Probability Theory

IX Metrics on Probability Measures

Seongho Joo

Bounded Lipschitz functions

Bounded Lipschitz functions

- For a function $f : S \to \mathsf{r}$, let us define a Lipschitz semi-norm by
$$\|f\|_{\boldsymbol{L}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$$
Clearly, $\|f\|_{\boldsymbol{L}} = 0$ if and only if f is constant so $\|f\|_{\boldsymbol{L}}$ is not a norm, even though it satisfies the triangle inequality.
Let us define a *bounded Lipschitz norm* by
$$\|f\|_{\mathsf{BL}} = \|f\|_{\boldsymbol{L}} + \|f\|_{\infty} \tag{1.1}$$
where $\|f\|_{\infty} = \sup_{s \in S} |f(s)|$. Let
$$BL(S,d) = \left\{ f : S \to \mathsf{r} \mid \|f\|_{\mathsf{BL}} < \infty \right\}$$
be the set of all bounded Lipschitz functions on (S,d). We will now prove several facts about these functions.

Bounded Lipschitz functions

- For a function $f : S \to$ r, let us define a Lipschitz semi-norm by

$$\|f\|_{\boldsymbol{L}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$$

Clearly, $\|f\|_{\boldsymbol{L}} = 0$ if and only if f is constant so $\|f\|_{\boldsymbol{L}}$ is not a norm, even though it satisfies the triangle inequality.
Let us define a *bounded Lipschitz norm* by

$$\|f\|_{\mathsf{BL}} = \|f\|_{\boldsymbol{L}} + \|f\|_{\infty} \tag{1.1}$$

where $\|f\|_{\infty} = \sup_{s \in S} |f(s)|$. Let

$$BL(S,d) = \left\{ f : S \to \mathrm{r} \mid \|f\|_{\mathsf{BL}} < \infty \right\}$$

be the set of all bounded Lipschitz functions on (S,d). We will now prove several facts about these functions.

Lemma 1

If $f, g \in BL(S,d)$ then $fg \in BL(s,d)$ and $\|fg\|_{\mathsf{BL}} \leq \|f\|_{\mathsf{BL}} \|g\|_{\mathsf{BL}}$.

Bounded Lipschitz functions

Notation. Let $* = \wedge$ or $\vee$.

Lemma 2

The following inequalities hold:

$$\|f_1 * \cdots * f_k\|_{\mathsf{L}} \leq \max_{1 \leq i \leq k} \|f_i\|_{\mathsf{L}} ,$$

$$\|f_1 * \cdots * f_k\|_{\mathsf{BL}} \leq 2 \max_{1 \leq i \leq k} \|f_i\|_{\mathsf{BL}} .$$

Another important fact is the following.

Bounded Lipschitz functions

Notation. Let $* = \wedge$ or $\vee$.

Lemma 2

The following inequalities hold:

$$\|f_1 * \cdots * f_k\|_{\mathsf{L}} \leq \max_{1\leq i\leq k} \|f_i\|_{\mathsf{L}}\,,$$

$$\|f_1 * \cdots * f_k\|_{\mathsf{BL}} \leq 2 \max_{1\leq i\leq k} \|f_i\|_{\mathsf{BL}}\,.$$

Another important fact is the following.

Theorem 1 (Extension theorem)

Given a set $A \subseteq S$ and a bounded Lipschitz function $f \in BL(A, d)$ on A, there exists an extension $h \in BL(S, d)$ such that $f = h$ on A and $\|h\|_{\mathsf{BL}} = \|f\|_{\mathsf{BL}}$.

Bounded Lipschitz functions

To prove the next property of bounded Lipschitz functions, let us first recall the following famous generalization of the Weierstras theorem.

Theorem 2 (Stone-Weierstrass)

Let (S, d) be a compact metric space and $\mathcal{F} \subseteq C(S)$ is such that

1. $\mathcal{F}$ is algebra, i.e. for all $f, g \in \mathcal{F}$, $c \in$ r, we have $cf + g \in \mathcal{F}, fg \in \mathcal{F}$.
2. $\mathcal{F}$ separates points, i.e. $x \neq y \in S$ then there exists $f \in \mathcal{F}$ such that $f(x) \neq f(y)$.
3. $\mathcal{F}$ contains constants.

Then $\mathcal{F}$ is dense in $(C(S), d_\infty)$.

Corollary 1

If (S, d) is a compact space then the set of bounded Lipschiz functions $BL(S, d)$ is dense in $(C(S), d_\infty)$.

Bounded Lipschitz functions

- We will also need another well-known result from analysis. A set $A \subseteq S$ is *totally bounded* if for any $\varepsilon > 0$ there exists a finite ε-cover of A, i.e. a set of points $a_1, \ldots a_N$ such that $A \subseteq \bigcup_{i \leq N} B(a_i, \varepsilon)$, where $B(a, \varepsilon)$ is a ball of radius ε centered at a.

Theorem 3 (Arzela-Ascoli)

If (S, d) is a compact metric space then a subset $\mathcal{F} \subseteq C(S)$ is totally bounded in d_∞ metric if and only if $\mathcal{F}$ is equicontinuous and uniformly bounded.

The following fact was used in the proof of the Selection Theorem.

Bounded Lipschitz functions

- We will also need another well-known result from analysis. A set $A \subseteq S$ is *totally bounded* if for any $\varepsilon > 0$ there exists a finite ε-cover of A, i.e. a set of points $a_1, \ldots a_N$ such that $A \subseteq \bigcup_{i \leq N} B(a_i, \varepsilon)$, where $B(a, \varepsilon)$ is a ball of radius ε centered at a.

Theorem 3 (Arzela-Ascoli)

If (S, d) is a compact metric space then a subset $\mathcal{F} \subseteq C(S)$ is totally bounded in d_∞ metric if and only if $\mathcal{F}$ is equicontinuous and uniformly bounded.

The following fact was used in the proof of the Selection Theorem.

Corollary 2

If (S, d) is a compact space then $C(S)$ is separable in d_∞.

Convergence of laws on metric spaces

Convergence of empirical measures

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X_1, X_2 \cdots : \Omega \to S$ be an i.i.d sequence of random variables with values in a metric space (S, d). Let μ be the law of X_i on S. Let us define the random empirical measures on μ_n on the Borel σ-algebra on the Borel σ-algebra $\mathcal{B}$ on S by

$$\mu_n(A)(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i(w) \in A), \ A \in \mathcal{B}$$

By the strong law of large numbers, for any $f \in C_b(S)$,

$$\int f \, \mathrm{d}\mu_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \mathbb{E} \, f(X_1) = \int f \, \mathrm{d}\mu \ a.s.$$

However, the set of measures zero where this convergence is violated *depends on* f and it is not right away clear that the convergence holds for all $f \in C_b(S)$ with probability one. We will need the following lemma:

Lemma 3

If (S, d) is separable then there exists a metric e on S such that (S, e) it totally bounded, and e and d define the same topology, i.e. $e(s_n, s) \to 0$ if and only if $d(s_n, s) \to 0$.

Convergence of empirical measures

Theorem 4 (Varadarajan)

Let (S, d) be a separable metric space. Then μ_n converges to μ weakly almost surely,

$$\mathbb{P}(w : \mu_n(\cdot)(w) \to \mu \text{ weakly }) = 1.$$

Convergence of laws on metric spaces

- Next, we will introduce two metrics on the set of all probability measures on (S, d) with the Borel σ-algebra $\mathcal{B}$ and, under some mild conditions, prove that they metrize the weak convergence. For a set $A \subseteq S$, let us denote by

$$A^{\varepsilon} = \{y \in S \,|\, d(x, y) < \varepsilon \text{ for some } x \in A\}$$

its open ε-neighborhood. If $\mathbb{P}$ and $\mathbb{Q}$ are probability distributions on S then

$$\rho(\mathbb{P}, \mathbb{Q}) = \inf\{\varepsilon > 0 \,|\, \mathbb{P}(A) \leq \mathbb{Q}(A^{\varepsilon}) + \varepsilon \text{ for all } A \in \mathcal{B}\}$$

is called the *Levy-Prohorov* distance between $\mathbb{P}$ and $\mathbb{Q}$.

Lemma 4

ρ is a metric on the set of probability laws on $\mathcal{B}$.

Convergence of laws on metric spaces

- Given probability distributions $\mathbb{P}, \mathbb{Q}$ on the metric space (S, d), we define the *bounded Lipschitz distance* between them by

$$\beta(\mathbb{P}, \mathbb{Q}) = \sup \left\{ \left| \int f \, \mathrm{d}\mathbb{P} - \int f \, \mathrm{d}\mathbb{Q} \right| \mid \|f\|_{\mathsf{BL}} \leq 1 \right\}$$

Lemma 5

β is a metric on the set of probability laws on $\mathcal{B}$.

Convergence of laws on metric spaces

• Let us now show that on separable metric space, the metric ρ and β metrize weak convergence. Before We prove this, let us recall the statement of Ulam's theorem. Namely, every probability law $\mathbb{P}$ on a complete separable metric space (S, d) is tight, which means that for any $\varepsilon > 0$ there exists a compact $K \subseteq S$ such that $\mathbb{P}(S \setminus K) \leq \varepsilon$.

Theorem 5

If (S, d) is separable or $\mathbb{P}$ is tight then the following are equivalent:

1. $\mathbb{P}_n \to \mathbb{P}$.
2. For all $f \in BL(S, d)$, $\int f \, \mathrm{d}\, \mathbb{P}_n \to \int f \, \mathrm{d}\, \mathbb{P}$.
3. $\beta(\mathbb{P}_n, \mathbb{P}) \to 0$.
4. $\rho(\mathbb{P}_n, \mathbb{P}) \to 0$.

Convergence and uniform tightness

Next, we will make a connection between the above metrics and uniform tightness. First, we will show that, in some case, uniform tightness is necessary for the convergence of laws.

Theorem 6

If $\mathbb{P}_n \to \mathbb{P}_0$ weaky and each $\mathbb{P}_n$ is tight for $n \geq 0$, then $(\mathbb{P}_n)_{n \geq 0}$ is uniformly tight.

In particular, by Ulam's theorem, any convergent sequence of laws on a complete separable metric space is uniformly tight.

Convergence and uniform tightness

- Next, on complete separable metric spaces, we will complement the Selection Theorem by showing how uniform tightness can be expressed in the above metrics.

Theorem 7

Let (S, d) be a complete separable metric space and $\mathscr{P}$ be a subset of probability laws on S. Then the following are equivalent.

1. $\mathscr{P}$ is uniformly tight.
2. For any sequence $\mathbb{P}_n \in \mathscr{P}$ there exists a converging subsequence $\mathbb{P}_{n(k)} \to \mathbb{P}$ where $\mathbb{P}$ is a law on S.
3. $\mathscr{P}$ has the compact closure on the space of probability laws equipped with the Levy-Prohorov or bounded Lipschitz metrics ρ or β.
4. $\mathscr{P}$ is totally bounded with respect to ρ or β.

Convergence of laws on metric spaces

Theorem 8 (Prokohorov)

The set of probability laws on a complete separable metric space is complete with respect to the metrics ρ and β.

Strassen's theorem. Relationships between metrics

Metric for convergence in probability

- Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, (S, d) - a separable metric space and $X, Y : \Omega \to S$ - random variables with values in S. The quantity

$$\alpha(X, Y) = \inf \{\varepsilon > 0 \,|\, \mathbb{P}(d(x, y) > \varepsilon) \leq \varepsilon)\}$$

is called the *Ky Fan* metric on the set $\mathcal{L}^0(\Omega, S)$ of classes of equivalence of such random variables.

Metric for convergence in probability

- Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, (S, d) - a separable metric space and $X, Y : \Omega \to S$ - random variables with values in S. The quantity

$$\alpha(X, Y) = \inf \{\varepsilon > 0 \,|\, \mathbb{P}(d(x, y) > \varepsilon) \leq \varepsilon)\}$$

is called the *Ky Fan* metric on the set $\mathcal{L}^0(\Omega, S)$ of classes of equivalence of such random variables.

Lemma 6

The *Ky Fan* metric α on $\mathcal{L}^0(\Omega, S)$ metrizes convergence in probability.

Lemma 7

For $X, Y \in \mathcal{L}^0(\Omega, S)$, the Levy-Prohorov metric ρ satisfies

$$\rho(\mathcal{L}(X), \mathcal{L}(Y)) \leq \alpha(X, Y).$$

Metric for convergence in probability

Question: Can we construct random variables s_1 and s_2 with laws $\mathbb{P}$ and $\mathbb{Q}$, that are defined on the same probability space and are close to each other in the Ky Fan metric α?

The following result will be a key tool in the proof of the main result of this section. Consider two sets X and Y. Given a subset $K \subseteq X \times Y$ and $A \subseteq X$ we define a K-image of A by

$$A^K = \{y \in Y \mid \exists x \in A, (x, y) \in K\}.$$

A K-matching f of X into Y is one-to-one function (injection) $f : X \to Y$ such that $(x, f(X)) \in K$. We will need the following well-known matching theorem.

Theorem 9 (Hall's marriage theorem)

If X, Y are finite and, for all $A \subseteq X$,

$$\text{card}(A^K) \geq \text{card}(A)$$

then there exists a K-matching f of X into Y.

Strassen's theorem. Relationships between metrics

Theorem 10 (Strassen)

Suppose that (S, d) is a separable metric space and $\alpha, \beta > 0$. Suppose the laws $\mathbb{P}$ and $\mathbb{Q}$ are such that, for all measurable sets $F \subseteq S$,

$$\mathbb{P}(F) \leq \mathbb{Q}(F^{\alpha}) + \beta \tag{3.1}$$

Then for any $\varepsilon > 0$ there exist two non-negative measures η, γ on $S \times S$ such that

1. $\mu = \eta + \gamma$ is a law on $S \times S$ with marginals $\mathbb{P}$ and $\mathbb{Q}$.
2. $\eta(d(x, y) > \alpha + \varepsilon) = 0$.
3. $\gamma(S \times S) \leq \beta + \varepsilon$.
4. μ is a finite sum of product measures.

Remark. In the above statement, it is enough to assume that (3.1) holds only for closed sets or only for open sets F; moreover, one can replace an open α-neighbourhood $F^{\alpha} = \{s \in S \,|\, d(s, F) < \alpha\}$ by a closed $\alpha-$ neighborhood $F^{\alpha]}$. This is because the set F^{ε} is open and $F^{\varepsilon]}$ is closed, and, for example the condition (3.1) for closed set implies

$$\mathbb{P}(F) \leq \mathbb{P}(F^{\varepsilon]}) \leq \mathbb{Q}((F^{\varepsilon]})^{\alpha}) + \beta \leq \mathbb{Q}(F^{\alpha+2\varepsilon}) + \beta$$

for all measurables sets F, which simply replaces α by $\alpha + 2\varepsilon$ in (3.1).

Strassen's theorem. Relationships between metrics

The following relationship between Ky Fan and Levy-Prohoorv metrics is an immediate consequence of Strassen's theorem.

Theorem 11

If (S, d) is a separable metric space and $\mathbb{P}, \mathbb{Q}$ are laws on S then, for any $\varepsilon > 0$, there exists random variables X and Y on the same probability space with the distributions $\mathcal{L}(X) = \mathbb{P}$ and $\mathcal{L}(Y) = \mathbb{Q}$ such that

$$\alpha(X, Y) \leq \rho(\mathbb{P}, \mathbb{Q}) + \varepsilon \tag{3.2}$$

If $\mathbb{P}, \mathbb{Q}$ are tight, one can take $\varepsilon = 0$.

Strassen's theorem. Relationships between metrics

- There is also relationship between the bounded Lipschitz metric β and Levy-Prohorov metic ρ.

Lemma 8

If (S, d) is a separable metric space then

$$\beta(\mathbb{P}, \mathbb{Q}) \leq 2\rho(\mathbb{P}, \mathbb{Q}) \leq 4\sqrt{\beta(\mathbb{P}, \mathbb{Q})}$$

Wasserstein distance and Kantorovich-Rubinstein theorem

Wasserstein distance and Kantorovich-Rubinstein theorem

- Let (S,d) be separable metric space. Denote by $\mathscr{P}_1$ the set of all laws on S such that for some $z \in S$

$$\int_S d(x,z)\,\mathrm{d}\,\mathbb{P}(x) < +\infty$$

Let us consider a set

$$M(\mathbb{P},\mathbb{Q}) = \{\mu \,|\, \mu \text{ is a law on } S \times S \text{ with marginals } \mathbb{P} \text{ and } Q\}$$

For $\mathbb{P},\mathbb{Q} \in \mathscr{P}_1$, the qunaity

$$W(\mathbb{P},\mathbb{Q}) = \inf\left\{\int d(x,y)\,\mathrm{d}\mu(x,y) \,|\, \mu \in M(\mathbb{P},\mathbb{Q})\right\}$$

is called the *Wasserstein distance* between $\mathbb{P}$ and $\mathbb{Q}$. If $\mathbb{P}$ and $\mathbb{Q}$ are tight, this infimum is attained.

Wasserstein distance and Kantorovich-Rubinstein theorem

Given any two laws $\mathbb{P}$ and $\mathbb{Q}$ on S, let us define

$$\gamma(\mathbb{P},\mathbb{Q}) = \sup\left\{\left|\int f \,\mathrm{d}\,\mathbb{P} - \int f \,\mathrm{d}\mathbb{Q}\right| \mid \|f\|_{\mathsf{L}} \leq 1\right\}$$

and

$$m_d(\mathbb{P},\mathbb{Q}) = \sup\left\{\int f \,\mathrm{d}\,\mathbb{P} + \int g \,\mathrm{d}\mathbb{Q} \mid f,g \in C(S), f(x)+g(x) < d(x,y)\right\}$$

Notice that, for $\mathbb{P},\mathbb{Q} \in \mathscr{P}_1(S)$, both $\gamma(\mathbb{P},\mathbb{Q}), m_d(\mathbb{P},\mathbb{Q}) < \infty$. Let us show that these two quantities are equal.

Lemma 9

We have $\gamma(\mathbb{P},\mathbb{Q}) = m_d(\mathbb{P},\mathbb{Q})$.

Wasserstein distance and Kantorovich-Rubinstein theorem

Below, we will need the following version of the Hahn-Banach theorem.

Theorem 12 (Hahn-Banach)

Let V be a normed vector space, E- a linear subspace of V and U- an open convex set in V such that $U \cap E \neq \emptyset$. If $r : E \to$ r is a linear non-zero functional on E then there exists a linear functional $\rho : V \to$ r such that $\rho|_E = r$ and $\sup_U \rho(x) = \sup_{U \cap E} r(x)$.

Using this, we will prove the following Kantorovich-Rubinstein theorem for compact metric spaces.

Wasserstein distance and Kantorovich-Rubinstein theorem

Below, we will need the following version of the Hahn-Banach theorem.

Theorem 12 (Hahn-Banach)

Let V be a normed vector space, E- a linear subspace of V and U- an open convex set in V such that $U \cap E \neq \emptyset$. If $r : E \to$ r is a linear non-zero functional on E then there exists a linear functional $\rho : V \to$ r such that $\rho|_E = r$ and $\sup_U \rho(x) = \sup_{U \cap E} r(x)$.

Using this, we will prove the following Kantorovich-Rubinstein theorem for compact metric spaces.

Theorem 13

If S is a compact metric space then $W(\mathbb{P}, \mathbb{Q}) = m_d(\mathbb{P}, \mathbb{Q}) = \gamma(\mathbb{P}, \mathbb{Q})$ for $\mathbb{P}, \mathbb{Q} \in \mathscr{P}_1$.

Wasserstein distance and entropy

Wasserstein distance and entropy

- In this section we will make several connections between the Wasserstein distance and other classical objects, with application to Gaussian concentration

Theorem 14 (Brunn-Minkowski inequality on $\mathbb{R}$)

If γ is the Lebesgue measure and A, B are two non-empty Borel sets on $\mathbb{R}$ then $\gamma(A+B) \geq \gamma(A) + \gamma(B)$, where $A + B\,\{a+b \,|\, a \in A, b \in B\}$.

Using this, we will prove another classical inequality.

Theorem 15 (Prekopa-Leindler inequality)

Consider nonnegative integrable functions $w, u, v \to \mathbb{R}^n \to [0, \infty)$ such that for some $\lambda \in [0, 1]$,

$$w(\lambda x + (1-\lambda)y) \geq u(x)^{\lambda} v(y)^{1-\lambda} \quad \text{for all } x, y \in \mathbb{R}^n$$

Then,

$$\int w \,\mathrm{d}x \geq \left(\int u \,\mathrm{d}x\right)^{\lambda} \left(\int v \,\mathrm{d}x\right)^{1-\lambda}$$

Cont.

Using the Prekopa-Leindler inequality one can prove that the Lebesgue measure γ on r^n satisfies the Brunn-Minkowski inequality

$$\gamma(A)^{1/n} + \gamma(B)^{1/B} \leq \gamma(A+B)^{1/n}. \tag{5.1}$$

From this one can easily deduce the famous isoperimetric property of Euclidean balls with respect to the Lebesgue measure. If B is a unit open ball in $\mathbb{R}^d$ and $\gamma(A) = \gamma(B)$ then, by (5.1),

$$\begin{aligned}\gamma(A^\varepsilon) &= \gamma(A+\varepsilon B) \geq (\gamma(A)^{1/n} + \gamma(\varepsilon B)^{1/n})^n \\ &= (\gamma(B)^{1/n} + \gamma(\varepsilon B)^{1/n})^n = (1+\varepsilon)^n \gamma(B) = \gamma(B^\varepsilon)\end{aligned}$$

In other words, volume of A grows faster than B as we expand the sets, which means that the surface area of B is smaller.

Wasserstein distance and entropy

Entropy and the Kullback-Leibler divergence. Consider a probability measure $\mathbb{P}$ on some measurable space and a nonnegative function $u : \Omega \to \mathbb{R}_+$. We define the entropy of u with respect to $\mathbb{P}$ by

$$\mathsf{Ent}_{\mathbb{P}}(u) = \int u \log u \, \mathrm{d}\,\mathbb{P} - \int u \, \mathrm{d}\,\mathbb{P} \cdot \log \int u \, \mathrm{d}\,\mathbb{P},$$

Notice that $\mathsf{Ent}_{\mathbb{P}}(u) \geq 0$ by Jensen's inequality, since $u \log u$is a convex function. Entropy has the following variational representation, know in physics and the Gibss variational principle.

Lemma 10

The enropy can be written as

$$\mathsf{Ent}_{\mathbb{P}}(u) = \sup \left\{ \int uv \, \mathrm{d}\,\mathbb{P} \mid \int e^v \, \mathrm{d}\,\mathbb{P} \leq 1 \right\}. \tag{5.2}$$

Talagrand's cost inequality for Gaussian measures.

- In this subsection we consider a non-degerate normal distribution $N(0, C)$ with the covariance matrix C such that $\det(C) \neq= 0$. We know that this distribution has density $e^{-V(x)}$, where

$$V(x) = \frac{1}{2}\left\langle C^{-1}x, x\right\rangle + const$$

If We denote $A = C^{-1}/2$ then, for any $t \in [0, 1]$,

$$\begin{aligned}
&tV(x) + (1-t)V(y) - V(tx + (1-t)y) \\
&= t\langle Ax, x\rangle + (1-t)\langle Ay, y\rangle - \langle A(tx + (1-t)y), (tx + (1-t)y)\rangle \\
&= t(1-t)\langle A(x-y), x-y\rangle \\
&\geq \frac{1}{2\lambda_{\mathsf{max}}(C)} t(1-t)|x-y|^2 = Kt(1-t)\|x-y\|^2
\end{aligned}$$

where $\lambda_{\mathsf{max}}(C)$ is the largest eigenvalue of C and $K = 1/(2\lambda_{\mathsf{max}}(C))$. We will use this to prove the following useful inequality for the Wasserstein distance W_2.

Talagrand's cost inequality for Gaussian measures.

Theorem 16

If $\mathbb{P} = N(0, c)$ and Q is absolutely continuous with respect to $\mathbb{P}$ with $\int |x|^2 \, \mathrm{d}\mathbb{Q}(x)$ then

$$W_2(\mathbb{Q}, \mathbb{P})^2 \leq 2\lambda_{\mathsf{max}}(C) D(\mathbb{Q} || \, \mathbb{P}). \tag{5.3}$$

Concentration of Gaussian measure.

Given a measurable set $A \subseteq \mathbb{R}^n$ with $\mathbb{P}(A) > 0$, define the distribution $\mathbb{P}_A$ by

$$\underset{A}{\mathbb{P}}(C) = \frac{\mathbb{P}(C \cap A)}{\mathbb{P}(A)}$$

Then, clearly, the Radon-Nikodym derivative

$$\frac{\mathrm{d}\,\mathbb{P}_A}{\mathrm{d}\,\mathbb{P}} = \frac{1}{\mathbb{P}(A)} \mathbf{1}_A$$

and the Kullback-Leibler divergence

$$D(\underset{A}{\mathbb{P}} \,||\, \mathbb{P}) = \int_A \log \frac{1}{\mathbb{P}(A)} \,\mathrm{d}\,\underset{A}{\mathbb{P}} = \log \frac{1}{\mathbb{P}(A)}.$$

Since W_2 is a metric, for any two Borel sets A and B,

$$W_2(\underset{A}{\mathbb{P}}, \underset{B}{\mathbb{P}}) \leq W_2(\underset{A}{\mathbb{P}}, \mathbb{P}) + W_2(\underset{B}{\mathbb{P}}, \mathbb{P}) \leq \sqrt{2\lambda_{\mathsf{max}}(C)} \left(\log^{1/2} \frac{1}{\mathbb{P}(A)} + \log^{1/2} \frac{1}{\mathbb{P}(B)} \right)$$

using (5.3). Suppose that the sets A and B are apart from each other by distance t, i.e. $d(A, B) \geq t > 0$. Then any two points in the support of measures $\mathbb{P}_A$ and $\mathbb{P}_B$ are at a distance at least t from each other, which implies that the transportation distance $W_2(\mathbb{P}_A, \mathbb{P}_B) \geq t$.

Cont.

Therefore,

$$t \le W_2(\underset{A}{\mathbb{P}}, \underset{B}{\mathbb{P}}) \le \sqrt{2\lambda_{\mathsf{max}}(C)} \left(\log^{1/2} \frac{1}{\mathbb{P}(A)} + \log^{1/2} \frac{1}{\mathbb{P}(B)} \right)$$
$$\le \sqrt{4\lambda_{\mathsf{max}}(C)} \log^{1/2} \frac{1}{\mathbb{P}(A)\,\mathbb{P}(B)}$$

Therefore,

$$\mathbb{P}(B) \le \frac{1}{\mathbb{P}(A)} \exp\left(-\frac{t^2}{4\lambda_{\mathsf{max}}(C)} \right).$$

In particular, if $B = \{x \mid d(x, A) \ge t\}$ then

$$\mathbb{P}(d(x, A) \ge t) \le \frac{1}{\mathbb{P}(A)} \exp\left(-\frac{t^2}{4\lambda_{\mathsf{max}}(C)} \right).$$

If the set A is not too small, e.g. $\mathbb{P}(A) \ge 1/2$, this implies that

$$\mathbb{P}(d(x, A) \ge t) \le 2 \exp\left(-\frac{t^2}{4\lambda_{\mathsf{max}}(C)} \right).$$

This shows that the Gaussian measure is exponentially concentrated near any "large" enough set.

Gaussian concentration via infimum-convolution

If we denote $c := 1/\lambda_{\mathsf{max}}(C)$ then setting $t = 1/2$ in,

$$V(x) + V(y) - 2V(\frac{x+y}{2}) \geq \frac{c}{4}|x-y|^2.$$

Given a function f on $\mathbb{R}^n$, let us define its infimum-convolution by

$$g(y) = \inf_x \left(f(x) + \frac{c}{4}|x-y|^2 \right).$$

Then, for all x and y, we have the inequality

$$g(y) - f(x) \leq \frac{c}{4}|x-y|^2 \leq V(x) + V(y) - 2V\left(\frac{x+y}{2}\right). \tag{5.4}$$

If we consider the functions $u(x) = e^{-f(x)-V(x)}, v(y) = e^{g(y)-V(y)}$ and $w(z) = e^{-V(z)}$, then (5.4) implies that

$$w\left(\frac{x+y}{2}\right) \geq u(x)^{1/2} v(y)^{1/2}$$

and the Prekopa-Leindler inequality with $\lambda = 1/2$ implies that

$$\int e^g \,\mathrm{d}\,\mathbb{P} \int e^{-f} \,\mathrm{d}\,\mathbb{P} \leq 1. \tag{5.5}$$

Gaussian concentration via infimum-convolution

Given a measurable set A let f be equal to 0 on A and $+\infty$ on the complement of A. Then $g(y) = \frac{c}{4}d(x,A)^2$ and (5.5) implies

$$\int \exp \frac{c}{4} d(x,A)^2 \,\mathrm{d}\,\mathbb{P}(x) \leq \frac{1}{\mathbb{P}(A)}.$$

By chebyshev's inequaltiy,

$$\mathbb{P}(d(x,A) \geq t) \leq \frac{1}{\mathbb{P}(A)} \exp\left(-\frac{ct^2}{4}\right) = \frac{1}{\mathbb{P}(A)} \exp\left(-\frac{t^2}{4\lambda_{\mathsf{max}}(C)}\right),$$

which is the same Gaussian concentration inequality we proved above.

Discrete metric and total variation

- The *total variation* distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $(S, \mathcal{B})$ is defined by

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{B}} |\,\mathbb{P}(A) - \mathbb{Q}(A)|$$

- Using the Hahn-Jordan decomposition, we can represent a signed measure $\mu = \mathbb{P} - \mathbb{Q}$ as $\mu = \mu^+ - \mu^-$ such that, for some set $D \in \mathcal{B}$ and for any set $E \in \mathcal{B}$,

$$\mu^+(E) = \mu(ED) \geq 0 \text{ and } \mu^-(E) = -\mu(ED^c) \geq 0$$

Therefore, for any $A \in \mathcal{B}$,

$$\mathbb{P}(A) - \mathbb{Q}(A) = \mu^+(AD) - \mu^-(AD^c),$$

which makes it clear that

$$\sup_{A \in \mathcal{B}} |\,\mathbb{P}(A) - \mathbb{Q}(A)| = \mu^+(D).$$

Discrete metric and total variation

- Let us describe some connections of the total variation distance to the Kullback-Leibler divergence and the Kantorovich-Rubinstein theorem.

Lemma 11

If f is a meausrable function on S such that $|f| \leq 1$ and $\int f \, \mathrm{d}\, \mathbb{P} = 0$ than for any $\lambda \in$ r,

$$\int e^{\lambda f} \, \mathrm{d}\, \mathbb{P} \leq e^{\frac{\lambda^2}{2}}.$$

Discrete metric and total variation

- Let us now consider a discrete metric on S given by

$$d(x, y) = \mathbf{1}_{x \neq y}$$

Then a 1-Lipschitz function f w.r.t the metric d, $\|f\|_{\mathsf{L}} \geq 1$, is defined by the condition that for all $x, y \in S$,

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

Formally, the Kantorovich-Rubinstein theorme in this case would state that

$$\begin{aligned} W(\mathbb{P}, \mathbb{Q}) &= \inf \left\{ \int \mathbf{1}_{x \neq y} \, \mathrm{d}\mu(x, y) \,|\, \nu \in M(\mathbb{P}, \mathbb{Q}) \right\} \\ &= \sup \left\{ \left| \int f \, \mathrm{d}\mathbb{Q} - \int f \, \mathrm{d}\,\mathbb{P} \right| \,|\, \|f\|_{\mathsf{L}} \leq 1 \right\} := \gamma(\mathbb{P}, \mathbb{Q}) \end{aligned}$$

- However, since any uncountable set S is not separable w.r.t. the discrete metric d, we can not apply the Kantorovich-Rubinstein theorem directly. In this case, one can use the Hahn-Jordan decomposition to show that W coincides with the total variation distance $W(\mathbb{P}, \mathbb{Q}) = \mathsf{TV}(\mathbb{P}, \mathbb{Q})$. One can also check that $\gamma(\mathbb{P}, \mathbb{Q}) = \mathsf{TV}(\mathbb{P}, \mathbb{Q})$. Thus, for the disctet metric d,

$$W(\mathbb{P}, \mathbb{Q}) = \mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \gamma(\mathbb{P}, \mathbb{Q}).$$

Discrete metric and total variation

- We have the following analogue of the Kullback-Leiber divergence bound for the Gaussian measure in the previsou theorem.

Theorem 17 (Pinsker's inequality)

If $\mathbb{Q}$ is absolutely continous with respect to $\mathbb{P}$ then

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{2D(\mathbb{Q}||\,\mathbb{P})}.$$

Hamming metric on a product space

- Let us consider a finite set A and, given integer $n \geq 1$, consider the folowing Hamming metirc on A^n,

$$d_H(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_(x_i \neq y_i)$$

For two measurs μ and ν on A^n, consider corredsponding Wasserstein distance

$$\begin{aligned} 1W_H(\mu, \nu) &= \inf \left\{ \int d_H(x, y) \, \mathrm{d}\lambda(x, y) \,|\, \lambda \in M(\mu, \nu) \right\} \\ &= \inf \left\{ \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i \neq y_i) \,|\, \lambda \in M(\mu, \nu) \right\}. \end{aligned}$$

Since $d_H(x, y) \leq \mathbf{1}(x \neq y)$, by Pinsker's inequality,

$$W_H(\mu, \nu) \leq \mathsf{TV}(\mu, \nu) \leq \sqrt{2D(\mu||\nu)}. \tag{5.6}$$

- On the other hand, on the product space A^n, one is often interested to understand how different a measure μ is from some (or any) prodcut measure $\nu = \nu_1 \times \cdots \times \nu_n$ w.r.t the above Wasserstein metric. Consider the function

$$\phi_A(x) = -x \log x - (1 - x) \log(1 - x) + x \log \mathsf{card}(A), \ x \in [0, 1].$$

This function concave, $\phi_A(x) \geq 0$ and it is equal to zero only at $x = 0$.

Hamming metric on a product space

The following reverse analog of the above's Pinsker's inequality holds.

Lemma 12

If $\mu_1 \ldots \mu_n$ are the marginals of μ then, for any product measures $\nu = \nu_1 \times \cdots \times \nu_n$ on A^n,

$$\phi_A(W_H(\mu, \nu)) \geq \frac{1}{2n} D(\mu || \mu_1 \times \cdots \times \mu_n). \tag{5.7}$$

In particular,

$$\phi_A(W_H(\mu, \mu_1 \times \cdots \times \mu_n)) \geq \frac{1}{2n} D(\mu || \mu_1 \times \cdots \times \mu_n).$$

Remark. The inequality then shows that the KL divergence between μ and the product measures with the same marginals $\mu_1 \times \cdots \times \mu_n$, can be used to control from below the Wasserstein distance $W_H(\mu, \nu)$ between μ and an arbitrary product measure ν with w.r.t the Hamming distance d_H.