))\right| \rho_\beta(T(\boldsymbol{x})) \tag{3}$$

$$\left|\det(T'(\boldsymbol{x}))\right| = \frac{\rho_\alpha(\boldsymbol{x})}{\rho_\beta(T(\boldsymbol{x}))}$$

- **Definition** (***Optimal Transport Problem, Monge Problem***) [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]
  Let $(\mathcal{X}, \mathscr{B}_X)$ and $(\mathcal{Y}, \mathscr{B}_Y)$ be two measurable spaces, where $\mathcal{X}$ and $\mathcal{Y}$ are *complete separable metric spaces*. Denote $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ as the space of probabiilty measures on $\mathcal{X}$ and $\mathcal{Y}$. Define a ***cost function*** $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ as non-negative real-valued measurable functions on $\mathcal{X} \times \mathcal{Y}$. ***The optimal transport problem*** by *Monge* (i.e. ***Monge Problem***) is defined as follows: given two probability measures $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$, find a *continuous measurable map* $T : \mathcal{X} \to \mathcal{Y}$ so that

$$\begin{aligned} &\inf_{T} \int_{\mathcal{X}} c(x, T(x)) d\mathbb{P}(x) \\ &\text{s.t. } \mathbb{Q} = T_{\#}\mathbb{P} \end{aligned}$$

  The optimal solution T is also called an ***optimal transportation plan***.

- **Definition** (***Optimal Transport Problem, Kantorovich Relaxation***) [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]
  The optimal transport problem by *Kantorovich* (i.e. ***Kantorovich Relxation***) is defined as follows: given two probability measures $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$, find a *joint probability measure* $\gamma \in \Pi(\mathbb{P}, \mathbb{Q})$ so that

$$\begin{aligned} &\inf_{\gamma} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y) \\ &\text{s.t. } \gamma \in \Pi(\mathbb{P}, \mathbb{Q}) := \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi_{\mathcal{X},\#}\gamma = \mathbb{P},\ \pi_{\mathcal{Y},\#}\gamma = \mathbb{Q}\} \end{aligned}$$

where $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is the space of joint probability measure on $\mathcal{X} \times \mathcal{Y}$, $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ are the coordinate projection onto $\mathcal{X}$ and $\mathcal{Y}$. $\pi_{\mathcal{X},\#}\gamma = \mathbb{P}$ means that $\mathbb{P}$ is the marginal distribution of γ on $\mathcal{X}$. Similarly $\mathbb{Q}$ is the marginal distribution of γ on $\mathcal{Y}$.

Equivalently, let X and Y are *random variables* taking values in $\mathcal{X}$ and $\mathcal{Y}$. The *joint distribution* of (X, Y) is γ with marginal distribution of X and Y being $\mathbb{P}$ and $\mathbb{Q}$. Then the problem is

$$\min_{\gamma \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{\gamma}\left[c(X,Y)\right]$$

The joint distribution $\gamma \in \Pi(\mathbb{P},\mathbb{Q})$ such that $X_{\#}\gamma = \mathbb{P}$ and $Y_{\#}\gamma = \mathbb{Q}$ is called ***a coupling***.

- **Proposition 1.1** *(**Existance of Solution**) [Santambrogio, 2015]*
  *Let $\mathcal{X}, \mathcal{Y}$ be **complete separable spaces**, $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{+}$ be **lower semi-continuous function**. Then the Kantorovich relaxation of optimal transport problem **admits a solution**.*

- **Definition** (***Dual Problem of Kantorovich Problem***) [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]
  The ***dual problem*** of *Kantorovich problem* is described as below:

$$\mathcal{L}_c(\mathbb{P},\mathbb{Q}) = \max_{(\varphi,\psi)\in \mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \psi(y) d\mathbb{Q}(y)$$
$$\text{s.t. } \varphi(x) + \psi(y) \le c(x,y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

  Here, (φ, ψ) is a pair of *continuous functions* on $\mathcal{X}$ and $\mathcal{Y}$ respectively and they are also the ***Kantorovich potentials***. The feasible region is

$$\mathcal{R}(c) := \{(\varphi,\psi) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \varphi \oplus \psi \le c\}$$

  where $(\varphi \oplus \psi)(x, y) = \varphi(x) + \psi(y)$.

  In other words, the dual optimization problem is

$$\max_{(\varphi,\psi)\in\mathcal{R}(c)} \mathbb{E}_{\mathbb{P}}\left[\varphi(X)\right] + \mathbb{E}_{\mathbb{Q}}\left[\psi(Y)\right]$$

- **Proposition 1.2** *(**Strong Duality**) [Santambrogio, 2015]*
  *Let $\mathcal{X}, \mathcal{Y}$ be **complete separable spaces**, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{+}$ be **lower semi-continuous and bounded from below**. Then the optimal value of* primal *and* dual problems *are the same*

$$\min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[c(X,Y)\right] = \mathcal{L}_c(\mathbb{P},\mathbb{Q}) = \max_{(\varphi,\psi)\in\mathcal{R}(c)} \mathbb{E}_{\mathbb{P}}\left[\varphi(X)\right] + \mathbb{E}_{\mathbb{Q}}\left[\psi(Y)\right].$$

1.2 Wasserstein Distance

- **Definition** (***Wasserstein Distance***)
  Let $((\mathcal{X}, d), \mathscr{B})$ be *a metric measurable space* with *Borel σ-algebra* induced by metric d. Let X, Y be two random variables taking values in $\mathcal{X}$ with distribution $\mathbb{P}$ and $\mathbb{Q}$. ***The d-Wasserstein distance*** between *probability distributions* $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathcal{W}_d(\mathbb{P},\mathbb{Q}) := \min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[d(X,Y)\right] \tag{4}$$

- **Remark** (d-***Wasserstein Distance is a Metric in*** $\mathcal{P}(\mathcal{X})$)
  The ***d-Wasserstein distance*** $\mathcal{W}_d(\mathbb{P},\mathbb{Q}) := \min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[d(X,Y)\right]$ is a well-defined *metric* in $\mathcal{P}(\mathcal{X})$: for all $\mathbb{P},\mathbb{Q},\mathbb{M} \in \mathcal{P}(\mathcal{X})$,

  1. (*Non-Negativity*): $W_d(\mathbb{P},\mathbb{Q}) \geq 0$.
  2. (*Definiteness*): $W_d(\mathbb{P},\mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
  3. (*Symmetric*): $W_d(\mathbb{P},\mathbb{Q}) = W_d(\mathbb{Q},\mathbb{P})$
  4. (*Triangular inequality*): $W_d(\mathbb{P},\mathbb{Q}) \leq W_d(\mathbb{P},\mathbb{M}) + W_d(\mathbb{M},\mathbb{Q})$

- **Remark** *The Wasserstein distance, or Optimal Transport (OT)*, $\mathcal{W}_d(\alpha,\beta)$ depends on the *distance definition* d on the base measurable space $\mathcal{X}$. In other word, OT can be seen as automatically "***lifting***" a ground metric d in $\mathcal{X}$ to a *metric* between ***measures*** on $\mathcal{X}$

- **Remark** (***Convergence in Wasserstein Space*** $\Leftrightarrow$ ***Weak Convergence***) [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]
  One of most ***important*** propery of *Wasserstein distance* is that it is a ***weak distance***, i.e. it allows one to compare singular distributions (for instance, discrete ones) whose **supports *do not overlap*** and to quantify the spatial shift between the supports of two distributions.

  In fact, $\mathcal{W}_p$ is a way to quantify the ***weak* convergence*** or ***convergence in distribution*** *(in law)* [Villani, 2009]:

  Definition On a compact domain $\mathcal{X}$, $(\alpha_k)_k$ converges ***weakly*** to α in $\mathcal{M}^1_+(\mathcal{X})$ (denoted $\alpha_n \stackrel{d}{\to} \alpha$) if and only if *for any* ***continuous*** *function* $g \in \mathcal{C}(\mathcal{X})$, $\int_{\mathcal{X}} g d\alpha_k \to \int_{\mathcal{X}} g d\alpha$. One needs to add additional decay conditions on g on noncompact domains.

  This notion of weak convergence corresponds to the **convergence in the distribution** of random vectors. Note the any random variable X_n is a continous function on Ω, and its distribution is the push-forward measure $\alpha_n = X_{n\#}\mathbb{P}$. Therefore, $\alpha_n \rightharpoonup \alpha$ is equivalent to $X_n \stackrel{d}{\to} X$. This convergence can be shown (see [Villani, 2009, Santambrogio, 2015]) to be equivalent to

  $$\alpha_n \rightharpoonup \alpha \;\;\Leftrightarrow\;\; \mathcal{W}_d(\alpha_n,\alpha) \to 0.$$

  Thus we can also write the weak convergance as $\alpha_n \xrightarrow{\mathcal{W}_d} \alpha$.

1.3 Dual Formulation of Wasserstein Distance

- **Theorem 1.3** ***(Kantorovich-Rubenstein duality)*** *[Villani, 2009]*
  Let $\mathcal{X}$ be a ***Polish space****, i.e. $\mathcal{X}$ a complete separable metric space equipped with a Borel σ-algebra induced by metric d, and $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $\mathcal{X}$. Let Lip_1 be the space of all 1-**Lipschitz** function with respect to metric d such that*

  $$\|f\|_L := \sup_{x,y\in\mathcal{X}} \left\{\frac{|f(x)-f(y)|}{d(x,y)}\right\} \leq 1.$$

  Then

  $$\mathcal{W}_d(\mathbb{P},\mathbb{Q}) = \sup_{f\in Lip_1} \left\{\mathbb{E}_{\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{\mathbb{Q}}\left[f(Y)\right]\right\}. \tag{5}$$

- **Example** (***Total Variation as $\mathcal{W}_d$ with respect to Hamming distance*** d_H)
  When $d(x,y) = \sum_i \mathbb{1}\{x_i \neq y_i\} = d_H(x,y)$ Hamming distance, the $\mathcal{W}_d$ becomes

$$\mathcal{W}_{d_H}(\mathbb{P},\mathbb{Q}) = \sup_{f:\mathcal{X}\to[0,1]} \int_{\mathcal{X}} f\,(d\mathbb{P} - d\mathbb{Q}) = \sup_{A\subset\mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)| := \|\mathbb{P} - \mathbb{Q}\|_{TV}$$

- **Example** ($\mathcal{W}_1$ ***with respect to*** L_1 ***Norm***)
  When $d(x,y) = |x-y|$ in $\mathbb{R}$, and F_α, F_β are cumulative distribution function of α, β, then $\mathcal{W}_1$ distance becomes

$$\begin{aligned}\mathcal{W}_1(\alpha,\beta) = \|F_\alpha - F_\beta\|_1 &:= \int_{-\infty}^{\infty} \|F_\alpha(x) - F_\beta(x)\|_1\,dx \\ &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{x} d(\alpha - \beta) \right|\end{aligned}$$

  which shows that $\mathcal{W}_1$ on $\mathbb{R}$ is a **norm**. An optimal Monge map T such that $T_\#\alpha = \beta$ is then defined by

$$T = F_\beta^{-1} \circ F_\alpha$$

  where $F_\beta^{-1} = \inf\{t : F_\beta \geq t\}$.

2 The Transportation Method

2.1 Concentration via Transportation Cost Inequality

- **Remark** (***Equivalence of Transportation Cost Inequality and Sub-Gaussian***) [Boucheron et al., 2013]
  Let X be a real-valued integrable random variable. Let ϕ be a ***convex*** and ***continuously differentiable*** function on a (possibly unbounded) interval $[0,b)$ and assume that $\phi(0) = \phi'(0) = 0$. Define, for every $x \geq 0$, ***the Legendre transform*** $\phi^*(x) = \sup_{\lambda\in(0,b)}(\lambda x - \phi(\lambda))$, and let, for every $t \geq 0$, $\phi^{*-1}(t) = \inf\{x \geq 0 : \phi^*(x) > t\}$, i.e. the ***the generalized inverse*** of ϕ^*. Then the following two statements are equivalent:

  1. for every $\lambda \in (0,b)$,

$$\psi_{X-\mathbb{E}[X]}(\lambda) \leq \phi(\lambda)$$

     where $\psi_X(\lambda) := \log \mathbb{E}_Q\left[e^{\lambda X}\right]$ is the logarithm of moment generating function;

  2. for any probability measure P absolutely continuous with respect to Q such that $\mathbb{KL}(P \,\|\, Q) < \infty$,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \phi^{*-1}\left(\mathbb{KL}(P\,\|\,Q)\right). \tag{6}$$

  In particular, given $\nu > 0$, X follows a ***sub-Gaussian distribution***, i.e.

$$\psi_{X-\mathbb{E}[X]}(\lambda) \leq \frac{\nu\lambda^2}{2}$$

for every $\lambda > 0$ ***if and only if*** for any probability measure P absolutely continuous with respect to Q and such that $\mathbb{KL}(P \,\|\, Q) < \infty$,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \le \sqrt{2\nu \mathbb{KL}(P \,\|\, Q)}. \tag{7}$$

- **Definition** (d-***Transportation Cost Inequality***) [Wainwright, 2019]
  Let $(\mathcal{X}, d)$ be a *metric space* with metric d, and $(\mathcal{X}, \mathscr{B})$ be a *measurable space*, where $\mathscr{B}$ is *the Borel σ-algebra* induced by metric d, ***the probability measure*** $\mathbb{P}$ is said to satisfy a <u>d-***transportation cost inequality***</u> with parameter $\nu > 0$ if

  $$\mathbb{E}_{\mathbb{Q}}[X] - \mathbb{E}_{\mathbb{P}}[X] \le \sqrt{2\nu \mathbb{KL}(\mathbb{Q} \,\|\, \mathbb{P})} \tag{8}$$

  for all probability measure $\mathbb{Q} \ll \mathbb{P}$ on $\mathscr{B}$.

- **Theorem 2.1** ***(Isoperimetric Inequality via Transportation Cost)****[Wainwright, 2019]*
  Consider a metric measure space $(\mathcal{X}, \mathscr{B}, \mathbb{P})$ with metric d, and suppose that $\mathbb{P}$ satisfies the d-***transportation cost inequality***

  $$\mathbb{E}_{\mathbb{Q}}[X] - \mathbb{E}_{\mathbb{P}}[X] \le \sqrt{2\nu \mathbb{KL}(\mathbb{Q} \,\|\, \mathbb{P})}$$

  for all probability measure $\mathbb{Q} \ll \mathbb{P}$ on $\mathscr{B}$. Then its ***concentration function*** *satisfies the bound*

  $$\alpha_{\mathbb{P},(\mathcal{X},d)}(t) \le 2\exp\left(-\frac{t^2}{2\nu}\right) \tag{9}$$

  Moreover, for any $Z \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the ***concentration inequality***

  $$\mathbb{P}\left\{|f(Z) - \mathbb{E}[f(Z)]| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2\nu L^2}\right). \tag{10}$$

2.2 Tensorization for Transportation Cost

2.3 Bounded Difference Inequality via Transportation Methods

2.4 Conditional Transportation Inequality

2.5 Convex Distance Inequality via Conditional Transportation Cost

2.6 Talagrand's Gaussian Transportation Inequality

2.7 Transportation Cost Inequalities for Markov Chains

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.