High Dimensional Probability and Statistics **2nd Semester, 2022-2023**

Lecture 3: Lipschitz Concentration and Transportation Method

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2023/03/20)*

Recap and Motivation: In this lecture, we take a reverse route by fixing the family of functions and then asking for what kind of random variables concentration phenomena will display. In particular, we will consider Lipschitz functions.

Definition 3.1 *Letting $(\mathcal{X}, d)$ be a (measurable) metric space, we say a function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if $|f(x) - f(y)| \le Ld(x, y)$ for all $x, y \in \mathcal{X}$.*

Remark 3.2 *Note that in this lecture $\mathcal{X}$ can used to denote a single metric space or a product metric space, which should be clear from its context. In addition, the Lipschitz condition is indeed equivalent to $f(x) - f(y) \le Ld(x, y)$ for all $x, y \in \mathcal{X}$ since $d(x, y) = d(y, x)$.*

In the last lecture we have already seen a result about concentration of Lipchitz functions - Gaussian concentration. That is, letting $X_1, \cdots, X_n$ be i.i.d standard Gaussian random variables taking values in $\mathcal{X} = \mathbb{R}$, if $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz function with respect to the $\|\cdot\|_2$ norm, then $f(X_1, \cdots, X_n)$ presents the Gaussian type concentration. It is natural to study the general principal for the concentration phenomena of Lipschitz functions and answer the following question:

> *When $f(X_1, \cdots, X_n)$ is sub-Gaussian for f being a Lipschitz function under certain metric defined on $\bigotimes_{k=1}^n \mathcal{X}_k$?*

The answer to this question essentially relies on the distribution[1] of $(X_1, \cdots, X_n)$. As in the entropy method, there are two key ingredients in the analysis: 1) a new characterization of the sub-Gaussian property based on the Wasserstein distance (known as transportation lemma), 2) a tensorization property (known as Marton theorem) which can transfer the problem from the general n case to the $n = 1$ case.

In fact we can also express the bounded difference inequality as the concentration of Lipschitz functions under a properly chosen metric. Let $(X_1, \cdots, X_n)$ be a vector of independent random variables taking values in $\bigotimes_{k=1}^n \mathcal{X}_k := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any function $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ satisfying the bounded difference property

$$|f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \le L_k,$$

$f(X_1, \cdots, X_n)$ has the Gaussian type concentration. To rephrase this result into a concentration result of Lipschitz functions, first define the following *weighted Hamming metric* on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$,

$$d_L(x, y) = \sum_{k=1}^n L_k 1_{\{x_k \neq y_k\}}, \quad \text{where } 1_{\{x_k \neq y_k\}} = \begin{cases} 1 & \text{if } x_k \neq y_k \\ 0 & \text{if } x_k = y_k. \end{cases}$$

[1] Or relies on the property of the probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$ induced by $(X_1, \cdots, X_n)$ since the distribution of the random variable $f(X_1(\omega), \cdots, X_n(\omega))$, where ω is in a probability space $(\Omega, \mathcal{F}, \mu)$, is completely determined by the probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$ induced by $(X_1, \cdots, X_n)$. Thus, we may equivalently ask: under what probability measure on $\bigotimes_{k=1}^n \mathcal{X}_k$, $f(x_1, \cdots, x_n)$ is sub-Gaussian?

Exercise 3.3 *Verify that $d_L(\cdot,\cdot)$ is a metric.*

Assuming f satisfies the bounded difference property, it follows that

$$\begin{aligned} f(x) - f(y) &= \sum_{k=1}^{n} \left(f(x_1, \cdots, x_{k-1}, x_k, y_{k+1}, \cdots, y_n) - f(x_1, \cdots, x_{k-1}, y_k, y_{k+1}, \cdots, y_n)\right) \\ &\leq \sum_{k=1}^{n} L_k 1_{\{x_k \neq y_k\}} \\ &= d_L(x, y). \end{aligned}$$

That is, f is 1-Lipschitz under with respect to the metric d_L. Therefore the bounded difference inequality can be rephrased as: Assume f is 1-Lipschitz under with respect to the metric d_L. Then for any independent random variables $X_1, \cdots, X_n$, $f(X_1, \cdots, X_n)$ is sub-Gaussian.

Agenda:

- KL divergence, Wasserstein distance
- Transportation lemma and tensorization
- Talagrand concentration inequality
- Short summary

3.1 KL Divergence, Wasserstein Distance

In this section we introduce two notions, KL divergence and Wasserstein distance, to measure the divergence or distance between the two probability distributions. These two distances are not only useful here but actually widely used in machine learning. Of course there are other divergence measures which will be introduced in the due course.

3.1.1 KL Divergence

Definition 3.4 (Kullback-Leibler (KL) Divergence) *Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$, the KL divergence (or relative entropy) of $\mathbb{Q}$ with respect to $\mathbb{P}$ is defined as*

$$D(\mathbb{Q}\|\mathbb{P}) = \begin{cases} \mathrm{Ent}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] & \text{if } \mathbb{Q} \ll \mathbb{P} \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathrm{Ent}_{\mathbb{P}}[\cdot]$ means computing the entropy (see Definition 2.1 of Lecture 2) of $d\mathbb{Q}/d\mathbb{P}$ under the probability distribution $\mathbb{P}$.

Remark 3.5 *The KL divergence quantifies the difference of $\mathbb{P}$ and $\mathbb{Q}$ using the randomness of $\mathbb{Q}$ relative to $\mathbb{P}$. Thus, KL divergence is also known as relative entropy. Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$, $\mathbb{Q} \ll \mathbb{P}$ means $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, namely, there exists a (real-valued) nonnegative random variable $Y(x)$ with $\mathbb{E}_{\mathbb{P}}[Y] = \int_{\mathcal{X}} Y d\mathbb{P} = 1$ such that*

$$\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}[Y 1_A] \quad \text{for any measurable } A.$$

If $\mathbb{P}$ and $\mathbb{Q}$ have densities with respect to some underlying measure μ (e.g., take $\mu = \mathbb{P}+\mathbb{Q}$), denoted $p(x)$ and $q(x)$, then

$$\mathbb{Q}(A) = \int_A q(x)\mu(dx) = \int_A \frac{q(x)}{p(x)} p(x)\mu(dx) = \mathbb{E}_{\mathbb{P}}\left[\frac{q(x)}{p(x)}\right],$$

and thus Y can be chosen to be $Y(x) = q(x)/p(x)$.

Remark 3.6 (Equivalent definition of KL-divergence) *By the definition of entropy[2], we have*

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) &= \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} \log \frac{d\mathbb{Q}}{d\mathbb{P}}\right] - \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \log \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right]. \end{aligned} \tag{3.1}$$

If both $\mathbb{P}$ and $\mathbb{Q}$ have densities with respect to some underlying measure μ, then

$$D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \mu(dx). \tag{3.2}$$

In particular, when x is a discrete space and $\mathbb{P}$ and $\mathbb{Q}$ are discrete probability distributions, we have

$$D(\mathbb{Q}\|\mathbb{P}) = \sum_{x\in\mathcal{X}} q(x) \log \frac{q(x)}{p(x)}. \tag{3.3}$$

Note that $D(\mathbb{Q}\|\mathbb{P})$ is not a metric ($D(\mathbb{Q}\|\mathbb{P}) \neq D(\mathbb{P}\|\mathbb{Q})$ in general, **give an example!**). However, we do have the following lemma.

Lemma 3.7 *$D(\mathbb{Q}\|\mathbb{P}) \geq 0$ and the equality holds if and only if $\mathbb{P} = \mathbb{Q}$ (almost everywhere).*

Proof: Use the definition in (3.1) and Jensen's inequality (noting that $\log x$ is strictly convex). ■

Example 3.8 *Let $\mathbb{P} = \mathcal{N}(\mu_1, \sigma^2)$ and $\mathbb{Q} = \mathcal{N}(\mu_2, \sigma^2)$. Then*

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) &= \mathbb{E}_{X\sim\mathbb{Q}}\left[\frac{-(X-\mu_2)^2}{2\sigma^2} + \frac{-(X-\mu_1)^2}{2\sigma^2}\right] \\ &= \mathbb{E}_{x\sim\mathbb{Q}}\left[\frac{\mu_1^2 - \mu_2^2 - 2(\mu_1-\mu_2)x}{2\sigma^2}\right] \\ &= \frac{\mu_1^2 - \mu_2^2 - 2(\mu_1-\mu_2)\mu_2}{2\sigma^2} \\ &= \frac{(\mu_1-\mu_2)^2}{2\sigma^2}. \end{aligned}$$

Exercise 3.9 *Compute the KL divergence between two multivariate Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$.*

The following lemmas establishes a connection between moment generating function and the KL divergence, showing the duality property between them.

[2]We will always assume $\mathbb{Q} \ll \mathbb{P}$ without specifying this next.

Lemma 3.10 (Duality between KL and MGF) *We have*

$$\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] = \sup_{\mathbb{Q}\ll\mathbb{P}} \left\{\mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})\right\}.$$

Proof: Let $Z = f(X) - \log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right]$. Since $\mathbb{E}_{\mathbb{P}}\left[e^{Z}\right] = 1$, by the variational form of entropy in Lecture 2, we have

$$D(\mathbb{Q}\|\mathbb{P}) = \mathrm{Ent}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] \geq \mathbb{E}_{\mathbb{P}}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}Z\right] = \mathbb{E}_{\mathbb{Q}}\left[Z\right] = \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - \log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right],$$

or equivalently that

$$\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] \geq \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P}).$$

Since this holds for any $\mathbb{Q} \ll \mathbb{P}$, it follows that $\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] \geq \sup_{\mathbb{Q}\ll\mathbb{P}} \left\{\mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})\right\}$.

In addition, if we define $\mathbb{Q}$ by

$$d\mathbb{Q} = \frac{e^{f(X)}}{\mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right]} d\mathbb{P},$$

a direct calculation can show that $\log \mathbb{E}_{\mathbb{P}}\left[e^{f(X)}\right] = \mathbb{E}_{\mathbb{Q}}\left[f(X)\right] - D(\mathbb{Q}\|\mathbb{P})$ (**check this!**). This completes the proof. ■

Lemma 3.11 (Chain rule of KL) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures that define the joint distribution of random variables (X_1, X_2). Then,*

$$D(\mathbb{Q}\|\mathbb{P}) = D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1}\left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right)\right],$$

where $\mathbb{P}_1$ and $\mathbb{Q}_2$ are marginal distributions of X_1 under the joint distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively, and $\mathbb{P}_2(\cdot|X_1)$ and $\mathbb{Q}_2(\cdot|X_1)$ are the conditional distribution of X_2 given X_1 under the joint distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Proof: It follows from the Bayes formula that

$$d\mathbb{P} = d\mathbb{P}_1 \cdot d\mathbb{P}_2(\cdot|X_1) \quad \text{and} \quad d\mathbb{Q} = d\mathbb{Q}_1 \cdot d\mathbb{Q}_2(\cdot|X_1).$$

Consequently,

$$\begin{aligned} D(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log\frac{d\mathbb{Q}}{d\mathbb{P}}\right] &= \mathbb{E}_{\mathbb{Q}}\left[\log\frac{d\mathbb{Q}_1}{d\mathbb{P}_1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\log\frac{d\mathbb{Q}_2(\cdot|X_1)}{d\mathbb{P}_2(\cdot|X_1)}\right] \\ &= \mathbb{E}_{\mathbb{Q}_1}\left[\log\frac{d\mathbb{Q}_1}{d\mathbb{P}_1}\right] + \mathbb{E}_{\mathbb{Q}_1}\left[\mathbb{E}_{\mathbb{Q}_2(\cdot|X_1)}\left[\log\frac{d\mathbb{Q}_2(\cdot|X_1)}{d\mathbb{P}_2(\cdot|X_1)}\right]\right] \\ &= D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1}\left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right)\right] \end{aligned}$$

as claimed. ■

Remark 3.12 *To have a better understanding of the above lemma, let us particularly consider the case when (X_1, X_2) are real-valued and $\mathbb{P}$ and $\mathbb{Q}$ have continuous densities $p(x_1, x_2)$ and $q(x_1, x_2)$ with respect to the Lebesgue measure, respectively. Let $p_1(x_1)$ and $q_1(x_1)$ be the marginal distribution of X_1 and X_2 under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Let $p_2(x_2|x_1)$ and $q_2(x_2|x_1)$ be the conditional probability densities of X_2 given X_1 under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then*

$$p_2(x_2|x_1) = \frac{p(x_1, x_2)}{\int_{\mathbb{R}} p(x_1, x_2)dx_2} \quad \textit{and} \quad q_2(x_2|x_1) = \frac{q(x_1, x_2)}{\int_{\mathbb{R}} q(x_1, x_2)dx_2}.$$

It follows that

$$\begin{aligned}
D(\mathbb{Q}\|\mathbb{P}) &= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q(x_1, x_2)}{p(x_1, x_2)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_2(x_2|x_1) \left(\int_{\mathbb{R}} q(x_1, x_2)dx_2\right)}{p_2(x_2|x_1) \left(\int_{\mathbb{R}} p(x_1, x_2)dx_2\right)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{\left(\int_{\mathbb{R}} q(x_1, x_2)dx_2\right)}{\left(\int_{\mathbb{R}} p(x_1, x_2)dx_2\right)} dx_1 dx_2 + \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_1 dx_2 \\
&= \int_{\mathbb{R}\times\mathbb{R}} q(x_1, x_2) \log \frac{q_1(x_1)}{p_1(x_1)} dx_1 dx_2 + \int_{\mathbb{R}\times\mathbb{R}} q_1(x_1) q_2(x_2|x_1) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_1 dx_2 \\
&= \int_{\mathbb{R}} q_1(x_1) \log \frac{q_1(x_1)}{p_1(x_1)} dx_1 + \int_{\mathbb{R}} q_1(x_1) \left(\int_{\mathbb{R}} q_2(x_2|x_1) \log \frac{q_2(x_2|x_1)}{p_2(x_2|x_1)} dx_2 \right) dx_1 \\
&= D(\mathbb{Q}_1\|\mathbb{P}_1) + \mathbb{E}_{\mathbb{Q}_1} \left[D\left(\mathbb{Q}_2(\cdot|X_1)\|\mathbb{P}_2(\cdot|X_1)\right) \right].
\end{aligned}$$

Exercise 3.13 *Let $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$ and $\mathbb{Q} = \mathbb{Q}_1 \times \mathbb{Q}_2$ two product probability measures that define the joint distribution of random variables (X_1, X_2). Show that*

$$D(\mathbb{Q}\|\mathbb{P}) = D(\mathbb{Q}_1\|\mathbb{P}_1) + D(\mathbb{Q}_2\|\mathbb{P}_2).$$

Exercise 3.14 *Generalize the result in Lemma 3.11 and Exercise 3.13 to the case of joint distributions of n random variables.*

3.1.2 Wasserstein Distance

Wasserstein distance is a distance defined over probability measures within the framework of optimal transport. Roughly speaking, it is the least cost of transporting/redistributing a source probability measure to a target probability measure. Optimal transport was first introduced in the Monge formulation. Then Kantorovich relaxed it by allowing the mass splitting in the source based on the notation of coupling.

Definition 3.15 (Coupling) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two given probability measures on $\mathcal{X}$. We say a probability measures π on $\mathcal{X} \times \mathcal{X}$ is a coupling of $\mathbb{P}$ and $\mathbb{Q}$ if the marginal distributions of π in the first and second coordinate coincides with $\mathbb{P}$ and $\mathbb{Q}$, respectively. In addition, we denote by $\mathcal{C}(\mathbb{P}, \mathbb{Q})$ the set of all the couplings of $\mathbb{P}$ and $\mathbb{Q}$.*

Exercise 3.16 *Does the coupling always exist? Are there always more than two couplings for a fixed pair of probability measures?*

Definition 3.17 (Wasserstein Distance) *Let $(\mathcal{X}, d)$ be a metric space. Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$, their Wasserstein distance is defined as*

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X} \times \mathcal{X}} d(x, y) d\pi(x, y) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] \tag{3.4}$$

If we view the joint distribution π as a transport plan, meaning a scheme for reshuffling the probability mass $\mathbb{P}$ to another probability mass $\mathbb{Q}$, and view $d(\cdot, \cdot)$ as the unit transport cost, then $\mathbb{E}_\pi \left[d(X, Y) \right]$ can be interpreted as the transport cost of associated with the plan π. Seeking the the transportation plan that minimizes the transport cost is the optimal transport problem. The solution to the optimal transport problem measures how far we have to move the mass of $\mathbb{P}$ and turn it into $\mathbb{Q}$ and thus is a natural way to define the distance between two probability measures.

Remark 3.18

1. *If d is a distance on $\mathcal{X}$, $W_1(\mathbb{P}, \mathbb{Q})$ is indeed a distance. Namely, it satisfies the three conditions required for a distance, especially the triangular inequality. A proof of this can be found in [4] and references therein.*

2. *It is evident that we can express the Wasserstein distance as*

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{(X, Y)} \{ \mathbb{E} \left[d(X, Y) \right], \ X \sim \mathbb{P}, \ Y \sim \mathbb{Q} \}.$$

3. *What we have in* (3.4) *is actually the* 1*-Wasserstein distance, hence there is subscript* 1 *in the notation. In general, we may also define p-Wasserstein distance as follows:*

$$W_p = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \left(\mathbb{E}_\pi \left[[d(X, Y)]^p \right] \right)^{1/p}. \tag{3.5}$$

4. *Computing the Wasserstein distance is generally difficult and relies on some numerical solvers. A detailed discussion of the computations is beyond the scope of this lecture.*

If viewing (3.4) as an optimization problem (equality constrained) on the infinity dimensional probability measure space, we can compute its dual problem. In addition, since every probability measure corresponds a linear functional over the function space on $\mathcal{X}$, we can also define the distance between probability measures based on the perspective of linear functional (similar to operator norm). This provides another duality for Wasserstein distance which plays an important role in this lecture.

Theorem 3.19 (Duality) *Under mild conditions, we have*

$$W_1(\mathbb{P}, \mathbb{Q}) \triangleq \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] = \sup_{f \in \mathrm{Lip}(\mathcal{X}, d)} \left| \mathbb{E}_\mathbb{P} \left[f(X) \right] - \mathbb{E}_\mathbb{Q} \left[f(Y) \right] \right|. \tag{3.6}$$

The proof of this theorem can be found in [4], and we only consider a simple example here.

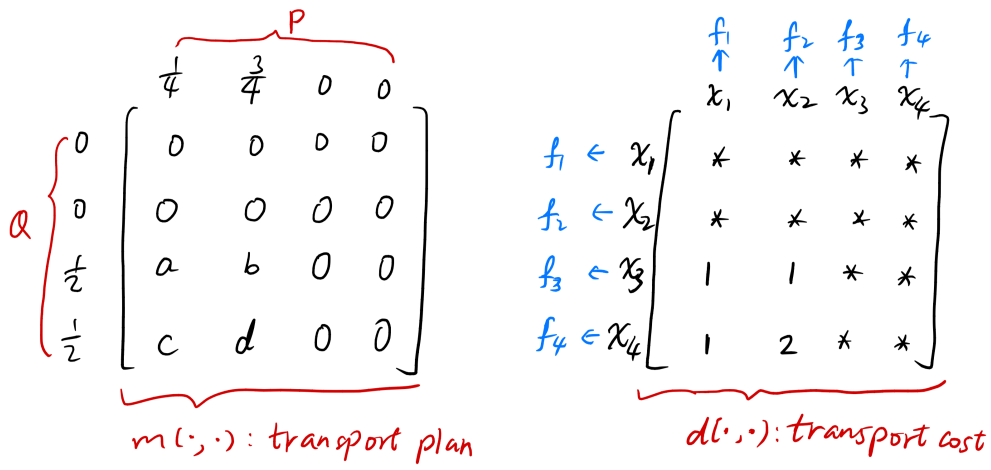

Figure 3.1: Problem setup of Example 3.20.

Example 3.20 (Discrete example) *Here we consider* $\mathbb{P}$ *and* $\mathbb{Q}$ *defined on a discrete space* $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$*, see Figure 3.1 for the problem setup. In order for* π *to be a coupling of* $\mathbb{P}$ *and* $\mathbb{Q}$ *we must have*

$$a + c = 1/4,\ b + d = 3/4,\ a + b = 1/2,\ c + d = 1/2,\ a, b, c, d \geq 0.$$

For example, if we take $\pi = \mathbb{P} \otimes \mathbb{Q}$*, then*

$$a = 1/8,\ b = 3/8,\ c = 1/8,\ d = 3/8,$$

with the total transport cost

$$c_1 = \mathbb{E}_\pi\left[d(X, Y)\right] = 1/8 + 3/8 + 1/8 + 6/8 = 11/8.$$

There are also exists other coupling of $\mathbb{P}$ *and* $\mathbb{Q}$ *(or transport plan) in addition to* $\pi_2 = \mathbb{P} \otimes \mathbb{Q}$*. For example, we may let*

$$a = 0,\ b = 1/2,\ c = 1/4,\ d = 1/4,$$

with the total transport cost

$$c_2 = \mathbb{E}_\pi\left[d(X, Y)\right] = 0 + 1/2 + 1/4 + 1/2 = 5/4 < c_1.$$

In fact, this is the minimum total transport cost we can achieve over all the possible couplings of $\mathbb{P}$ *and* $\mathbb{Q}$ *(**why?**), i.e.,*

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi\left[d(X, Y)\right] = 5/4.$$

Moreover, let $f = (f_1, f_2, f_3, f_3)$ *be a function defined on* $\mathcal{X}$*. Then* f *is 1-Lipschitz under* $d(\cdot, \cdot)$ *if and only if*

$$|f_1 - f_3| \leq d(x_1, x_3) = 1,\ |f_2 - f_3| \leq d(x_2, x_3) = 1,\ |f_1 - f_4| \leq d(x_1, x_4) = 1,\ |f_2 - f_4| \leq d(x_2, x_4) = 2.$$

Given a 1-Lipschitz f, letting π (with (a,b,c,d)) be any coupling of $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| &= \left|\frac{1}{4}f_1 + \frac{3}{4}f_2 - \frac{1}{2}f_3 - \frac{1}{2}f_4\right| \\
&= |(a+c)f_1 + (b+d)f_2 - (a+b)f_3 - (c+d)f_4| \\
&= |a(f_1 - f_3) + b(f_2 - f_3) + c(f_1 - f_4) + d(f_2 - f_4)| \\
&\leq \mathbb{E}_{\pi}[d(X,Y)].
\end{aligned}$$

It follows that,

$$\sup_{f \in \mathrm{Lip}(\mathcal{X},d)} |\mathbb{E}_{\mathbb{P}}[f] - \mathbb{E}_{\mathbb{Q}}[f]| \leq \inf_{\pi \in \mathcal{C}(\mathbb{P},\mathbb{Q})} \mathbb{E}_{\pi}[d(X,Y)].$$

The equality can be achieved for example by taking $f = (f_1, f_2, f_3, f_4) = (0, 1, 0, -1)$. Thus, we have verified the duality theorem using this example.

Next we study a special case of the Wasserstein distance with the trivial metric $d(x,y) = 1_{\{x \neq y\}}$. In this case it can be shown that the Wasserstein distance is none other than the total variation distance which itself is interesting in many applications.

Definition 3.21 (Total variation distance) *The total variation distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

Example 3.22 (Wasserstein distance for $d(x,y) = 1_{\{x \neq y\}}$) *In this case it is not hard to see that f is 1-Lipschitz if and only if*

$$|f(x) - f(y)| \leq 1.$$

Since the Wasserstein distance is invariant to constant offsets of the function, we have

$$W_1(\mathbb{P},\mathbb{Q}) = \sup_{0 \leq f \leq 1} |\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Y)]|.$$

Let $d\mathbb{P}(x) = p(x)\mu(dx)$ and $d\mathbb{Q}(x) = q(x)\mu(dx)$ for some underlying measure μ. Then we can rewrite the Wasserstein distance as

$$\begin{aligned}
W_1(\mathbb{P},\mathbb{Q}) &= \sup_{0 \leq f \leq 1} \left|\int_{\mathcal{X}} f(x)(p(x) - q(x))\mu(dx)\right| \\
&= \int_{\mathcal{X}} [p(x) - q(x)]_+ \mu(dx). \qquad (3.7)
\end{aligned}$$

On the other hand, since

$$|\mathbb{P}(A) - \mathbb{Q}(A)| = \left|\int_A (p(x) - q(x))\mu(dx)\right|,$$

it is not hard to see that the supremum is attained at $A_1 = \{x : \ p(x) \geq q(x)\}$ *or* $A_2 = \{x : \ q(x) \geq p(x)\}$ ***(show that*** $|\mathbb{P}(A) - \mathbb{Q}(A)|$ ***have the same value on these two sets!)****. It follows that*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \int_{\mathcal{X}} [p(x) - q(x)]_+ \, \mu(dx).$$

Therefore we have

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi [X \neq Y] = \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}. \tag{3.8}$$

In this case we can also construct the optimal coupling/transport plan explicitly. We refer interested readers to Chapter 4.2 of [2] for details.

3.2 Transportation Lemma and Tensorization

3.2.1 Transportation Lemma

There is a necessary and sufficient condition in terms of the two probability divergence measures to characterize the sub-Gaussian property of Lipschitz functions. We begin with a sub-Gaussian characterization in terms of the KL divergence.

Lemma 3.23 (Sub-Gaussian in terms of KL) *Letting* $X \sim \mathbb{P}$*, then* $f(X)$ *is* ν^2*-sub-Gaussian if and only if*

$$|\mathbb{E}_{\mathbb{Q}} [f(Y)] - \mathbb{E}_{\mathbb{P}} [f(X)]| \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \textit{for all } \mathbb{Q} \ll \mathbb{P}.$$

Proof: By the definition $f(X)$ is sub-Gaussian if and only if

$$\log \mathbb{E}_{\mathbb{P}} \left[e^{\lambda (f(X) - \mathbb{E}_{\mathbb{P}}[f(X))]} \right] \leq \frac{\lambda^2 \nu^2}{2} \quad \text{for all } \lambda \in \mathbb{R}.$$

Then, by the duality between DL divergence and MGF, this is equivalent to

$$\lambda \left(\mathbb{E}_{\mathbb{Q}} [f(Y) - \mathbb{E}_{\mathbb{P}} [f(X)]] \right) - D(\mathbb{Q}\|\mathbb{P}) - \frac{\lambda^2 \nu^2}{2} \leq 0 \quad \text{for all } \lambda \in \mathbb{R} \text{ and } \mathbb{Q} \ll \mathbb{P}.$$

Taking the supremum of the left hand side yields the claim. ■

Lemma 3.24 (Transportation lemma) *Let* $\mathbb{P}$ *be a probability measure defined on a metric space* $(\mathcal{X}, d)$*. Then the following are equivalent:*

1. *Letting* $X \sim \mathbb{P}$*,* $f(X)$ *is* ν^2*-sub-Gaussian for every* $f \in \mathrm{Lip}\,(\mathcal{X}, d)$*.*
2. $W_1(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})}$ *for all probability measures* $\mathbb{Q} \ll \mathbb{P}$*.*

Proof: By the last lemma we can see that that the property 1 can be stated as

$$|\mathbb{E}_{\mathbb{Q}} [f(Y)] - \mathbb{E}_{\mathbb{P}} [f(X)]\,| \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all } \mathbb{Q} \ll \mathbb{P}. \quad \text{for all } f \in \mathrm{Lip}\,(\mathcal{X}, d) \text{ and } \mathbb{Q} \ll \mathbb{P}.$$

Taking the supremum of the left hand side with respect to $f \in \mathrm{Lip}\,(\mathcal{X}, d)$ yields that the above expression is equivalent to

$$W_1(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all } \mathbb{Q} \ll \mathbb{P},$$

which concludes the proof. ■

Exercise 3.25 *Lemma 3.24 works for f being 1-Lipschitz functions. What about general L-Lipschitz functions?*

Our first consequence of Lemma 3.24 is a useful inequality known as *Pinsker inequality*

Proposition 3.26 (Pinsker inequality) *Let $\mathbb{P}$ and $\mathbb{Q}$ are probability measures satisfying $\mathbb{Q} \ll \mathbb{P}$. Then*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})}. \tag{3.9}$$

Proof: First we have shown that f is 1-Lipschitz with respect to $d(x, y) = 1_{\{x \neq y\}}$ if and only if $|f(x) - f(y)| \leq 1$. Thus, for any $X \sim \mathbb{P}$, we known that $f(X)$ is in an interval of length bounded by 1. Consequently, $f(X)$ is $\frac{1}{4}$-sub-Gaussian. Therefore, applying Lemma 3.24 yields the result since we have already shown that $W_1(\mathbb{P}, \mathbb{Q}) = \|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}$ for the trivial metric. ■

Of course, if we could give an independent proof of Pinsker inequality (there are indeed direct proofs), we can use Lemma 3.24 to provide an alternative proof of the sub-Gaussian property of bounded variables (by taking $f = 1_{[a,b]}(x)$).

Exercise 3.27 *Let $X \sim \mathbb{P}$ be ν^2-sub-Gaussian. Show that $W_1(\mathbb{Q}, \mathbb{P}) \lesssim \sqrt{\nu^2 D(\mathbb{Q}\|\mathbb{P})}$.*

Due to Theorem 3.19, in order to show $f(X)$ is sub-Gaussian for $f \in \mathrm{Lip}\,(\mathcal{X}, d)$, it suffices to show that

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}, \mathbb{Q})} \mathbb{E}_\pi \left[d(X, Y) \right] \leq \sqrt{2\nu^2 D(\mathbb{Q}\|\mathbb{P})}. \tag{3.10}$$

Inequalities of this type are usually called *transportation (cost) inequalities* which play an key role in establishing some useful concentration inequalities that cannot be established by the previous methods.

3.2.2 Tensorization and Bounded Difference Inequality Revisited

Given metric spaces $(\mathcal{X}_k, d_k),\ k = 1, \cdots, n$, there are different ways to define a metric on $\bigotimes_{k=1}^n \mathcal{X}_k$, for example,

$$d_L(x, y) = \sum_{k=1}^n L_k d_k(x_k, y_k), \quad L_k > 0 \tag{3.11}$$

or

$$d_2(x, y) = \sqrt{\sum_{k=1}^n d_k(x_k, y_k)^2}. \tag{3.12}$$

Rather that considering the tensorizaton in a specific setting, the following theorem provides a general tensorization principle. The proof of the theorem is by induction and is omitted. Details of the proof can be found in [2] and [3].

Theorem 3.28 (Marton) *Let $\bigotimes_{k=1}^n \mathbb{P}_k$ be a product measure on a product measure space $\bigotimes_{k=1}^n \mathcal{X}_k$. Let $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function, and let $c_k : \mathcal{X}_k \times \mathcal{X}_k \to \mathbb{R}_+$ be positive weight function. Suppose that for $k = 1, \cdots, n$ and for every probability measure $\mathbb{Q}_k$ which is absolutely continuous with respect to $\mathbb{P}_k$,*

$$\inf_{\pi \in \mathcal{C}(\mathbb{P}_k, \mathbb{Q}_k)} \phi\left(\mathbb{E}_\pi\left[c_k(X_k, Y_k)\right]\right) \leq 2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k).$$

Then for any probability measure $\mathbb{Q}$ that is absolutely continuous with respect to $\bigotimes_{k=1}^n \mathbb{P}_k$, we have

$$\inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n \phi(\mathbb{E}_\pi\left[c_k(X_k, Y_k)\right]) \leq 2\nu^2 D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right). \tag{3.13}$$

Note that the left hand of (3.13) is not necessarily a W_1 distance. Thus, for different metric, we need to choose suitable ϕ and c_k in order to obtain a result associated with W_1 distance. The following proposition considers the $d_L(\cdot, \cdot)$ metric in (3.11).

Proposition 3.29 *Let $\bigotimes_{k=1}^n \mathbb{P}_k$ be a product measure on a product measure space $\bigotimes_{k=1}^n (\mathcal{X}_k, d_k)$. If for each univariate probability measure,*

$$W_1(\mathbb{Q}_k, \mathbb{P}_k) \leq \sqrt{2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k)} \quad \textit{for all } \mathbb{Q}_k \ll \mathbb{P}_k. \tag{3.14}$$

Then

$$W_1\left(\mathbb{Q}, \bigotimes_{k=1}^n \mathbb{P}_k\right) \leq \sqrt{2\nu^2 \left(\sum_{k=1}^n L_k^2\right) D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right)} \quad \textit{for all } \mathbb{Q} \ll \bigotimes_{k=1}^n \mathbb{P}_k,$$

where the W_1 is defined using the distance $d_L(x, y) = \sum_{k=1}^n L_k d_k(x_k, y_k)$. Hence, for any 1-Lipschitz function $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ under this metric, $f(X)$ is sub-Gaussian if $X \sim \bigotimes_{k=1}^n \mathbb{P}_k$.

Proof: First we have

$$\begin{aligned}
W_1\left(\mathbb{Q}, \bigotimes_{k=1}^n \mathbb{P}_k\right) &= \inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \mathbb{E}_\pi\left[\sum_{k=1}^n L_k d_k(X_k, Y_k)\right] \\
&= \inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n L_k \mathbb{E}_\pi\left[d_k(X_k, Y_k)\right] \\
&\leq \sqrt{\sum_{k=1}^n L_k^2} \sqrt{\inf_{\pi \in \mathcal{C}(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q})} \sum_{k=1}^n \left(\mathbb{E}_\pi\left[d_k(X_k, Y_k)\right]\right)^2}.
\end{aligned}$$

Noting that the assumption implies

$$W_1(\mathbb{Q}_k, \mathbb{P}_k)^2 = \inf_{\pi \in \mathcal{C}(\mathbb{P}_k, \mathbb{Q}_k)} \left(\mathbb{E}_\pi\left[d_k(X_k, Y_k)\right]\right)^2 \leq 2\nu^2 D(\mathbb{Q}_k \| \mathbb{P}_k),$$

the claim follows immediately from Theorem 3.28 by taking $\phi(x) = x^2$ and $c_k(\cdot, \cdot) = d_k(\cdot, \cdot)$. ∎

Example 3.30 (Bounded difference inequality revisited) *For any $X_k \sim \mathbb{P}_k$, under the metric $d_k(x_k, y_k) = 1_{\{x_k \neq y_k\}}$, the Pinsker inequality implies that*

$$W_1(\mathbb{Q}_k, \mathbb{P}_k) \leq \sqrt{\frac{1}{2} D(\mathbb{Q}_k \| \mathbb{P}_k)}.$$

Thus, for f being 1-Lipschitz under the metric $d_L(x, y) = \sum_{k=1}^n L_k 1_{\{x_k \neq y_k\}}$ (equivalent to the bounded difference property as mentioned at the beginning of this lecture), the application of Proposition 3.29 yields that

$$f(X_1, \cdots, X_k) \text{ is } \tfrac{\sum_{k=1}^n L_k^2}{4}\text{-sub-Gaussian},$$

which recovers the bounded difference inequality. More precisely, we have

$$\inf_{\pi \in \mathcal{C}\left(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q}\right)} \sum_{k=1}^n \left(\mathbb{E}_\pi \left[d_k(X_k, Y_k)\right]\right)^2 = \inf_{\pi \in \mathcal{C}\left(\bigotimes_{k=1}^n \mathbb{P}_k, \mathbb{Q}\right)} \sum_{k=1}^n \left(\pi \left[X_k \neq Y_k\right]\right)^2 \leq \frac{1}{2} D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right). \tag{3.15}$$

3.3 Talagrand Concentration Inequality

Up to this point, the transportation method did not yield any new results yet. In this section we will use a type of asymmetric transportation cost inequalities based on a one-sided variant of the trivial metric to establish the following remarkable Talagrand concentration inequality.

Theorem 3.31 (Talagrand) *Let $f : \bigotimes_{k=1}^n \mathcal{X}_k \to \mathbb{R}$ be a function satisfying*

$$f(y) - f(x) \leq \sum_{k=1}^n a_k(y) 1_{\{x_k \neq y_k\}} \quad \text{for all } x, y.$$

Assume $\sup_y \left(\sum_{k=1}^n a_k^2(y)\right) \leq \nu^2$. Then, for any independent random variables $X = (X_1, \cdots, X_n)$ taking values in $\bigotimes_{k=1}^n \mathcal{X}_k$, $f(X)$ is ν^2-sub-Gaussian.

Proof: Assume $X \sim \bigotimes_{k=1}^n \mathbb{P}_k$. By Lemma 3.23, we need to show that

$$\left|\mathbb{E}_Q \left[f(Y)\right] - \mathbb{E}_{\bigotimes_{k=1}^n \mathbb{P}_k} \left[f(X)\right]\right| \leq \sqrt{2\nu^2 D\left(\mathbb{Q} \| \bigotimes_{k=1}^n \mathbb{P}_k\right)}. \tag{3.16}$$

Letting $\pi \in \mathcal{C}\left(\mathbb{Q}, \mathbb{P}\right)$, a simple calculation yields that (again the goal is to reduce the problem about the concentration of f to the problem of comparing the divergences of two probability measures)

$$\begin{aligned}
\mathbb{E}_Q \left[f(Y)\right] - \mathbb{E}_{\bigotimes_{k=1}^n \mathbb{P}_k} \left[f(X)\right] &= \mathbb{E}_\pi \left[f(Y) - f(X)\right] \\
&\leq \mathbb{E}_\pi \left[\sum_{k=1}^n c_k(Y) 1_{\{X_k \neq Y_k\}}\right] \\
&= \sum_{k=1}^n \mathbb{E}_\pi \left[c_k(Y) 1_{\{X_k \neq Y_k\}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \mathbb{E}_\pi \left[c_k(Y) \pi \left[X_k \neq Y_k | Y \right] \right] \\
&\leq \sum_{k=1}^{n} \left(\mathbb{E}_\pi \left[(c_k(Y))^2 \right] \right)^{1/2} \left(\mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2} \\
&\leq \left(\mathbb{E}_\pi \left[\sum_{k=1}^{n} (c_k(Y))^2 \right] \right)^{1/2} \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2} \\
&\leq \nu \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right] \right)^{1/2}.
\end{aligned}
$$

Similarly, we have (**check this!**)

$$
\mathbb{E}_{\bigotimes_{k=1}^{n} \mathbb{P}_k} \left[f(X) \right] - \mathbb{E}_Q \left[f(Y) \right] \leq \nu \left(\sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | X \right] \right)^2 \right] \right)^{1/2}.
$$

Therefore, in order to show (3.16), it suffices to show (since π is arbitrary above)

$$
\max \left\{ \inf_{\pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | Y \right] \right)^2 \right], \inf_{\pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \sum_{k=1}^{n} \mathbb{E}_\pi \left[\left(\pi \left[X_k \neq Y_k | X \right] \right)^2 \right] \right\} \leq 2D \left(\mathbb{Q} \| \bigotimes_{k=1}^{n} \mathbb{P}_k \right), \tag{3.17}
$$

which actually holds. Thus the proof is complete. ■

Remark 3.32 *Here we have used* (3.17) *without proof. Note that* (3.17) *can be viewed as an asymmetric version of* (3.15)*. The details of proof can be found in [2] and [3], which uses a conditional version of Theorem 3.28 for tensorization.*

Corollary 3.33 (Talagrand) *Let $X = (X_1, \cdots, X_n)$ be a vector of independent random variables taking values in $[a, b]$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-Lipschitz with respect to the Euclidean norm. Then, $f(X)$ is $L^2(b-a)^2$-sub-Gaussian.*

Proof: The first order condition for convexity implies

$$
f(y) - f(x) \leq \nabla f(y)^\top (y - x) \quad \text{for all } x, y.
$$

Since $|x_k - y_k| \leq (b - a)$, we have

$$
\begin{aligned}
f(y) - f(x) &\leq \sum_{k=1}^{n} \partial_k f(y)(y_k - x_k) \\
&\leq \sum_{k=1}^{n} (b-a) \left| \partial_k f(y) \right| 1_{\{x_k \neq y_k\}}.
\end{aligned}
$$

The result follows immediately from Theorem 3.31 by further noting that $\|\nabla f(y)\|_2 \leq L$. ■

Remark 3.34 *a) Only upper tail can be established in this scenario based on the entropy method in Lecture 2. b) The convexity property of f is indispensable to establish the sub-Gaussian property. There exists nonconvex 1-Lipschitz functions Corollary 3.33 even fails for symmetric Bernoulli variables, see for example Problem 4.9 of [2].*

Example 3.35 (Rademacher complexity revisited) *In Example 2.22 of Lecture 2, we have established upper tail bound of the Rademacher complexity of a set $\mathcal{A}$ in terms of the width of the set $\sup_{a\in\mathcal{A}}\|a\|_2$. In the example we actually show that the function $f(x_1,\cdots,x_n)=\sup_{a\in\mathcal{A}}[\sum_{k=1}^n a_k x_k]$ is convex and $\sup_{a\in\mathcal{A}}\|a\|_2$-Lipschitz continuous. Thus, it follows from Corollary 3.33 that the Rademacher complexity $\sup_{a\in\mathcal{A}}[\sum_{k=1}^n a_k\varepsilon_k]$ is $4\sup_{a\in\mathcal{A}}\|a\|_2^2$ and hence*

$$\mathbb{P}\left[\left|\sup_{a\in\mathcal{A}}\left[\sum_{k=1}^n a_k\varepsilon_k\right]-\mathbb{E}\left[\sup_{a\in\mathcal{A}}\left|\sum_{k=1}^n a_k\varepsilon_k\right|\right]\right|\geq t\right]\leq 2\exp\left(-\frac{t^2}{8\sup_{a\in\mathcal{A}}\|a\|_2^2}\right).$$

3.4 Gaussian Concentration Revisited

So far, we have established certain concentration inequalities (i.e., sub-Gaussian properties) for Lipschitz functions under the ℓ_1-type metric,

$$f(y)-f(x)\leq L\sum_{k=1}^n d_k(x_k,y_k)\quad\text{or}\quad f(y)-f(x)\leq\sum_{k=1}^n c(y)d_k(x_k,y_k). \tag{3.18}$$

However, we have already seen that independent Gaussian random variables exhibits dimension free concentration for Lipschitz functions under the ℓ_2 metric,

$$f(y)-f(x)\leq L\sqrt{\sum_{k=1}^n[d_k(x_k,y_k)]^2}. \tag{3.19}$$

That is if f satisfies (3.19) and X is a vector of i.i.d standard Gaussian random variables, then $f(X)$ is L^2-sub-Gaussian.

Since ℓ_2-metric is less than the ℓ_1-metric, if f satisfies (3.19), it naturally satisfies the first inequality in (3.18). Lets first attempt to use Proposition 3.29 to establish the Gaussian concentration. Since for each univariate standard Gaussian random variables, we have $W_1(\mathbb{Q}_k,\mathbb{P}_k)\lesssim\sqrt{D(\mathbb{Q}_k\|\mathbb{P}_k)}$ (see Exercise 3.27), by Proposition 3.29 we can conclude that for any Lipschitz functions satisfying (3.19), $f(X)$ is nL^2-sub-Gaussian. *However, it is much weaker than the Gaussian concentration we have previously seen which does have the multiple factor n.* Thus, we need a direct route for the concentration of ℓ_2-Lipschitz functions.

Let $\mathcal{X}\sim\bigotimes_{k=1}^n\mathbb{P}_k$ and assume f is 1-Lipschitz under the ℓ_2-metric. Then, in order to establish the sub-Gaussian property of $f(X)$, by the concentration lemma and the Monge-Kantorovich duality, we need to show that for any $\mathbb{Q}\ll\bigotimes_{k=1}^n\mathbb{P}_k$,

$$\begin{aligned}W_1\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)&=\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)}\mathbb{E}_\pi\left[\sqrt{\sum_{k=1}^n[d_k(X_k,Y_k)]^2}\right]\\&\leq\sqrt{2\nu^2D\left(\mathbb{Q}\|\bigotimes_{k=1}^n\mathbb{P}_k\right)}.\end{aligned} \tag{3.20}$$

Since by Jensen's inequality ($\sqrt{z}$ is concave) we evidently have

$$\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\right)}\mathbb{E}_\pi\left[\sqrt{\sum_{k=1}^n\left[d_k(X_k,Y_k)\right]^2}\right]\leq\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)}\sqrt{\mathbb{E}_\pi\left[\sum_{k=1}^n\left[d_k(X_k,Y_k)\right]^2\right]}.$$

Thus it suffices to show

$$\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)}\sqrt{\mathbb{E}_\pi\left[\sum_{k=1}^n\left[d_k(X_k,Y_k)\right]^2\right]}\leq\sqrt{2\nu^2D\left(\mathbb{Q}\|\bigotimes_{k=1}^n\mathbb{P}_k\right)},$$

or equivalently

$$\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)}\mathbb{E}_\pi\left[\sum_{k=1}^n\left[d_k(X_k,Y_k)\right]^2\right]\leq2\nu^2D\left(\mathbb{Q}\|\bigotimes_{k=1}^n\mathbb{P}_k\right).\tag{3.21}$$

That is, what we need is a characterization based on W_2 distance instead of W_1 distance. It turns out that if (3.21) is satisfied each univariate $\mathbb{P}_k$, then it is also satisfied for the product measure $\bigotimes_{k=1}^n\mathbb{P}_k$.

Theorem 3.36 *Let $\bigotimes_{k=1}^n\mathbb{P}_k$ be a product probability measure on $\bigotimes_{k=1}^n(\mathcal{X}_k,d_k)$. Assume*

$$\inf_{\pi\in\mathcal{C}(\mathbb{Q}_k,\mathbb{P}_k)}\mathbb{E}_\pi\left[\left[d_k(X_k,Y_k)\right]^2\right]\leq2\nu^2D\left(\mathbb{Q}_k\|\mathbb{P}_k\right)\quad\textit{for all }\mathbb{Q}_k\ll\mathbb{P}_k.\tag{3.22}$$

holds for each k. Then we have

$$\inf_{\pi\in\mathcal{C}\left(\mathbb{Q},\bigotimes_{k=1}^n\mathbb{P}_k\right)}\mathbb{E}_\pi\left[\sum_{k=1}^n\left[d_k(X_k,Y_k)\right]^2\right]\leq2\nu^2D\left(\mathbb{Q}\|\bigotimes_{k=1}^n\mathbb{P}_k\right)\quad\textit{for all }\mathbb{Q}\ll\bigotimes_{k=1}^n\mathbb{P}_k.$$

Proof: Apply Theorem 3.28 with $\phi(x)=x$ and $c_k(x_k,y_k)=\left[d_k(x_k,y_k)\right]^2$. ■

Remark 3.37 *a) It can be shown that the standard Gaussian distribution satisfies* (3.22) *(see for example [3]). Thus, we can establish the dimension free Gaussian concentration based on the transportation method from the above analysis, which has been established by the Herbst argument and Gaussian log-Sobolev inequality in Lecture 2. b)* (3.22) *cannot be established for arbitrary sub-Gaussian random variables (otherwise it will contradicts with the counter example for Corollary 3.33). c) In fact it turns out* (3.22) *is not only sufficient, but also necessary for establishing* (3.20)*, i.e., for establishing the dimension-free concentration result, see for example [2].*

3.5 Short Summary

We have discussed three methods for establishing concentration inequalities of functions of independent random variables:

- Chernoff method (Lecture 1)
- Entropy method (Lecture 2)

- Transportation method (Lecture 3).

In both the entropy method and the transportation method, the variational formulations (duality in the transportation method) play an important role in showing the tensorization property. Here are a list of concentration inequalities we have presented:

- Hoeffding inequality (Lecture 1, for sum of independent sub-Gaussian random variables)
- Bernstein equality (Lecture 1, for sum of independent sub-exponential random variables)
- Bounded difference inequality (Lecture 1, for function obeying bounded difference property)
- General bounded difference inequality (Lecture 2, for function obeying asymmetric bounded difference property)
- Gaussian concentration inequality (Lecture 2 and 3, Lipschitz function concentration of Gaussian random variables)
- Talagrand inequality (Lecture 3, Lipschitz and convex function concentration of bounded random variables)

In addition, there is another method – geometric method based on isoperimetric inequalities we did not cover. This method works for certain types of probability measures such as Gaussian measure and uniform measure on the sphere.

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 3.3.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 4.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Chapters 8.

[4] Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Chapter 2, 8.