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:

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

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 Wasserstein distance*** between *probability distributions* $\mathbb{P}$ and $\mathbb{Q}$ induced by d is defined as

$$\mathcal{W}_1(\mathbb{P},\mathbb{Q}) \equiv \mathcal{W}_d(\mathbb{P},\mathbb{Q}) := \min_{X\sim\mathbb{P}, Y\sim\mathbb{Q}} \mathbb{E}\left[d(X,Y)\right] \tag{4}$$

In general, for $p \in [1, \infty)$, we can define ***Wasserstein*** p***-distance*** as

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := \left(\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}\left[(d(X,Y))^p \right] \right)^{1/p}. \tag{5}$$

- **Remark** Not to confuse the 2***-Wasserstein distance*** with ***the Wasserstein distance induced by*** L_2 ***norm***:

$$\mathcal{W}_{\|\cdot\|_2}(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{1,\|\cdot\|_2}(\mathbb{P}, \mathbb{Q}) := \min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}\left[\|X - Y\|_2 \right]$$
$$\mathcal{W}_2(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{2,d}(\mathbb{P}, \mathbb{Q}) := \sqrt{\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}\left[d(X,Y)^2 \right]}$$

- **Remark** (***Wasserstein*** p***-Distance is a Metric in*** $\mathcal{P}(\mathcal{X})$)
The ***Wasserstein*** p***-distance*** $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := \left(\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}\left[(d(X,Y))^p \right] \right)^{1/p}$ 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*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \geq 0$.
    2. (*Definiteness*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$
    3. (*Symmetric*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) = \mathcal{W}_{p,d}(\mathbb{Q}, \mathbb{P})$
    4. (*Triangular inequality*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \leq \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{M}) + \mathcal{W}_{p,d}(\mathbb{M}, \mathbb{Q})$

- **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 $\alpha_n \overset{d}{\to} \alpha$) 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 $X_n \overset{d}{\to} X$. This convergence can be shown (see [Villani, 2009, Santambrogio, 2015]) to be equivalent to

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

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

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 $\mathcal{X}$. For fixed $p \in [1, \infty)$, let Lip_1 be the space of all 1-**Lipschitz** function with respect to metric d such that*

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

Then

$$\mathcal{W}_d(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{1,d}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in Lip_1} \left\{ \mathbb{E}_{\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] \right\}. \tag{6}$$

- **Remark** This theorem only applies for *Wasserstein* 1-*distance*, i.e. $p = 1$.

- **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 $\mathcal{W}_{1,d}$ becomes

$$\begin{aligned}\mathcal{W}_{1,d_H}(\mathbb{P}, \mathbb{Q}) &= \min_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \gamma\{X \neq Y\} \\ &= \sup_{f:\mathcal{X} \to [0,1]} \int_{\mathcal{X}} f\,(d\mathbb{P} - d\mathbb{Q}) \\ &= \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)| := \|\mathbb{P} - \mathbb{Q}\|_{TV}\end{aligned}$$

- **Example** ($\mathcal{W}_1$ ***in 1-dimensional space*** $\mathbb{R}$)
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 $F_\beta^{-1} = \inf\{t : F_\beta \ge t\}$.

2 The Transportation Method

2.1 Concentration via Transportation Cost Inequality

- **Lemma 2.1** *(**Transportation Lemma**) [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 $\psi_X(\lambda) := \log \mathbb{E}_{\mathbb{P}}\left[e^{\lambda X}\right]$ *is the logarithm of moment generating function;*

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

$$\mathbb{E}_{\mathbb{Q}}[X] - \mathbb{E}_{\mathbb{P}}[X] \leq \phi^{*-1}\left(\mathbb{KL}(\mathbb{Q} \,\|\, \mathbb{P})\right). \tag{7}$$

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* $\mathbb{Q}$ *absolutely continuous with respect to* $\mathbb{P}$ *such that* $\mathbb{KL}(\mathbb{Q} \,\|\, \mathbb{P}) < \infty$*,*

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

- **Remark** (***Concentration via Transportation Methods***)
  Let $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$ be the product measure for $Z := (Z_1, \ldots, Z_n)$ on $\mathcal{X}^n$ and $f : \mathcal{X}^n \to \mathbb{R}$ be 1-*Lipschitz function*. Consider a probability measure $\mathbb{Q}$ on $\mathcal{X}^n$, absolutely continuous with respect to $\mathbb{P}$ and let Y be a random variable (defined on the same probability space as $\mathcal{X}$) such that Y has distribution $\mathbb{Q}$.

  The lemma above suggests that one may prove *sub-Gaussian concentration inequalities* for $X = f(Z_1, \ldots, Z_n)$ by proving a "*transportation*" *inequality* as above. The key to achieving this relies on *coupling*. In particular, *the Kantorovich-Rubenstein duality* for $\mathcal{W}_{1,d}$ suggests that

$$\mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(Z)] \leq \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_{\gamma}[d(Y, Z)] := \mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P})$$

  Thus, it suffices to *upper bound* the 1-*Wasserstein distance* between $\mathbb{Q}$ and $\mathbb{P}$.

- **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 ***d-transportation cost inequality*** with parameter $\nu > 0$ if

$$\mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu \mathbb{KL}(\mathbb{Q} \,\|\, \mathbb{P})} \tag{9}$$

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

- **Theorem 2.2** ***(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*** *with parameter $\nu/2 > 0$ in (9) Then its* ***concentration function*** *satisfies the bound*

$$\alpha_{\mathbb{P},(\mathcal{X},d)}(t) \le \exp\left(-\frac{(t-t_0)_+^2}{2\nu}\right), \text{ for } t \ge t_0 \tag{10}$$

where $t_0 := \sqrt{2\nu \log 2}$. Moreover, for any $Z \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the ***concentration inequality***

$$\mathbb{P}\left\{|f(Z) - \mathbb{E}\left[f(Z)\right]| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2\nu L^2}\right). \tag{11}$$

Proof: We begin by proving the bound (10). For any set A with $\mathbb{P}(A) \ge 1/2$ and a given $t > 0$, consider the set

$$A_t^c = \left\{x \in \mathcal{X} : d(x, A) \ge t\right\}.$$

If $\mathbb{P}(A_t) = 1$, then the proof is complete, so that we may assume that $P(A_t^c) > 0$. By construction, we have $d(A, A_t^c) := \inf_{x \in A_t^c} \inf_{y \in A} d(x, y) \ge t$. On the other hand, let $\mathbb{P}_A := \mathbb{P}(\cdot|A)$ and $\mathbb{P}_{A_t} := \mathbb{P}(\cdot|A_t^c)$ denote the distributions of $\mathbb{P}$ conditioned on A and A_t^c, and let γ denote any *coupling* of this pair. Since the marginals of γ are supported on A and A_t^c, respectively, we have

$$d(A, A_t^c) \le \int_{\mathcal{X}\times\mathcal{X}} d(x, x') d\gamma(x, x').$$

Taking the *infimum* over all *couplings*, we conclude that

$$t \le d(A, A_t^c) \le \inf_{\gamma \in \Pi(\mathbb{P}_A, \mathbb{P}_{A_t^c})} \int_{\mathcal{X}\times\mathcal{X}} d(x, x') d\gamma(x, x') := \mathcal{W}_{1,d}(\mathbb{P}_A, \mathbb{P}_{A_t^c})$$

Now applying the triangle inequality, we have

$$\begin{aligned} t \le \mathcal{W}_{1,d}(\mathbb{P}_A, \mathbb{P}_{A_t^c}) &\le \mathcal{W}_{1,d}(\mathbb{P}_A, \mathbb{P}) + \mathcal{W}_{1,d}(\mathbb{P}, \mathbb{P}_{A_t^c}) \\ &\le \sqrt{2\nu \mathbb{KL}\left(\mathbb{P}_A \,\|\, \mathbb{P}\right)} + \sqrt{2\nu \mathbb{KL}\left(\mathbb{P}_{A_t^c} \,\|\, \mathbb{P}\right)} \end{aligned}$$

It remains to compute the *Kullback-Leibler divergences*. For any measurable set C, we have

$$\begin{aligned} \mathbb{P}_A(C) &= \frac{\mathbb{P}(C \cap A)}{\mathbb{P}(A)} \\ g = \frac{d\mathbb{P}_A}{d\mathbb{P}} &= \frac{1}{\mathbb{P}(A)} \mathbb{1}\left\{A\right\} \\ \mathbb{KL}\left(\mathbb{P}_A \,\|\, \mathbb{P}\right) &= \int \log\left(\frac{d\mathbb{P}_A}{d\mathbb{P}}\right) d\mathbb{P}_A = \log\frac{1}{\mathbb{P}(A)} \end{aligned}$$

Similarly, we have $\mathbb{KL}\left(\mathbb{P}_{A_t^c} \,\|\, \mathbb{P}\right) = \log\frac{1}{\mathbb{P}(A_t^c)}$. Combining the pieces, we have

$$t \le \mathcal{W}_{1,d}(\mathbb{P}_A, \mathbb{P}_{A_t^c}) \le \sqrt{2\nu \log\frac{1}{\mathbb{P}(A)}} + \sqrt{2\nu \log\frac{1}{\mathbb{P}(A_t^c)}}$$

Denote $u = \sqrt{2\nu \log \frac{1}{\mathbb{P}(A)}}$, we have

$$(t-u)_{+} \le \sqrt{2\nu \log \frac{1}{\mathbb{P}(A_t^c)}}$$

$$\mathbb{P}(A_t^c) \le \exp\left(-\frac{(t-u)_{+}^2}{2\nu}\right), \text{ for } t \ge u.$$

Since $\mathbb{P}(A) \ge 1/2$ so $u \le \sqrt{2\nu \log 2}$. Thus for $t \ge \sqrt{2\nu \log 2}$, the concentration function

$$\alpha_{\mathbb{P},(\mathcal{X},d)}(t) = \sup_{A \subset \mathcal{X}: \mathbb{P}(A) \ge 1/2} \mathbb{P}(A_t^c) \le \exp\left(-\frac{\left(t - \sqrt{2\nu \log 2}\right)_{+}^2}{2\nu}\right),$$

which proves (10).

To show (11), we see that for L-Lipschitz function:

$$\mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] - \mathbb{E}_{\mathbb{P}}\left[f(Z)\right] \le L \min_{\gamma \in \Pi(\mathbb{Q},\mathbb{P})} \mathbb{E}_{\gamma}\left[d(Y,Z)\right] = L\, \mathcal{W}(\mathbb{Q},\mathbb{P}) \le \sqrt{2L^2 \nu \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)}$$

where the first inequality follows *the Kantorovich-Rubenstein duality* and the second inequality follows the assumption. By *the transportation lemma*,

$$\psi_{f(Z)-\mathbb{E}[f(Z)]}(\lambda) = \mathbb{E}_{\mathbb{P}}\left[e^{\lambda(f(Z)-\mathbb{E}[f(Z)])}\right] \le \frac{\nu L^2 \lambda^2}{2}$$

The upper tail bound thus follows by the Chernoff bound. The same argument can be applied to $-f$, which yields the lower tail bound. ∎

2.2 Tensorization for Transportation Cost

- **Proposition 2.3** ***(Tensorization for Transportation Cost)*** *[Boucheron et al., 2013]* *Suppose that, for each $k = 1, 2, \ldots, n$, the univariate distribution $\mathbb{P}_k$ satisfies a d_k**-transportation cost inequality** with parameter ν_k. Then **the product distribution** $\mathbb{P} = \otimes_{k=1}^{n} \mathbb{P}_k$ satisfies the transportation cost inequality*

$$\mathcal{W}_{1,d}(\mathbb{Q},\mathbb{P}) = \sqrt{2\left(\sum_{k=1}^{n} \nu_k\right) \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)}, \quad \textit{for all distributions } \mathbb{Q} \ll \mathbb{P} \tag{12}$$

where the Wasserstein metric is defined using the distance $d(x,y) := \sum_{k=1}^{n} d_k(x_k, y_k)$.

2.3 Induction Lemma

- **Lemma 2.4** *[Boucheron et al., 2013]* *Let $\mathbb{P} = \otimes_{i=1}^{n} \mathbb{P}_i$ be a **product probability measure** on a product measurable space $\mathcal{X}^n$ and let $\mathbb{Q}$ be a probability measure absolutely continuous with respect to $\mathbb{P}$ (i.e. $\mathbb{Q} \ll \mathbb{P}$). Let $w : \mathcal{X} \times \mathcal{X} \to [0,\infty)$ be a measurable function and let $\phi : [0,\infty) \to [0,\infty)$ be a **convex***

function. *Suppose that for every $i = 1, \ldots, n$ and for every probability measure $\nu \ll \mathbb{P}_i$ which is absolutely continuous with respect to $\mathbb{P}_i$,*

$$\min_{\gamma \in \Pi(\mathbb{P}_i, \nu)} \phi\left(\mathbb{E}_{\gamma}\left[w(X_i, Y_i)\right]\right) \le \mathbb{KL}\left(\nu \,\|\, \mathbb{P}_i\right)$$

Then

$$\min_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \sum_{i=1}^{n} \phi\left(\mathbb{E}_{\gamma}\left[w(X_i, Y_i)\right]\right) \le \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right).$$

2.4 Marton's Transportation Inequality

- **Theorem 2.5** ***(Marton's Transportation Inequality)*** *[Boucheron et al., 2013]*
  Let $\mathbb{P} = \otimes_{k=1}^{n} \mathbb{P}_k$ be a product probability measure on $\mathcal{X}^n$, and let $\mathbb{Q}$ be a probability measure absolutely continuous with respect to $\mathbb{P}$. Define two random vectors $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ in $\mathcal{X}^n$ with distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} \gamma^2 \left\{X_i \neq Y_i\right\} \le \frac{1}{2} \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right) \tag{13}$$

- **Proof:** (***Proof of Bounded Difference Inequality***)
  Any function with ***bounded difference property*** is ***Lipschitz function*** with respect to ***Hamming distance***. This implies that for all $x, y \in \mathcal{X}^n$,

$$f(y) - f(x) \le \sum_{i=1}^{n} L_i \mathbb{1}\left\{x_i \neq y_i\right\} \equiv d_{H,L}(x, y).$$

  Note that for coupling $\gamma \in \Pi(\mathbb{Q}, \mathbb{P})$ where $Y \sim \mathbb{Q}$ and $X \sim \mathbb{P}$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] - \mathbb{E}_{\mathbb{P}}\left[f(X)\right] &= \mathbb{E}_{\gamma}\left[f(Y) - f(X)\right] \\
&\le \sum_{i=1}^{n} L_i \mathbb{E}_{\gamma}\left[\mathbb{1}\left\{X_i \neq Y_i\right\}\right] \\
&\le \left(\sum_{i=1}^{n} L_i^2\right)^{1/2} \left(\sum_{i=1}^{n} (\mathbb{E}_{\gamma}\left[\mathbb{1}\left\{X_i \neq Y_i\right\}\right])^2\right)^{1/2}
\end{aligned}$$

  We want to prove the concentration using transportation cost inequality. That is, to bound the term

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} (\mathbb{E}_{\gamma}\left[\mathbb{1}\left\{X_i \neq Y_i\right\}\right])^2 = \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} \gamma^2 \left\{X_i \neq Y_i\right\}.$$

  We have shown that

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma\left\{X \neq Y\right\} = \mathcal{W}_{1, d_H}(\mathbb{Q}, \mathbb{P}) = \sup_{A \in \mathcal{X}} \left|\mathbb{Q}(A) - \mathbb{P}(A)\right| \equiv \left\|\mathbb{Q} - \mathbb{P}\right\|_{TV}.$$

For each independent variable X_i, Y_i, and their marginal distribution $\mathbb{P}_i, \mathbb{Q}_i$ where $\mathbb{Q}_i \ll \mathbb{P}_i$, by Pinsker's inequality,

$$\min_{\gamma \in \Pi(\mathbb{Q}_i, \mathbb{P}_i)} \gamma \{X_i \neq Y_i\} \leq \sqrt{\frac{1}{2} \mathbb{KL}\left(\mathbb{Q}_i \,\|\, \mathbb{P}_i\right)}$$

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma^2 \{X_i \neq Y_i\} \leq \left[\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma \{X_i \neq Y_i\}\right]^2 \leq \frac{1}{2} \mathbb{KL}\left(\mathbb{Q}_i \,\|\, \mathbb{P}_i\right)$$

Thus by induction lemma,

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} \gamma^2 \{X_i \neq Y_i\} \leq \frac{1}{2} \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)$$

which is the *Marton's transportation inequality.* Finally, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] - \mathbb{E}_{\mathbb{P}}\left[f(X)\right] &\leq \left(\sum_{i=1}^{n} L_i^2\right)^{1/2} \left(\sum_{i=1}^{n} (\mathbb{E}_{\gamma}\left[\mathbb{1}\{X_i \neq Y_i\}\right])^2\right)^{1/2} \\ &\leq \sqrt{\frac{\left(\sum_{i=1}^{n} L_i^2\right)}{2} \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)}. \end{aligned}$$

Then we can apply the transportation lemma with $\nu := \frac{1}{4} \sum_{i=1}^{n} L_i^2$, which proves the bounded difference inequality. ∎

- **Theorem 2.6** ***(Marton's Conditional Transportation Inequality)*** *[Boucheron et al., 2013]*
  Let $\mathbb{P} = \otimes_{k=1}^{n} \mathbb{P}_k$ be a product probability measure on $\mathcal{X}^n$, and let $\mathbb{Q}$ be a probability measure absolutely continuous with respect to $\mathbb{P}$. Define two random vectors $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ in $\mathcal{X}^n$ with distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_{\gamma}\left[\sum_{i=1}^{n} (\gamma^2 \{X_i \neq Y_i | X_i\} + \gamma^2 \{X_i \neq Y_i | Y_i\})\right] \leq 2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right) \tag{14}$$

- **Proposition 2.7** ***(Concentration of Lipschitz Function with Function Weighted Hamming Distance)*** *[Boucheron et al., 2013]*
  Let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function and let $Z_1, \ldots, Z_n$ be independent random variables taking their values in $\mathcal{X}$. Define $X = f(Z_1, \ldots, Z_n)$. Assume that there exist ***measurable functions*** *$c_i : \mathcal{X}_n \to [0, \infty)$ such that for all $x, y \in \mathcal{X}^n$,*

$$f(y) - f(z) \leq \sum_{i=1}^{n} c_i(z) \mathbb{1}\{y_i \neq z_i\}.$$

  Setting

$$\nu = \mathbb{E}\left[\sum_{i=1}^{n} c_i^2(Z)\right] \qquad \text{and} \qquad \nu_{\infty} = \sup_{z \in \mathcal{X}^n} \sum_{i=1}^{n} c_i^2(z)$$

  for all $\lambda > 0$, we have

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \frac{\nu \lambda^2}{2} \qquad \text{and} \qquad \psi_{-X + \mathbb{E}[X]}(\lambda) \leq \frac{\nu_{\infty} \lambda^2}{2}$$

In particular, for all $t > 0$,

$$\mathbb{P}\{X \geq \mathbb{E}[X] + t\} \leq \exp\left(-\frac{t^2}{2\nu}\right)$$
$$\mathbb{P}\{X \leq \mathbb{E}[X] - t\} \leq \exp\left(-\frac{t^2}{2\nu_\infty}\right). \tag{15}$$

- **Remark** The condition in above proposition covers
  1. *Lipschitz functions* such as *functions with bounded difference*,
  2. ***self-bounding functions*** including ***configuration functions***: Let f be such a configuration function. For any $z \in \mathcal{X}^n$, fix a *maximal sub-sequence* $(z_{i,1}, \ldots, z_{i,m})$ satisfying property Π (so that $f(z) = m$). Let $c_i(z)$ denote *the indicator that z_i belongs to the sub-sequence* $(z_{i,1}, \ldots, z_{i,m})$. Thus,

     $$\sum_{i=1}^n c_i^2(z) = \sum_{i=1}^n c_i(z) = f(z).$$

     It follows from the definition of a configuration function that for all $z, y \in \mathcal{X}^n$,

     $$f(y) \geq f(z) - \sum_{i=1}^n c_i(z)\mathbb{1}\{z_i \neq y_i\}$$

     So $g = -f$ satisfies the condition in above proposition.
  3. ***weakly self-bounding functions***
  4. ***convex distance function***

     $$d_T(z, A) := \sup_{\alpha \in \mathbb{R}_+^n : \|\alpha\|_2 = 1} \inf_{y \in A} \sum_{i=1}^n \alpha_i \mathbb{1}\{z_i \neq y_i\}$$

     Denote by $c(z) = (c_1(z), \ldots, c_n(z)) = \alpha^*$ the vector of nonnegative components *in the unit ball* for which *the supremum is achieved*. Thus

     $$\begin{aligned} d_T(z, A) - d_T(y, A) &\leq \inf_{z' \in A} \sum_{i=1}^n c_i(z)\mathbb{1}\{z_i \neq z_i'\} - \inf_{y' \in A} \sum_{i=1}^n c_i(z)\mathbb{1}\{y_i \neq y_i'\} \\ &\leq \sum_{i=1}^n c_i(z)\mathbb{1}\{z_i \neq y_i\} \end{aligned}$$

2.5 Talagrand's Gaussian Transportation Inequality

- **Theorem 2.8** ***(Talagrand's Gaussian Transportation Inequality)*** *[Boucheron et al., 2013]*
  Let $\mathbb{P}$ be be the standard Gaussian probability measure on $\mathbb{R}^n$, and let $\mathbb{Q}$ be a probability measure absolutely continuous with respect to $\mathbb{P}$. Define two random vectors $X = (X_1, \ldots, X_n), Y =$

$(Y_1, \ldots, Y_n)$ *in* $\mathcal{X}^n$ *with distribution* $\mathbb{P}$ *and* $\mathbb{Q}$ *respectively. Then*

$$\mathcal{W}_{2,d}(\mathbb{Q},\mathbb{P}) := \sqrt{\min_{\gamma\in\Pi(\mathbb{Q},\mathbb{P})} \sum_{i=1}^{n} \mathbb{E}_{\gamma}\left[(X_i - Y_i)^2\right]} \le \sqrt{2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)} \tag{16}$$

$$\Leftrightarrow \min_{\gamma\in\Pi(\mathbb{Q},\mathbb{P})} \sum_{i=1}^{n} \mathbb{E}_{\gamma}\left[(X_i - Y_i)^2\right] \le 2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)$$

- **Remark** (***Gaussian Transportation Inequality*** $\Rightarrow$ ***Gaussian Concentration Inequality***) [Boucheron et al., 2013]
  Talagrand's ***Gaussian transportation inequality*** implies *the Tsirelson-Ibragimov-Sudakov inequality* (i.e. ***the dimension-free concentration*** of *Lipschitz function* of Gaussian vectors), which we proved based on *the Gaussian logarithmic Sobolev inequality* and *Herbst's argument*.

  Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a *Lipschitz function* with respect to *Euclidean distance*, that is, for all $x, y \in \mathbb{R}^n$,

  $$f(y) - f(x) \le L\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

  Then, by *Jensen's inequality*, for every *coupling* $\gamma \in \Pi(\mathbb{P},\mathbb{Q})$, one has

  $$\begin{aligned}
  \mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] - \mathbb{E}_{\mathbb{P}}\left[f(X)\right] &= \mathbb{E}_{\gamma}\left[f(Y) - f(X)\right] \\
  &\le L\mathbb{E}_{\gamma}\left[\left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)^{1/2}\right] \\
  &\le L\left(\sum_{i=1}^{n}\mathbb{E}_{\gamma}\left[(X_i - Y_i)^2\right]\right)^{1/2} = L\,\mathcal{W}_2(\mathbb{Q},\mathbb{P}) \\
  &\le \sqrt{2L^2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)} \quad \text{by Gaussian Transportation Inequality}
  \end{aligned}$$

  By transportation lemma, we show that $f(X) - \mathbb{E}\left[f(X)\right]$ is *sub-Gaussian distributed* with parameter L^2. This implies ***the Gaussian concentration inequality***.

2.6 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.