Concentration of measure

Ergan Shang, Han Zhang

USTC

September 17, 2022

Overview

1. Concentration by entropic techniques
   - Entropy and bounds
   - Separately convex functions and the entropic method
   - Tensorization and separately convex functions
2. A geometric perspective on concentration
3. Wasserstein distances and information inequalities
   - Wasserstein distance
   - Tensorization
   - For Markov Chain
   - Asymmetric coupling cost
4. Tail bounds for empirical processes
   - Functional Hoeffding inequality
   - Functional Bernstein inequality

Overview

1. Concentration by entropic techniques
   - Entropy and bounds
   - Separately convex functions and the entropic method
   - Tensorization and separately convex functions
2. A geometric perspective on concentration
3. Wasserstein distances and information inequalities
   - Wasserstein distance
   - Tensorization
   - For Markov Chain
   - Asymmetric coupling cost
4. Tail bounds for empirical processes
   - Functional Hoeffding inequality
   - Functional Bernstein inequality

Entropy and its properties

Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$, we define

$$H_\phi(\boldsymbol{X}) = \mathbb{E}[\phi(\boldsymbol{X})] - \phi(\mathbb{E}[\boldsymbol{X}]) \overset{Jensen}{\geq} 0$$

where $\boldsymbol{X} \sim \mathbb{P}$. Easy to see $H_\phi(\boldsymbol{X}) = 0 \Leftrightarrow \boldsymbol{X}$ is equal to its expectation $\mathbb{P}$-a.e.
Choose the convex function $\phi(u) = ulogu : [0, \infty) \to \mathbb{R}$ and $\phi(0) = 0$ for any non-negative random variable $Z = e^{\lambda \boldsymbol{X}} \geq 0$, we have

$$H(Z) = H(e^{\lambda \boldsymbol{X}}) = \lambda \varphi'(\lambda) - \varphi(\lambda) log \varphi(\lambda)$$

where $\varphi(\lambda) = \mathbb{E}[e^{\lambda \boldsymbol{X}}] \Rightarrow \varphi'(\lambda) = \mathbb{E}[\boldsymbol{X} e^{\lambda \boldsymbol{X}}]$.
So that if we know the moment generating function of $\boldsymbol{X}$, it is straightforward to compute the entropy $H(e^{\lambda \boldsymbol{X}})$.

Herbst argument and its extensions

Herbst argument

Suppose that $H(e^{\lambda \boldsymbol{X}})$ satisfies $H(e^{\lambda \boldsymbol{X}}) \leq \frac{1}{2}\sigma^2\lambda^2\varphi_{\boldsymbol{X}}(\lambda)$ for all $\lambda \in I$ where I can be $[0, \infty)$ or $\mathbb{R}$. Then $\boldsymbol{X}$ satisfies the bound

$$log\mathbb{E}\left[e^{\lambda(\boldsymbol{X}-\mathbb{E}[\boldsymbol{X}])}\right] \leq \frac{1}{2}\lambda^2\sigma^2 \text{ for all } \lambda \in I.$$

RMK:

1. For $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2)$, we have $H(e^{\lambda \boldsymbol{X}}) = \frac{1}{2}\lambda^2\sigma^2\varphi_{\boldsymbol{X}}(\lambda)$.
2. When $I = \mathbb{R}$, the Prop. is equivalent to asserting the sub-Gaussian via Chernoff Ineq.

$$log\mathbb{P}(\boldsymbol{X} - \mu \geq t) \leq \inf_{\lambda}\left\{log\mathbb{E}\left[e^{\lambda(\boldsymbol{x}-\mu)}\right] - \lambda t\right\} \leq \inf_{\lambda}\left\{\frac{1}{2}\lambda^2\sigma^2 - \lambda t\right\}.$$

Proof

Using the condition that $\lambda\varphi'(\lambda) - \varphi(\lambda) log\varphi(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2\varphi(\lambda)\ \forall\lambda \geq 0$, we define the function $G(\lambda) = \frac{1}{\lambda} log\varphi(\lambda)$ with $G(0) := \lim_{\lambda\to 0} G(\lambda) = \mathbb{E}\boldsymbol{X}$. So that

$$G'(\lambda) = \frac{1}{\lambda}\frac{\varphi'(\lambda)}{\varphi(\lambda)} - \frac{1}{\lambda^2} log\varphi(\lambda) \overset{Condition}{\Rightarrow} \lambda^2 G'(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2.$$

Clear λ^2 and integrate both sides we have

$$G(\lambda) - G(\lambda_0) \leq \frac{1}{2}\sigma^2(\lambda - \lambda_0).$$

Let $\lambda_0 \to 0^+$, we have $G(\lambda) - \mathbb{E}\boldsymbol{X} \leq \frac{1}{2}\sigma^2\lambda$ which finishes the proof.

Bernstein entropy bound

Suppose that there are positive constants b, σ such that

$$H(e^{\lambda \boldsymbol{X}}) \leq \lambda^2 \{ b\varphi'_{\boldsymbol{X}}(\lambda) + \varphi_{\boldsymbol{X}}(\lambda)(\sigma^2 - b\mathbb{E}\boldsymbol{X}) \} \text{ for all } \lambda \in [0, 1/b).$$

Then $\boldsymbol{X}$ satisfies the bound

$$log\mathbb{E}\left[e^{\lambda(\boldsymbol{X}-\mathbb{E}\boldsymbol{X})}\right] \leq \sigma^2\lambda^2(1-b\lambda)^{-1} \text{ for all } \lambda \in [0, 1/b).$$

RMK:

1. Also we can derive the Bernstein-type bound via Chernoff Ineq.:
$$\mathbb{P}(\boldsymbol{X} \geq \mathbb{E}[\boldsymbol{X}] + \delta) \leq exp\left(-\frac{\delta^2}{4\sigma^2 + 2b\delta}\right) \text{ for all } \delta \geq 0.$$
2. WLOG we can assume $\mathbb{E}\boldsymbol{X} = 0, b = 1$(**See the next page for Entropy Rescaling Proposition**). Then we define $G(\lambda) = \frac{1}{\lambda} log\varphi(\lambda)$ and imitate the last proof.

Rescaling

Entropy Rescaling

R.V. $\boldsymbol{X}$ satisfies the Bernstein entropy bound

$$H(e^{\lambda \boldsymbol{X}}) \leq \lambda^2 \{b\varphi'_{\boldsymbol{X}}(\lambda) + \varphi_{\boldsymbol{X}}(\lambda)(\sigma^2 - b\mathbb{E}\boldsymbol{X})\} \text{ for all } \lambda \in [0, 1/b)$$

if and only if $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}\boldsymbol{X}$ satisfies

$$H(e^{\lambda \tilde{\boldsymbol{X}}}) \leq \lambda^2 \{b\varphi'_{\tilde{\boldsymbol{X}}}(\lambda) + \varphi_{\tilde{\boldsymbol{X}}}(\lambda)\sigma^2\} \ \forall \lambda \in [0, 1/b). \tag{3.90}$$

Moreover, if $\mathbb{E}\boldsymbol{X} = 0$ and $\boldsymbol{X}$ satisfies 3.90 if and only if $\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X}}{b}$ satisfies

$$H(e^{\lambda \tilde{\boldsymbol{X}}}) \leq \lambda^2 \{\varphi'_{\tilde{\boldsymbol{X}}}(\lambda) + \tilde{\sigma^2}\varphi_{\tilde{\boldsymbol{X}}}(\lambda)\} \ \forall \lambda \in [0, 1),$$

where $\tilde{\sigma^2} = \sigma^2/b^2$.

Proof: By observing the following facts: $\varphi_{a\boldsymbol{X}+b}(\lambda) = e^{\lambda b}\varphi(a\lambda)$ and $\varphi'_{a\boldsymbol{X}+b}(\lambda) = e^{\lambda b}[a\varphi'(a\lambda) + b\varphi(a\lambda)]$.

Separately convex functions and the entropic method

Definition(Separately convex and Lipschitz class)

We say that a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is separately convex if for each index $k \in [n]$, the univariate function $y_k \mapsto f(x_1, x_2, \cdots, x_{k-1}, y_k, x_{k+1}, \cdots, x_n)$ is convex for $\mathbb{R}^{n-1}$ fixed.
A function f is L-Lipschitz wrt Euclidean norm if

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n.$$

Theorem 3.4

Let $\{\boldsymbol{X}_i\}_{i=1}^n$ be independent r.v., each supported on $[a, b]$ and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be **separately convex** and **L-Lipschitz** wrt Euclidean norm. Then $\forall \delta > 0$, we have

$$\mathbb{P}\left(f(\boldsymbol{X}) \geq \mathbb{E}\left[f(\boldsymbol{X})\right] + \delta\right) \leq exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right).$$

Example

We apply the theorem to **Rademacher complexity**.

For $\mathcal{A} \subset \mathbb{R}^n$ is bounded and define $\mathbb{Z} = \sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k \epsilon_k$ where $\epsilon_k \in \{-1, 1\}$ i.i.d. Rademacher r.v. and let us view $\mathbb{Z}$ as a function of ϵ. Observing that

1. $\mathbb{Z} = \mathbb{Z}(\epsilon)$ is the maximum of linear functions, so that it is jointly convex.
2. Let $\mathbb{Z}' = \mathbb{Z}(\epsilon')$ is a second vector of Rademacher random variables. For any $\boldsymbol{a} \in \mathcal{A}$, we have $\langle \boldsymbol{a}, \epsilon \rangle - \mathbb{Z}' = \langle \boldsymbol{a}, \epsilon \rangle - \sup_{\boldsymbol{a}' \in \mathcal{A}} \langle \boldsymbol{a}', \epsilon' \rangle \leq \langle \boldsymbol{a}, \epsilon - \epsilon' \rangle \leq \|\boldsymbol{a}\|_2 \|\epsilon - \epsilon'\|_2$, where the first inequality is by the definition of sup by choosing $\boldsymbol{a}' = \boldsymbol{a}$. Take sup from both sides, we have

$$\mathbb{Z} - \mathbb{Z}' \leq \left(\sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2 \right) \|\epsilon - \epsilon'\|_2.$$

Example

Denote $W(\mathcal{A}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2$ and apply Theorem 3.4, we get

$$\mathbb{P}(\mathbb{Z} \geq \mathbb{E}\mathbb{Z} + t) \leq exp\left(\frac{t^2}{16W^2(\mathcal{A})}\right).$$

Two Lemmas

In order to proof Theorem 3.4, we state the following two lemmas.

Lemma 3.7

Let $X, Y \sim \mathbb{P}$ be a pair of i.i.d. variates. Then for any function $g : \mathbb{R} \to \mathbb{R}$, we have

$$H(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbb{I}(g(X) \geq g(Y))] \ \forall \lambda > 0. \quad (3.20a)$$

If in addition X is supported on $[a, b]$, and g is convex and Lipschtiz, then

$$H(e^{\lambda g(X)}) \leq \lambda^2 (b - a)^2 \mathbb{E}[(g'(X))^2 e^{\lambda g(X)}] \ \forall \lambda > 0. \quad (3.20b)$$

RMK: If g is L-Lipschitz, we are guaranteed by 3.20a that $\|g'\|_\infty \leq L$, so that

$$H(e^{\lambda g(X)}) \leq \lambda^2 L^2 (b - a)^2 \mathbb{E}[e^{\lambda g(X)}] \ \forall \lambda > 0$$

if $X, Y \in [a, b]$.

By definition

$$
\begin{aligned}
H\left(e^{\lambda g(X)}\right) &= \mathbb{E}\left[\lambda g(X) e^{\lambda g(X)}\right] - \mathbb{E}\left[e^{\lambda g(X)}\right] log \mathbb{E}\left[e^{\lambda g(Y)}\right] \\
&\overset{log-concave}{\leq} \mathbb{E}\left[\lambda g(X) e^{\lambda g(X)}\right] - \mathbb{E}\left[e^{\lambda g(X)} \lambda g(Y)\right] \\
&\overset{\text{i.i.d.}}{=} \frac{1}{2}\mathbb{E}\left[\lambda (g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)})\right] \\
&\overset{symmetry}{=} \lambda \mathbb{E}\left[(g(X) - g(Y))(e^{\lambda g(X)} - e^{\lambda g(Y)}) \mathbb{I}\{g(X) \geq g(Y)\}\right]
\end{aligned}
$$

By convexity of e^x, we have $e^s - e^t \leq e^s(s-t)\ \forall s,t \in \mathbb{R}$. So that $(s-t)(e^s - e^t)\mathbb{I}\{s \geq t\} \leq (s-t)^2 e^s \mathbb{I}\{s \geq t\}$. Taking $s = \lambda g(X)$ and $t = \lambda g(Y)$ finishes the proof.

Tensorization property of entropy

Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$, an index $k \in [n]$ and a vector $x_{\backslash k} = (x_i, i \neq k) \in \mathbb{R}^{n-1}$, we define the conditional entropy in coordinate k via

$$H\left(e^{\lambda f_k(X_k)}|x_{\backslash k}\right) := H\left(e^{\lambda f(x_1,\cdots,x_{k-1},X_k,x_{k+1},\cdots,x_n)}\right),$$

where $f_k : \mathbb{R} \to \mathbb{R}; x_k \mapsto f(x_1, \cdots, x_k, \cdots, x_n)$.

When $x_{\backslash k}$ is not fixed, we view the entropy $H\left(e^{\lambda f_k(X_k)}|X^{\backslash k}\right)$ as a random variable.

Lemma 3.8(Tensorization of entropy)

let $f\colon \mathbb{R}^n \to \mathbb{R}$, let $\{X_k\}_{k=1}^n$ be independent random variables. Then

$$H\left(e^{\lambda f(X_1,\cdots,X_n)}\right) \leq \mathbb{E}\left[\sum_{k=1}^n H\left(e^{\lambda f_k(X_k)}|X^{\backslash k}\right)\right] \quad \forall \lambda > 0. \tag{3.21}$$

Exercise 3.9

In order to prove Lemma 3.8, we claim the following truth.

$$H\left(e^{\lambda f(X)}\right) = \sup_{g}\left\{\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] | \mathbb{E}\left[e^{g(X)}\right] \leq 1\right\} \qquad \text{(Exercise 3.9)}$$

Proof: For g such that $\mathbb{E}\left[e^{g(X)}\right]$, define the measure $\mathbb{P}_g(A) = \mathbb{E}\left[\mathbb{I}_A e^{g(X)}\right]$. Then, under the new measure, Jensen's inequality still holds(WHY), so the entropy is still non-negative. Therefore,

$$\begin{aligned} H\left(e^{f(X)}\right) - \mathbb{E}\left[g(X)e^{f(X)}\right] \\ &= \mathbb{E}\left[(f(X) - g(X))e^{f(X)}\right] - \mathbb{E}\left[e^{f(X)}\right] log\mathbb{E}\left[e^{f(X)}\right] \\ &= \mathbb{E}_g\left[(f(X) - g(X))e^{f(X)-g(X)}\right] - \mathbb{E}_g\left[e^{f(X)-g(X)}\right] log\mathbb{E}_g\left[e^{f(X)-g(X)}\right] \\ &= H_g\left(e^{f(X)-g(X)}\right) \geq 0 \end{aligned}$$

Then by choosing $g(x) = f(x) - log\mathbb{E}\left[e^{f(X)}\right]$(r.v.=expectation a.e.) to attain the sup.

Why the Jensen's inequality still hold?

Let the new measure be μ, say $d\mu = g(x)d\nu(x)$ which ν is the base measure. And f is a convex function, so that there exists $a, b \in \mathbb{R}$ such that $f(x) \geq ax + b$. Let $x_0 := \int h(x)d\mu(x)$. We have $f(h(x)) \geq ah(x) + b \quad \forall x$. Therefore

$$\int f(h)d\mu \geq \int (ah + b)d\mu = ax_0 + b = f(x_0) = f(\int hd\mu).$$

Under the language of expectation, this means that

$$f(\mathbb{E}_g[h(X)]) \leq \mathbb{E}_g\left[f(h(X))\right].$$

partcularly, we choose $h(x) = x$, we obtain

$$f(\mathbb{E}_g[X]) \leq \mathbb{E}_g\left[f(X)\right],$$

which is exactly the Jensen's inequality under the new measure $\mathbb{P}_g(A)$.

Proof of 3.8

We define $X_j^n = (X_j, \cdots, X_n)$. Let g be any function such that $\mathbb{E}\left[e^{g(X)}\right] \leq 1$. Then define $\{g^1, \cdots, g^n\}$ via

$$g^1(X_1, \cdots, X_n) = g(X) - log\mathbb{E}\left[e^{g(X)}|X_2^n\right]$$

and

$$g^k(X_k, \cdots, X_n) = log\frac{\mathbb{E}\left[e^{g(X)}|X_k^n\right]}{\mathbb{E}\left[e^{g(X)}|X_{k+1}^n\right]} \text{ for } k = 2, \cdots, n.$$

So that we have

$$\sum_{k=1}^{n} g^k(X_k, \cdots, X_n) = g(X) - log\mathbb{E}\left[e^{g(X)}\right] \geq g(X) \tag{3.25}$$

and moreover, $\mathbb{E}[exp(g^k(X_k, X_{k+1}, \cdots, X_n))|X_{k+1}^n] = 1$ by conditional expectation on X_k.

Proof of 3.8

$$
\begin{aligned}
\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] &\overset{3.25}{\leq} \sum_{k=1}^{n} \mathbb{E}\left[g^k(X_k,\cdots,X_n)e^{\lambda f(X)}\right] \\
&= \sum_{k=1}^{n} \mathbb{E}_{X_{\setminus k}}\left[\mathbb{E}_{X_k}\left[g^k(X_k,\cdots,X_n)e^{\lambda f(X)}|X_{\setminus k}\right]\right] \\
&\overset{14}{\leq} \sum_{k=1}^{n} \mathbb{E}_{X_{\setminus k}}\left[H\left(e^{\lambda f_k(X_k)}|X_{\setminus k}\right)\right].
\end{aligned}
$$

Finally take the supremum over the LHS to finish the proof via 14.

Proof of 8

f is separately convex means $f_k : \mathbb{R} \to \mathbb{R}$ is convex for fixed $x_{\backslash k} \in \mathbb{R}^{n-1}$. By 3.20b, we know

$$H\left(e^{\lambda f_k(X_k)} | x_{\backslash k}\right) \leq \lambda^2 (b-a)^2 \mathbb{E}_{X_k}\left[(f_k'(X_k))^2 e^{\lambda f_k(X_k)} | x_{\backslash k}\right].$$

Then by 3.21, we have

$$H\left(e^{\lambda f(X)}\right) \leq \lambda^2 (b-a)^2 \mathbb{E}\left[\sum_{k=1}^{n} \left(\frac{\partial f(X)}{\partial x_k}\right)^2 e^{\lambda f(X)}\right] \leq \lambda^2 (b-a)^2 L^2 \mathbb{E}\left[e^{\lambda f(X)}\right]$$

by the property of Lipschitz function. Finally 4 finishes the proof.

Overview

Concentration functions

Basic definitions

Given a set $A \subset \mathcal{X}$ and a point $x \in \mathcal{X}$, define $\rho(x, A) = \inf_{y \in A} \rho(x, y)$. The ϵ-entanglement of A is given by

$$A^\epsilon = \{x \in \mathcal{X} : \rho(x, A) < \epsilon\}$$

for a given ϵ.

The concentration function $\alpha : [0, \infty) \to \mathbb{R}_+$ associated with metric space $(\mathbb{P}, \mathcal{X}, \rho)$ is given by

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) = \sup_{A \subset \mathcal{X}} \left\{ 1 - \mathbb{P}[A^\epsilon] | \mathbb{P}[A] \geq \frac{1}{2} \right\} \in [0, \frac{1}{2}].$$

Connections to Lipschitz function

Propostion 3.11

Given a random variable $\boldsymbol{X} \sim \mathbb{P}$ and concentration function α. Any 1-Lipschitz function on $(\mathcal{X}, \rho)$ satisfies

$$\mathbb{P}(|f(x) - m_f| \geq \epsilon) \leq 2\alpha(\epsilon) \tag{3.39a}$$

where $\mathbb{P}(f(x) \geq m_f) \geq \frac{1}{2}$ and $\mathbb{P}(f(x) \leq m_f) \geq \frac{1}{2}$.

Conversely, suppose $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all 1-Lipschitz function on $(\mathcal{X}, \mathbb{P})$ satisfying

$$\mathbb{P}(f(x) \geq \mathbb{E}f(x) + \epsilon) \leq \beta(\epsilon) \text{ for all } \epsilon \geq 0 \tag{3.39b}$$

, then $\alpha(\epsilon) \leq \beta(\frac{\epsilon}{2})$.

RMK: This Prop. shows the connection between concentration around median and the concentration function concerning Lipschitz functions.

Proof of Prop.3.11

For 3.39a, define $A = \{x \in \mathcal{X} : f(x) \leq m_f\}$ and $\forall x \in A^{\epsilon/L}$, there exists $y \in A$ such that $\rho(x, y) \leq \epsilon/L$. When f is L-Lipschitz, we have $|f(x) - f(y)| \leq \epsilon \Rightarrow |f(y)| \leq |f(y) - f(x)| + |f(x)| \leq m_f + \epsilon$. So

$$A^{\epsilon} \overset{L=1}{=} A^{\epsilon/L} \subset \{x \in \mathcal{X} : f(x) < m_f + \epsilon\}$$

, leading to

$\{x : f(x) \geq m_f + \epsilon\} \subset (A^{\epsilon})^c \Rightarrow \mathbb{P}(f(x) \geq m_f + \epsilon) \leq 1 - \mathbb{P}(A^{\epsilon}) \overset{def}{\leq} \alpha(\epsilon)$.
Conversely, fix ϵ, let A be a set with $\mathbb{P}(A) \geq 1/2$ and define $f(x) := min\{\rho(x, A), \epsilon\}$ which is 1-Lipschitz.

Moreover, $\mathbb{E}[f(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X}) | \boldsymbol{X} \notin A]\mathbb{P}(\boldsymbol{X} \notin A) \leq (1 - \mathbb{P}(A))\epsilon \leq \epsilon/2$. Then we have

$$1 - \mathbb{P}(A^{\epsilon}) = \mathbb{P}(f(x) \geq \epsilon) \leq \mathbb{P}\left(f(\boldsymbol{X}) \geq \mathbb{E}f(\boldsymbol{X}) + \frac{\epsilon}{2}\right) \overset{cond}{\leq} \beta(\frac{\epsilon}{2}).$$

Finally, take the sup, we have $\alpha(\epsilon) \leq \beta(\epsilon/2)$.

From median to expectation

Concentrations around means and medians

Given a r.v. $\boldsymbol{X}$, there exists $c_1, c_2 > 0$ such that

$$\mathbb{P}(|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}| \geq t) \leq c_1 e^{-c_2 t^2} \ \forall\ t \geq 0. \tag{2.68}$$

Then

1. $var(\boldsymbol{X}) = \int_0^\infty \mathbb{P}(|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}| \geq t)dt \leq \frac{c_1}{c_2}$
2. For all median m_x,

$$\mathbb{P}(|\boldsymbol{X} - m_x| \geq t) \leq c_3 e^{-c_4 t^2} \tag{2.69}$$

   where $c_3 = 4c_1, c_4 = c_2/8$.
3. Conversely, if 2.69 holds, then 2.68 holds.

See the Proof.

Overview

Wasserstein distance and Transportation cost

For L-Lipschitz function class wrt metric ρ, we use $\|f\|_{Lip}$ to denote the smallest L for which the inequality

$$|f(x) - f(x')| \leq L\rho(x, x')$$

holds.

Given two probability distributions $\mathbb{Q}$ and $\mathbb{P}$ on $\mathcal{X}$, we define the Wasserstein distance between them

$$W_\rho(\mathbb{Q}, \mathbb{P}) = \sup_{\|f\|_{Lip} \leq 1} \left(\int f d\mathbb{Q} - \int f d\mathbb{P} \right).$$

Moreover, let $f : \mathcal{X} \to \mathbb{R}$ be any 1-Lipschitz function, and let $\mathbb{M}$ be any coupling of pair $(\mathbb{Q}, \mathbb{P})$, we have

$$\int \rho(x, x') d\mathbb{M} \overset{lip}{\geq} \int (f(x) - f(x')) d\mathbb{M}(x, x') \overset{def}{=} \int f(d\mathbb{P} - d\mathbb{Q})$$

Wasserstein distance and Transportation cost

In fact, by *Kantorovich–Rubinstein*, we have

$$W_\rho(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_{Lip}\le 1} \int f(d\mathbb{Q}-d\mathbb{P}) = \inf_{\mathbb{M}} \int \rho(x,x')d\mathbb{M}(x,x') = \inf_{\mathbb{M}} \mathbb{E}_{\mathbb{M}}[\rho(x,x')].$$

From this, we can tell the name of transportation cost where ρ can be interpreted as the cost with the transportation plan $\mathbb{M}$ from distribution $\mathbb{P}$ to $\mathbb{Q}$.

Definition 3.18

For a given metric ρ, the probability measure $\mathbb{P}$ is said to satisfy a ρ-transportation cost inequality with parameter $\gamma > 0$ if

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})} \tag{3.58}$$

for all probability measure $\mathbb{Q}$.

Here we utilize an exmaple to see the importance of this definition. Consider the Hamming metric $\rho(x, y) = \mathbb{I}(x \neq y)$ $x, y \in \{0, 1\}$ which is 1-Lipschitz.
We claim that

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = \sup_{f:\mathcal{X}\to[0,1]} \int f(p-q)d\mu$$

where $\|\mathbb{P} - \mathbb{Q}\|_{TV} \overset{def}{=} \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2}\int |d\mathbb{P} - d\mathbb{Q}|$.

Hamming metric

Consider $A = \{p \geq q\}, f : \mathcal{X} \to [0, 1]$. We have $\int f(p - q)d\mu \leq \int_A (p - q)d\mu = \mathbb{P}(A) - \mathbb{Q}(A) \leq \|\mathbb{P} - \mathbb{Q}\|_{TV}$. So that $\sup_{f:\mathcal{X}\to[0,1]} \int f(p - q)d\mu \leq \|\mathbb{P} - \mathbb{Q}\|_{TV}$.

Then for a Borel set B,
$\mathbb{P}(B) - \mathbb{Q}(B) = \int_B (p - q)d\mu \leq \sup_{f:\mathcal{X}\to[0,1]} \int f(p - q)d\mu$. Take the sup, we finish the proof.

Now, assuming the Hamming metric and $\|f\|_{Lip} \leq 1$, we have $f : \mathcal{X} \to [c, c + 1]$. So we may choose $c = 0$ and by the definition $W_\rho(\mathbb{Q}, \mathbb{P}) = \sup_{\|f\|_{Lip}\leq 1} \left(\int f d\mathbb{Q} - \int f d\mathbb{P}\right)$, we know that $W_\rho(\mathbb{P}, \mathbb{Q}) = \|\mathbb{Q} - \mathbb{P}\|_{TV} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})}$, where the final inequality is by *Pinsker-Csizar-Kullback* inequality. Therefore, the parameter $\gamma = \frac{1}{4}$.

We want to prove

$$\|\mathbb{P}-\mathbb{Q}\|_{TV} \leq \sqrt{\frac{1}{2}KL(\mathbb{P}\|\mathbb{Q})}.$$

Define $\psi(x) = xlogx - x + 1$ and $\psi(0) = 0$. The key of the proof is that

$$\left(\frac{4}{3}+\frac{2}{3}x\right)\psi(x) \geq (x-1)^2$$

, which is because $g(x) = (x-1)^2 - \left(\frac{4}{3}+\frac{2}{3}x\right)\psi(x) \leq 0$, via

$$g(1) = 0,\ g'(1) = 0,\ g''(x) = -\frac{4\psi(x)}{3x} \leq 0$$

and the Taylor expansion of g.

Pinsker-Csizar-Kullback Inequality

Finally

$$
\begin{aligned}
\|\mathbb{P}-\mathbb{Q}\|_{TV} &= \frac{1}{2}\int |p-q| = \frac{1}{2}\int \left|\frac{p}{q}-1\right| q \\
&\leq \frac{1}{2}\int q\sqrt{\left(\frac{4}{3}+\frac{2p}{3q}\right)\psi\left(\frac{p}{q}\right)} \\
&\overset{CS}{\leq} \frac{1}{2}\sqrt{\int \left(\frac{4q}{3}+\frac{2p}{3}\right)}\sqrt{\int q\psi\left(\frac{p}{q}\right)} \\
&= \sqrt{\frac{1}{2}\int plog\frac{p}{q}} = \sqrt{\frac{1}{2}KL(\mathbb{P}\|\mathbb{Q})}.
\end{aligned}
$$

From transportation cost to concentration

Theorem 3.19

Consider a metric measure space $(\mathbb{P}, \mathcal{X}, \rho)$, and suppose that $\mathbb{P}$ satisfies the ρ-transportation cost inequality 3.58. Then its concentration function satisfies the bound

$$\alpha(t) \leq 2exp\left(-\frac{t^2}{2\gamma}\right)$$

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

$$\mathbb{P}(|f(x) - \mathbb{E}f(x)| \geq t) \leq 2exp\left(-\frac{t^2}{2\gamma L^2}\right)$$

.

RMK: We can prove the concentration inequality via 3.39a and 2.68 but a worse constant.

Tensorization for transportation cost

Proposition 3.20

For each $k = 1, 2, \cdots, n$, the univariate distribution $\mathbb{P}_k$ satisfies $W_\rho(\mathbb{Q}, \mathbb{P}_k) \leq \sqrt{2\gamma_k D(\mathbb{Q}\|\mathbb{P}_k)}$ for all probability measure $\mathbb{Q}$. Then the product distribution $\mathbb{P} = \bigotimes_{k=1}^n \mathbb{P}_k$ satisfies

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\left(\sum_{k=1}^n \gamma_k\right) D(\mathbb{Q}\|\mathbb{P})} \ \forall \mathbb{Q}$$

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

Proof of Prop.3.20

For $j = 2, \cdots, n$, let $\mathbb{M}_1^j$ denote joint distribution of $(X_1^j, Y_1^j) = (X_1, \cdots, X_j, Y_1, \cdots, Y_j)$; let $\mathbb{M}_{j|j-1}$ denote the conditional distribution of (X_j, Y_j) given (X_1^{j-1}, Y_1^{j-1}) and let $\mathbb{M}_j$ denote the marginal distribution of (X_j, Y_j).

By conditional expectation, we have

$$W_\rho(\mathbb{Q}, \mathbb{P}) = \inf_{\mathbb{M}} \int \rho(x, x') d\mathbb{M}(x, x') \leq \mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1, Y_1)] + \sum_{j=1}^{n} \mathbb{E}_{\mathbb{M}_1^{j-1}}[\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)]].$$

We choose the optimal coupling of $(\mathbb{P}_1, \mathbb{Q}_1)$ as $\mathbb{M}_1$, so that $\mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1, Y_1)] \overset{opt}{=} W_{\rho_1}(\mathbb{Q}_1, \mathbb{P}_1) \leq \sqrt{2\gamma_1 D(\mathbb{Q}_1 \| \mathbb{P}_1)}$.

Similarly, we choose the optimal coupling of $(\mathbb{P}_j, \mathbb{Q}_{j|j-1})$ as $\mathbb{M}_{j|j-1}$, we have

$$\mathbb{E}_{\mathbb{M}_1^{j-1}}[\rho_j(X_j, Y_j)] \leq \sqrt{2\gamma_j D(\mathbb{Q}_{j|j-1}(\cdot|y_1^{j-1})\|\mathbb{P}_j)}.$$

Take expectation of both sides and use the concavity of $\sqrt{x}$ via Jensen's inequality, we have

$$\mathbb{E}_{M_1^{j-1}}[\mathbb{E}_{M_{j|j-1}}[\rho_j(X_j, Y_j)]] \leq \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_{j|j-1}(\cdot|Y_1^{j-1})\|\mathbb{P}_j)]}.$$

Therefore, by using Cauchy-Schwarze inequality, we conclude that

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\gamma_1 D(\mathbb{Q}_1\|\mathbb{P}_1)} + \sum_{j=2}^{n} \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_{j|j-1}(\cdot|Y_1^{j-1})\|\mathbb{P}_j)]}$$

$$\overset{CS}{\leq} \sqrt{2\left(\sum_{j=1}^{n} \gamma_j\right)} \sqrt{D(\mathbb{Q}_1\|\mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_{j|j-1}(\cdot|Y_1^{j-1})\|\mathbb{P}_j)]}$$

Proof of Prop.3.20

Finally, we claim the following lemma called **Chain Rules for KL divergence**.

Given two n-variate distributions $\mathbb{Q}$ and $\mathbb{P}$, we have the decomposition

$$D(\mathbb{Q}\|\mathbb{P}) = D(\mathbb{Q}_1\|\mathbb{P}_1) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}[D(\mathbb{Q}_j(\cdot|X_1^{j-1})\|\mathbb{P}_j(\cdot|X_1^{j-1}))]$$

which is proved here.

Example: McDiarmid's Inequality

We assume that $|f(x_1, \cdots, x_k, \cdots, x_n) - f(x_1, \cdots, x_k', \cdots, x_n)| \leq L_k$ for $x_k \neq x_k'$, and define the metric as $\rho(x, y) = \sum_{k=1}^{n} L_k \mathbb{I}(x_k \neq y_k)$.
By Page 26, we know that

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = \sup_{f:\mathcal{X}\to[0,1]} \int f(d\mathbb{Q} - d\mathbb{P}) = \frac{1}{L_k} W_\rho(\mathbb{Q}, \mathbb{P})$$

$$= \frac{1}{L_k} \sup_{\|f\|_{Lip}\leq 1} \int f(d\mathbb{Q} - d\mathbb{P}) = \sup_{f:\mathcal{X}\to[0,L_k]} \int \frac{f}{L_k}(d\mathbb{Q} - d\mathbb{P}).$$

Therefore, $\gamma_k = \frac{L_k^2}{4} \overset{Prop.3.20}{\Rightarrow} \gamma = \sum_{k=1}^{n} \frac{1}{4}\gamma_k^2$. Finally, we get

$$\mathbb{P}(|f(x) - \mathbb{E}f(x)| \geq t) \leq 2exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right),$$

which is the same as the result obtained in Chapter 2.

Transportation cost inequalities for Markov chains

Let $(X_1, \cdots, X_n)$ be random variables generated by a Markov chain, where each X_i takes value in a countable space $\mathcal{X}$ and the transition kernel

$$\mathbb{K}_{i+1}(x_{i+1}|x_i) = \mathbb{P}_{i+1}(X_{i+1} = x_{i+1}|X_i = x_i)$$

with $X_1 \sim \mathbb{P}_1$.

We define Markov chains that are β-contractive, which means there exists $\beta \in [0, 1)$, such that

$$\max_{i=1,\cdots,n-1} \sup_{x_i, x_i'} \|\mathbb{K}_{i+1}(\cdot|x_i) - \mathbb{K}_{i+1}(\cdot|x_i')\|_{TV} \leq \beta.$$

Transportation cost inequalities for Markov chains

Theorem 3.22

Let $\mathbb{P}$ be the distribution of a β-contractive Markov chain over the discrete space $\mathcal{X}^n$. Then for any other distribution $\mathbb{Q}$ over $\mathcal{X}^n$, we have

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \frac{1}{1-\beta}\sqrt{\frac{n}{2}D(\mathbb{Q}\|\mathbb{P})}$$

where the Wasserstein distance is defined wrt the Hamming norm $\rho(x, y) = \sum_{i=1}^n \mathbb{I}(x_i \neq y_i)$.

Now we consider an exmaple concerning "Parameter estimation for a binary Markov chain". Let $X_i \in \{0, 1\}^2$ specified by an initial distribution $\mathbb{P}_1$ that is uniform, and the transition kernel

$$\mathbb{K}_{i+1}(x_{i+1}|x_i) = \begin{cases} \frac{1}{2}(1+\delta) & \text{if } x_{i+1} = x_i, \\ \frac{1}{2}(1-\delta) & \text{if } x_{i+1} \neq x_i, \end{cases} \Rightarrow \beta = \delta$$

where $\delta \in [0, 1]$.

Parameter estimation for a binary Markov chain

Our goal is to estimate δ on $(X_1, \cdots, X_n)$. An unbiased estimate of $\frac{1}{2}(1+\delta)$ is given by the function

$$f(X_1, \cdots, X_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{I}(X_i = X_{i+1}) \Rightarrow \mathbb{E}[f(X_1, \cdots, X_n)] = \frac{1}{2}(1+\delta)$$

Moreover, function f is $\frac{2}{n-1}$-Lipschitz wrt the Hamming norm. So that

$$\mathbb{P}\left(|f(X) - \frac{1}{2}(1+\delta)| \geq t\right) \leq 2exp\left(-\frac{(n-1)^2(1-\delta)^2 t^2}{2n}\right).$$

Asymmetric coupling cost

Up to now, transportation cost is mentioned to derive concentration inequality, but with dimension n. We would like to capture inequalities free of dimensions.

We define

$$\mathcal{C}(\mathbb{Q}, \mathbb{P}) := \inf \sqrt{\int \sum_{i=1}^{n} (\mathbb{M}(Y_i \neq x_i | X_i = x_i))^2 d\mathbb{P}(x)}$$

$$\stackrel{ref}{=} \sqrt{\int \left| 1 - \frac{d\mathbb{Q}}{d\mathbb{P}}(x) \right|_{+}^{2} d\mathbb{P}(x)}.$$

Due to Samson's result [51], we have

$$max\{\mathcal{C}(\mathbb{Q}, \mathbb{P}), \mathcal{C}(\mathbb{P}, \mathbb{Q})\} \leq \sqrt{2D(\mathbb{Q} \| \mathbb{P})}.$$

Asymmetric coupling and Total Variation

In fact we have the following observation

$$2\|\mathbb{P}-\mathbb{Q}\|_{TV} = \int |p-q| = 2\int |p-q|_+$$

which the second equality is because

$$\int_{p>q} p - q dx + \int_{q>p} q - p dx = 2\int_{p>q} p - q dx$$

via $\mathbb{P}(A) - \mathbb{Q}(A) = \mathbb{Q}(A^c) - \mathbb{P}(A^c)$.
Eventually, the Pinsker inequality implies Samson's result.

Asymmetric coupling cost

Theorem 3.24

Consider a vector of independent random variables $(X_1, \cdots, X_n)$, each taking values in $[0, 1]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex , and L-Lipschitz wrt the Euclidean norm. Then for all $t \geq 0$, we have

$$\mathbb{P}(|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| \geq t) \leq 2exp\left(-\frac{t^2}{2L^2}\right).$$

Proof:

We want to prove that $W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2D(\mathbb{Q}\|\mathbb{P})}$. By convexity, we have $f(x) \geq f(y) + \nabla f(y)^\top (y - x) = f(y) + \sum_{i=1}^n \frac{\partial f}{\partial y_i}(y)(y_i - x_i) \Rightarrow$ $f(y) - f(x) \leq \sum_{j=1}^n |\frac{\partial f}{\partial y_j}|\mathbb{I}(x_j \neq y_j)$ because $x, y \in [0, 1]^n$.

Proof of Theorem 3.24

So that, by conditional expectation

$$
\begin{aligned}
\int f(y)d\mathbb{Q}(y) - \int f(x)d\mathbb{P}(x) &\leq \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}(x_j \neq y_j) d\mathbb{M}(x,y) \\
&= \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{M}(X_j \neq y_j | Y_j = y_j) d\mathbb{Q}(y) \\
&\overset{CS}{\leq} \int \|\nabla f(y)\| \sqrt{\sum_{j=1}^{n} \mathbb{M}^2(X_j \neq y_j | Y_j \neq y_j)} d\mathbb{Q}(y) \\
&= L\mathcal{C}(\mathbb{P},\mathbb{Q})
\end{aligned}
$$

Therefore $\int \frac{f}{L}(d\mathbb{Q} - d\mathbb{P}) \leq \mathcal{C}(\mathbb{P},\mathbb{Q}) \leq \sqrt{2D(\mathbb{Q}\|\mathbb{P})} \Rightarrow W_\rho(\mathbb{Q}\|\mathbb{P}) \leq \sqrt{2D(\mathbb{Q}\|\mathbb{P})} \Rightarrow \gamma = 1$, which finishes the proof.

Revist to Rademacher complexity

As previously mentioned in 9: For $\mathcal{A} \subset \mathbb{R}^n$, Rademacher complexity is defined as $Z = Z(\epsilon_1, \cdots, \epsilon_n) := \sup_{\boldsymbol{x} \in \mathcal{A}} \sum_{k=1}^{n} a_k \epsilon_k$. We have shown that $\epsilon \mapsto Z(\epsilon)$ is jointly convex and the Lipschitz parameter is $W(\mathcal{A}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2$. Then Theorem 3.24 implies

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq 2exp\left(-\frac{t^2}{8W^2(\mathcal{A})}\right),$$

which provides a two-sided bound, also sharper.

Please pay attention that the number 8 is different from the number in the textbook, because $\epsilon_i \in \{-1, +1\}$, whereas the conditions in the theorem ask for $X_n \in [0, 1]$. By reviewing the proof when $X_n \in [-1, 1]$, we conclude the "8".

Overview

1. Concentration by entropic techniques
   - Entropy and bounds
   - Separately convex functions and the entropic method
   - Tensorization and separately convex functions
2. A geometric perspective on concentration
3. Wasserstein distances and information inequalities
   - Wasserstein distance
   - Tensorization
   - For Markov Chain
   - Asymmetric coupling cost
4. Tail bounds for empirical processes
   - Functional Hoeffding inequality
   - Functional Bernstein inequality

Motivation

This section, we focus on the sample averages over function classes

$$Z := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\},$$

where $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, with $(X_1, \cdots, X_n)$ drawn from a product distribution $\mathbb{P} = \bigotimes_{i=1}^{n} \mathbb{P}_i$ and each $\mathbb{P}_i$ is supported on some set $\mathcal{X}_i \subset \mathcal{X}$.

What should be noticed is that if the goal is to obtain bounds on $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|$, we can consider the augmented function class $\tilde{\mathcal{F}} = \mathcal{F} \cup \{-\mathcal{F}\}$.

A functional Hoeffding inequality

Theorem 3.26(Functional Hoeffding theorem)

For each $f \in \mathcal{F}$ and $i = 1, \cdots$, assume that there are real numbers $a_{if} \le b_{if}$ such that $f(x) \in [a_{if}, b_{if}]$ for all $x \in \mathcal{X}_i$. Then for all $\delta \ge 0$, we have

$$\mathbb{P}(Z \ge \mathbb{E}[Z] + \delta) \le exp\left(-\frac{n\delta^2}{4L^2}\right), \tag{3.80}$$

where $L^2 = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (b_{if} - a_{if})^2$.

We begin with 3.20a and 3.21 for function via $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$, $x_j \mapsto Z_j(x_j) = Z(X_1, \cdots, X_{j-1}, x_j, X_{j+1}, \cdots, X_n)$:

$$H\left(e^{\lambda Z(X)}\right) \le \lambda^2 \mathbb{E}\left[\sum_{j=1}^n \mathbb{E}[(Z_j(X_j) - Z_j(Y_j))^2 \mathbb{I}(Z_j(X_j) \ge Z_j(Y_j)) e^{\lambda Z(X)} | X^{\setminus j}]\right].$$

(*)

Proof of Theorem 3.26

Define the set $\mathcal{A}(f) := \{(x_1, \cdots, x_n) \in \mathbb{R}^n : Z = \sum_{i=1}^n f(x_i)\}$, which is the set of realizations for which Z defined by sup **is achieved by** f. For any $x \in \mathcal{A}(f)$, we have

$$Z_j(x_j) - Z_j(y_j) = f(x_j) + \sum_{i \neq j}^n f(x_i) - \max_{\tilde{f} \in \mathcal{F}} \left(\tilde{f}(y_j) + \sum_{i \neq j}^n \tilde{f}(y_i) \right) \overset{choose \tilde{f}=f}{\leq} f(x_j) - f(y_j).$$

So that

$$\sum_{j=1}^n (Z_j(x_j) - Z_j(y_j))^2 \mathbb{I}(Z_j(x_j) \geq Z_j(y_j)) e^{\lambda Z(x)}$$

$$\leq \sum_{f \in \mathcal{F}} \mathbb{I}(x \in \mathcal{A}(f)) \sum_{j=1}^n (b_{jf} - a_{jf})^2 e^{\lambda Z(x)}$$

$$\leq \sup_{f \in \mathcal{F}} \sum_{j=1}^n (b_{jf} - a_{jf})^2 e^{\lambda Z(x)} := nL^2 e^{\lambda Z(x)}.$$

Proof of Theorem 3.26

Then plug into *, we have

$$H\left(e^{\lambda Z(x)}\right) \leq nL^2\lambda^2\mathbb{E}\left[e^{\lambda Z(x)}\right].$$

Finally, 4 finishes the proof.

A functional Bernstein inequality

Theorem 3.27(Talagrand concentration for empirical processes)

Consider a countable class if functions $\mathcal{F}$ uniformly bounded by b. Then for all $\delta > 0$, the random variable $Z = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ satisfies the upper tail bound

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + \delta) \leq 2exp\left(\frac{-n\delta^2}{8e\mathbb{E}[\Sigma^2] + 4b\delta}\right), \tag{3.83}$$

where $\Sigma^2 = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} f^2(X_i)$.

We also work with $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i)$ and recall the proof by * to get

$$H\left(e^{\lambda Z}\right) \leq \lambda^2 \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{E}\left[\sum_{f\in\mathcal{F}} \mathbb{I}(x \in \mathcal{A}(f))(f(X_j) - f(Y_j))^2 e^{\lambda Z} | X^{\setminus j}\right]\right]. \tag{**}$$

Proof of Theorem 3.27

By $(x+y)^2 \leq 2(x^2+y^2)$, we have

$$\sum_{j=1}^{n} \sum_{f \in \mathcal{F}} \mathbb{I}(X \in \mathcal{A}(f))(f(X_j) - f(Y_j))^2 \leq 2 \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2(X_i) + 2 \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2(Y_i)$$

$$:= 2(\Gamma(X) + \Gamma(Y))$$

where $\Gamma(X) = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2(X_i)$. By plugging into **, we get the entropy bound

$$H\left(e^{\lambda Z(X)}\right) \leq 2\lambda^2 \left(\mathbb{E}\left[\Gamma(X) e^{\lambda Z(X)}\right] + \mathbb{E}[\Gamma(Y)]\mathbb{E}\left[e^{\lambda Z(X)}\right]\right)$$

by independence. Via 50, multiply on both side by $e^{-\lambda \mathbb{E}[Z]}$, we have

$$H\left(e^{\lambda \tilde{Z}(X)}\right) \leq 2\lambda^2 \left(\mathbb{E}\left[\Gamma(X) e^{\lambda (Z(X) - \mathbb{E}[Z])}\right] + \mathbb{E}[\Gamma(Y)]\mathbb{E}\left[e^{\lambda (Z(X) - \mathbb{E}[Z])}\right]\right) \tag{***}$$

where $\tilde{Z} := Z - \mathbb{E}[Z]$.

Proof of Theorem 3.27

Lemma 3.28(Controlling the random variance)

For all $\lambda > 0$, we have

$$\mathbb{E}[\Gamma e^{\lambda \tilde{Z}}] \leq (e-1)\mathbb{E}[\Gamma]\mathbb{E}[e^{\lambda \tilde{Z}}] + \mathbb{E}[\tilde{Z} e^{\lambda \tilde{Z}}]. \tag{3.88}$$

Putting together with ***, we have

$$H\left(e^{\lambda \tilde{Z}}\right) \leq \lambda^2 \left(2e\mathbb{E}[\Gamma]\varphi(\lambda) + 2\varphi'(\lambda)\right)$$

Finally, combined with 6, we know the parameter in 6 is $b = 2, \sigma^2 = 2e\mathbb{E}[\Gamma]$ by noting that $\mathbb{E}[\tilde{Z}] = 0$. By choosing $\delta = nt$, we obtain

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq exp\left(-\frac{n^2t^2}{8e\mathbb{E}[\Gamma] + 4nt}\right) = exp\left(-\frac{nt^2}{8e\mathbb{E}[\Sigma^2] + 4t}\right)$$

which is the situation of $b = 1$ to which general cases could be reduced.

Exercise 3.4

We claim

$$H\left(e^{\lambda(X+c)}\right) = e^{\lambda c} H\left(e^{\lambda X}\right)$$

for r.v. X and constant $c \in \mathbb{R}$.

It is easy to check by expanding H.

References

1 Wainwright M J. High-dimensional statistics: A non-asymptotic viewpoint[M]. Cambridge University Press, 2019.

2 Proof for 49

3 Grimmett G, Stirzaker D. Probability and random processes[M]. Oxford university press, 2020.

4 Samson P M. Concentration of measure inequalities for Markov chains and Φ-mixing processes[J]. The Annals of Probability, 2000, 28(1): 416-461.

Thank you !