3. Concentration of measure

Yoji Tomita

May 12, 2021

Introduction

- 2 章を前提として, この章では tail bound や concentration inequalities を求めるためのより上級的な手法を紹介する.
- 3.1：Concentration by entropic techniques
- 3.2：A geometric perspective on concentration
- 3.3：Wasserstein distances and information inequalities
- 3.4：Tail bounds for empirical processes

3.1 Concentration by entropic techniques

- エントロピーと, 集中不等式導出のためのその関連テクニックに関する議論から始める.

3.1.1 Entropy and its properties

- 凸関数 $\phi : \mathbb{R} \to \mathbb{R}$ と, 確率変数 $X \sim \mathbb{P}$ に対して, ϕ-entropy を

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X])$$

  とする ($X, \phi(X)$ の有限期待値は仮定).

- Jensen の不等式より, ϕ-entropy は非負.
- これは X のばらつき加減を表す.
  - 極端な場合, X が a.s. で期待値と一致するなら, $\mathbb{H}_\phi(X) = 0$.

- 例 1：$\phi(u) = u^2$ なら $\mathbb{H}_\phi(X)$ は分散.

$$\mathbb{H}_\phi(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathrm{var}(X).$$

- 例 2：$\phi(u) = -\log u$, $Z := e^{\lambda X}$ とすると,

$$\mathbb{H}_\phi(e^{\lambda X}) = -\lambda \mathbb{E}[X] + \log \mathbb{E}[e^{\lambda X}] = \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]$$

となり, centerd cumulant generating function となる.

- この章では, 次の凸関数 $\phi : [0, \infty) \to \mathbb{R}$ に対する entropy を考える.

$$\phi(u) := \begin{cases} u \log u & \text{for } u > 0 \\ 0 & \text{for } u = 0 \end{cases}. \tag{3.1}$$

非負確率変数 Z に対して, ϕ-entropy は

$$\mathbb{H}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z], \tag{3.2}$$

となる (ただし関連する期待値の存在は仮定).

  - Shannon entropy や Kullback-Leibler divergence と関連がある (see Exercise 3.1).
  - 以後この entropy を考えるので, $\mathbb{H}_\phi$ の subscrript ϕ は省略.

- $Z = e^{\lambda X}$ とすると, $\mathbb{H}(e^{\lambda X})$ は X のモーメント母関数 $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ とその導関数 $\varphi'_X(\lambda)$ で表せる.

$$\mathbb{H}(e^{\lambda X}) = \lambda \varphi'_X(\lambda) - \varphi_X(\lambda) \log \varphi_X(\lambda). \tag{3.3}$$

Example 3.1 (Entropy of a Gauusian random variable)

- X は 1 次元正規分布 $X \sim \mathcal{N}(0, \sigma^2)$ とすると, $\varphi_X(\lambda) = e^{\lambda^2\sigma^2/2}, \varphi'_X(\lambda) = \lambda\sigma^2\varphi_X(\lambda)$ なので,

$$\mathbb{H}(e^{\lambda X}) = \lambda^2\sigma^2\varphi_X(\lambda) - \frac{1}{2}\lambda^2\sigma^2\varphi_X(\lambda) = \frac{1}{2}\lambda^2\sigma^2\varphi_X(\lambda). \tag{3.4}$$

- この節の残りで, このエントロピー (3.3) と tail bounds との関連性を説明していく.

3.1.2 Herbst argument and its extensions

- ある定数 $\sigma > 0$ が存在して, エントロピーが次の上限を満たすとする.

$$\mathbb{H}(e^{\lambda X}) \leq \frac{1}{2}\sigma^2\lambda^2\varphi_X(\lambda). \tag{3.5}$$

  - Example 3.1 より, 正規分布 $X \sim \mathcal{N}(0, \sigma^2)$ は任意の $\lambda \in \mathbb{R}$ に対し (3.5) をイコールで満たす.
  - また任意の bounded な確率変数も (3.5) を満たす (Exercise 3.7).
- このとき, その確率変数は sub-Gaussian となる.

Proposition 3.2 (Herbst argument)

エントロピー $\mathbb{H}(e^{\lambda X})$ が (3.5) を任意の $\lambda \in I$ (ただし $I = [0, \infty)$ or $\mathbb{R}$) について満たすとする. このとき,

$$\log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \frac{1}{2}\lambda^2\sigma^2 \quad \text{for all } \lambda \in I. \tag{3.6}$$

Remarks:

- $I = \mathbb{R}$ なら, (3.6) は $X - \mathbb{E}[X]$ がパラメータ σ の sub-Gauusian であることと同値.
- Chernoff argument より, $I = [0, \infty)$ で片側 tail-bound

$$\mathbb{P}[X \geq \mathbb{E}[X] + t] \leq e^{-\frac{t^2}{2\sigma^2}} \tag{3.7}$$

が得られ, $I = \mathbb{R}$ なら両側 tail bounds

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

となる.

Proof.

- $I = [0, \infty)$ の場合のみ示す ($I = \mathbb{R}$ は演習とする).
- エントロピーのモーメント母関数による表現 (3.3) と仮定 (3.5) より,

$$\mathbb{H}(e^{\lambda X}) = \lambda\varphi'(\lambda) - \varphi(\lambda)\log\varphi(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2\varphi(\lambda) \quad \forall\lambda \geq 0. \tag{3.8}$$

- 関数 G を $G(\lambda) := \frac{1}{\lambda}\log\varphi(\lambda)$ $(\lambda \neq 0)$ と定義し, $\lambda = 0$ では連続性を満たすように

$$G(0) := \lim_{\lambda\to 0} G(\lambda) = \mathbb{E}[X] \tag{3.9}$$

とする.

- $G'(\lambda) = \frac{1}{\lambda}\frac{\varphi'(\lambda)}{\varphi(\lambda)} - \frac{1}{\lambda^2}\log\varphi(\lambda)$ より, (3.8) は $G'(\lambda) \leq \frac{1}{2}\sigma^2$ となるので, $\lambda_0 (> 0)$ から λ まで両辺積分すると

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

- $\lambda_0 \downarrow 0$ とすると

$$G(\lambda) - \mathbb{E}[X] \leq \frac{1}{2}\sigma^2\lambda$$

となり, これは (3.6) と同値である. □

- 2 章と同様に, 次は sub-exponential tail をもつ確率変数を考える.

Proposition 3.3 (Bernsten entropy bound)

正整数 b, σ が存在して, エントロピー $\mathbb{H}(e^{\lambda X})$ は以下を満たすとする.

$$\mathbb{H}(e^{\lambda x}) \leq \lambda^2 \left\{ b\varphi_X'(\lambda) + \varphi_X(\lambda)(\sigma^2 - b\mathbb{E}[X]) \right\} \quad \text{for all } \lambda \in [0, 1/b). \tag{3.10}$$

このとき,

$$\log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \sigma^2 \lambda^2 (1 - b\lambda)^{-1} \quad \text{for all } \lambda \in [0, 1/b). \tag{3.11}$$

Remarks:

- Chernoff argument より, Prop.3.3 は以下の sub-exponential tails をもつ変数の上側 Bernstein-type bound を含意する.

$$\mathbb{P}[X \geq \mathbb{E}[X] + \delta] \leq \exp\left(-\frac{\delta^2}{4\sigma^2 + 2b\delta} \right) \quad \text{for all } \delta \geq 0. \tag{3.12}$$

Proof.

- 一般性を失わずに $\mathbb{E}[X] = 0$ と $b = 1$ を仮定できる (see Exercise 3.6).
- このとき (3.10) は次のように簡単化される.

$$\mathbb{H}(e^{\lambda X}) \leq \lambda^2 \left\{ \varphi'(\lambda) + \varphi(\lambda)\sigma^2 \right\} \quad \text{for all } \lambda \in [0, 1). \tag{3.13}$$

- Prop.3.2 の証明と同様に $G(\lambda) = \frac{1}{\lambda} \log \varphi(\lambda)$ を定義すると, (3.13) は $G' \leq \sigma^2 + \frac{\varphi'}{\varphi}$ と同値になる.
- $\lambda_0 > 0$ を任意にとり λ_0 から λ まで両辺積分すると,

$$G(\lambda) - G(\lambda_0) \leq \sigma^2(\lambda - \lambda_0) + \log \varphi(\lambda) - \log \varphi(\lambda_0).$$

- $\lambda_0 \downarrow 0$ とすると, $\lim_{\lambda \downarrow 0} G(\lambda_0) = \mathbb{E}[X]$ と $\log \varphi(0) = 0$ より,

$$G(\lambda) - \mathbb{E}[X] \leq \sigma^2 \lambda + \log \varphi(\lambda) \tag{3.14}$$

  となる.
- (3.14) に G と φ を入れて書き換えると (3.11) が得られる. □

3.1.3 Separately convex functions and the entropic method

- Entropy method は複数の確率変数からなる関数の concentration を考える時に強力.
- 関数 $f : \mathbb{R}^n \to \mathbb{R}$ が separately convex であるとは, 各 $k \in \{1, \dots, n\}$ について 1 変数関数

$$y_k \mapsto f(x_1, \dots, x_{k-1}, y_k, x_{k+1}, \dots, x_n)$$

が任意の $(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) \in \mathbb{R}^{n-1}$ に対して凸であることをいう.

- また関数 f がユークリッドノルムに対して L-Lipschitz であるとは,

$$\left|f(x) - f(x')\right| \le L\|x - x'\| \quad \text{for all } x, x' \in \mathbb{R}^n.$$

が成り立つことをいう.

Theorem 3.4

$\{X_i\}_{i=1}^n$ は独立な確率変数列でそれぞれのサポートは $[a,b]$ に含まれるものとし, $f:\mathbb{R}^n \to \mathbb{R}$ は separately convex かつ L-Lipschitz であるとする. このとき, 任意の $\delta > 0$ に対して

$$\mathbb{P}\left[f(X) \le \mathbb{E}[f(X)] + \delta\right] \le \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right) \tag{3.16}$$

が成り立つ.

Remarks:

- この結果は Gaussian へ変数の Lipscitz 関数の upper tail bound を求めた Thm.2.26 の analogue だが, 独立かつ bounded な変数に対して適用できる.
- ただし, separetely convexity の仮定は一般に取り除くことができない.
- f が jointly convex の場合は lower tail bound の導出に他のテクニックが使える (cf Thm.3.24).

Example 3.5 (Sharp bounds on Rademacher complexity)

- 有界部分集合 $\mathcal{A} \in \mathbb{R}^n$ は所与とし, 確率変数 $Z = \sup_{a \in \mathcal{A}} \sum_{k=1}^n a_k \epsilon_k$ を考える.
- $\epsilon_k \in \{-1, 1\}$ は i.i.d. な Rademacher variables.
- Z は線形関数の sup をとったものなので jointly (したがって separately) convex.
- 別の $Z' = Z(\epsilon')$ に対し, 任意の $a \in \mathcal{A}$ について

$$\langle a, \epsilon \rangle - Z' = \langle a, \epsilon \rangle - \sup_{a' \in \mathcal{A}} \langle a, \epsilon' \rangle \leq \langle a, \epsilon - \epsilon' \rangle \leq \|a\|_2 \|\epsilon - \epsilon'\|_2.$$

- $a \in \mathcal{A}$ について sup をとると $Z - Z' \leq \sup_{a \in \mathcal{A}} \|a\|_2 \|\epsilon - \epsilon'\|$.
- よって, Z は $\mathcal{W}(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$-Lipschitz.
- したがって, Theorem 3.4 より

$$\mathcal{P}[Z \leq \mathbb{E}[Z] + t] \leq \exp\left(-\frac{t^2}{16 \mathcal{W}^2(\mathcal{A})}\right). \tag{3.17}$$

- 通常 $\mathcal{W}^2(A)$ は $\sum_{k=1}^n \sup a_k^2$ よりずっと小さいので, Example2.25 より強い bound となる.

Example 3.6 (Operator norm of a random matrix)

- $\mathbf{X} \in \mathbb{R}^{n \times d}$ はランダム行列で, X_{ij} は平均 0, サポート $[-1, +1]$ の分布から i.i.d. でひかれるとする.
- $|||\mathbf{X}|||_2$ はスペクトルノルム（最大の特異値）, or ℓ_2-オペレータノルムとし, これは以下で与えられる.

$$|||\mathbf{X}|||_2 = \max_{\substack{v \in \mathbb{R}^d \\ \|v\|_2 = 1}} \|\mathbf{X}v\|_2 = \max_{\substack{v \in \mathbb{R}^d \\ \|v\|_2 = 1}} \max_{\substack{u \in \mathbb{R}^n \\ \|u\|_2 = 1}} u^{\mathrm{T}} \mathbf{X} v. \tag{3.18}$$

- $\mathbf{X} \mapsto |||\mathbf{X}|||_2$ を関数 $f : \mathbb{R}^{nd} \to \mathbb{R}$ と見ると, f は Theorem 3.4 の仮定を満たす.
  - f は線形な関数の supremum なので convex,
  - かつ

$$\Big| |||\mathbf{X}|||_2 - |||\mathbf{X}'|||_2 \Big| \overset{\text{(i)}}{\leq} |||\mathbf{X} - \mathbf{X}'|||_2 \overset{\text{(ii)}}{\leq} |||\mathbf{X} - \mathbf{X}'|||_{\mathrm{F}}, \tag{3.19}$$

    となる（(i) は三角不等式, (ii) はフロベニウスノルムは常にオペレータノルムを上からおさえることより）ので, f は $L = 1$ で Lipschitz.
- したがって, Theorem 3.4 より,

$$\mathbb{P}[|||\mathbf{X}|||_2 \geq \mathbb{E}[|||\mathbf{X}|||_2] + \delta] \leq e^{-\frac{\delta^2}{16}}.$$

3.1.4 Tensorization and separately convex functions

- 2 つの lemma をもとに Theorem 3.4 を証明する.

Lemma 3.7

$X, Y \sim \mathbb{P}$, i.i.d. とすると, 任意の関数 $g : \mathbb{R} \to \mathbb{R}$ に対し以下が成り立つ.

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 \mathbb{E}\left[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbb{I}[g(X) \ge g(Y)]\right] \quad \text{for all } \lambda > 0. \tag{3.20a}$$

さらに X のサポートが $[a, b]$ に含まれ, g が凸かつ Lipschitz なら,

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 (b-a)^2 \mathbb{E}\left[(g'(X))^2 e^{\lambda g(X)}\right] \quad \text{for all } \lambda > 0, \tag{3.20b}$$

が成り立つ (g' は g の導関数).

- Lemma 3.7 は, 凸かつ Lipschitz な関数はほとんどいたるところ微分可能であるという事実を使っている (Rademacher's Theorem).
- また, g が L-Lipschitz なら $\|g'\|_\infty \le L$[1] なので, (3.20b) は以下を含意する.

$$\mathbb{H}(e^{\lambda g(X)}) \le \lambda^2 L^2 (b-a)^2 \mathbb{E}[e^{\lambda g(X)}] \quad \text{for all } \lambda > 0.$$

- したがって Proposition 3.2 より

$$\mathbb{P}[g(X) \ge \mathbb{E}[g(X)] + \delta] \le e^{-\frac{\delta^2}{4L^2(b-a)^2}}$$

  となるので, Lemma 3.7 は Theorem 3.4 の 1 変数バージョンをただちに導く.
- しかし (3.20b) では L でなく g' とより強い bound となっており, これが Theorem 3.4 の導出において重要となる.

[1]関数 f に対して $\|f\|_\infty$ は f の本質的上限 $\|f\|_\infty := \inf\{C \ge 0 : |f(x)| \le C \text{ almost every } x.\}$.

Proof of Lemma 3.7

- エントロピーの定義より,

$$\begin{aligned}
\mathbb{H}(e^{\lambda g(X)}) &= \mathbb{E}_X[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}_X[e^{\lambda g(X)}] \log(\mathbb{E}_Y[e^{\lambda g(Y)}]) \\
&\leq \mathbb{E}_X[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}_{X,Y}[e^{\lambda g(X)} \lambda g(Y)] \qquad \text{(Jensen's inequality)} \\
&= \frac{1}{2} \mathbb{E}_{X,Y} \left[\lambda \{g(X) - g(Y)\} \{e^{\lambda g(X)} - e^{\lambda g(Y)}\} \right] \\
&= \lambda \mathbb{E} \left[\{g(X) - g(Y)\} \{e^{\lambda g(X)} - e^{\lambda g(Y)}\} \mathbb{I}[g(X) \geq g(Y)] \right]. \qquad (3.22)
\end{aligned}$$

となる (ただし最後の等式は X, Y の対称性より).

- 指数関数の凸性より, 任意の $s \geq t$ に対し $e^s - e^t \leq e^s(s-t)$ なので,

$$(s-t)(e^s - e^t)\mathbb{I}[s \geq t] \leq (s-t)^2 e^s \mathbb{I}[s \geq t]$$

- これを (3.22) に適用すると, (3.20a)

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbb{I}[g(X) \geq g(Y)]]. \tag{3.23}$$

  が得られる.
- さらに g が凸で $x, y \in [a, b]$ とすると, $g(x) - g(y) \leq g'(x)(x-y)$ より,

$$(g(x) - g(y))^2 \mathbb{I}[g(x) \geq g(y)] \leq (g'(x))^2 (x-y)^2 \leq (g'(x))^2 (b-a)^2$$

  となるので, これを (3.23) に適用すれば (3.20a) が得られる. □

Tensorization peoperty of entropy

- 関数 $f : \mathbb{R}^n \to \mathbb{R}, k \in \{1, \dots, n\}, x_{\setminus k} = (x_i, i \neq k) \in \mathbb{R}^{n-1}$ に対して, 条件付きエントロピーを
$$\mathbb{H}(e^{\lambda f_k(X_k)} \mid x_{\setminus k}) := \mathbb{H}(e^{\lambda f(x_1,\dots,x_{k-1},X_k,x_{k+1},\dots,x_n)})$$
と定義する（f_k は $x_k \mapsto f(x_1, \dots, x_k, \dots, x_n)$ なる 1 変数関数）.
- 確率変数 $X^{\setminus k} \in \mathbb{R}^{n-1}$ に対し, このエントロピー $\mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\setminus k})$ は確率変数となる.

Lemma 3.8 (Tensorization of entropy)

n 変数関数 $f : \mathbb{R}^n \to \mathbb{R}$ と, 独立な確率変数列 $\{X_k\}_{k=1}^n$ に対し,

$$\mathbb{H}(e^{\lambda f(X_1\dots,X_k)}) \leq \mathbb{E}\left[\sum_{k=1}^{n} \mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\setminus k})\right] \quad \text{for all } \lambda > 0. \tag{3.21}$$

Proof of Lemma 3.8

- まずエントロピーは以下のように表せる（Exercise 3.9）.

$$\mathbb{H}\left(e^{\lambda f(X)}\right) = \sup_{g} \left\{ \mathbb{E}\left[g(X)e^{\lambda f(X)}\right] \mid \mathbb{E}\left[e^{g(X)}\right] \leq 1 \right\}. \tag{3.24}$$

- 各 $j = 1, \ldots, n$ に対し, $X_j = (X_j \ldots, X_n)$ とする.
- g は $\mathbb{E}[e^{g(X)}] \leq 1$ を満たす関数とする.
- 関数列 $g_1, \ldots, g_n$ を以下で定義する.

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

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

- すると定義より,

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

で, さらに $\mathbb{E}[\exp(g^k(X_k, \ldots, X_n)) \mid X_{k+1}^n] = 1$.

- このとき,

$$\begin{aligned}\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] &\overset{\text{(i)}}{\leq} \sum_{k=1}^{n}\mathbb{E}\left[g^k\left(X_k,\ldots,X_n\right)e^{\lambda f(X)}\right]\\ &= \sum_{k=1}^{n}\mathbb{E}_{X_{\setminus k}}\left[\mathbb{E}_{X_k}\left[g^k\left(X_k,\ldots,X_n\right)e^{\lambda f(X)}\mid X_{\setminus k}\right]\right]\\ &\overset{\text{(ii)}}{\leq} \sum_{k=1}^{n}\mathbb{E}_{X_{\setminus k}}\left[\mathbb{H}\left(e^{\lambda f_k(X_k)}\mid X_{\setminus k}\right)\right]\end{aligned}$$

  となる, ただし (i) は (3.25), (ii) は (3.24) と $\mathbb{E}[\exp(g^k(X_k,\ldots,X_n))\mid X_{\setminus k}]=1$ より.
- これを関数 g s.t. $\mathbb{E}[e^{g(X)}]\leq 1$ に対して sup をとると, (3.24) より

$$\mathbb{H}(e^{\lambda f(X_1\ldots,X_k)})\leq \mathbb{E}\left[\sum_{k=1}^{n}\mathbb{H}(e^{\lambda f_k(X_k)}\mid X^{\setminus k})\right]$$

  が得られる. □

Proof of Theorem 3.4

- 各 $k = 1, \ldots, n$ と $x_{\setminus k} \in \mathbb{R}^{n-1}$ について, 仮定より f_k は凸かつ Lipschitz なので, Lemma 3.7 より

$$\begin{aligned}\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid x_{\setminus k}\right) &\leq \lambda^2(b-a)^2\mathbb{E}_{X_k}\left[\left(f_k'\left(X_k\right)\right)^2 e^{\lambda f_k(X_k)} \mid x_{\setminus k}\right] \\ &= \lambda^2(b-a)^2\mathbb{E}_{X_k}\left[\left(\frac{\partial f\left(x_1, \ldots, X_k, \ldots, x_n\right)}{\partial x_k}\right)^2 e^{\lambda f(x_1, \ldots, X_k, \ldots, x_n)}\right]\end{aligned}$$

- Lemma 3.8 より,

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

  となる, ただし (i) は Lipschitz 関数の性質 $\|\nabla f(x)\|_2^2 = \sum_{k=1}^{n}\left(\frac{\partial f(x)}{\partial x_k}\right)^2 \leq L^2$ より.

- したがって, Proposition 3.2 を $f(X)$ に適用すると, upper bound (3.16) が得られる. □

3.2 A geometric perspective on concentration

- Concentration of measure の幾何的な面の議論に移る.
- この章の結果は, 距離測度空間 (metric measure space)—つまり, 距離空間 $(\mathcal{X}, \rho)$ とそのボレル集合上の確率測度 $\mathbb{P}$—の言葉で示される.
- 距離空間の例としては,
  - the Euclid space $\mathcal{X} = \mathbb{R}^n$ with the usual Euclidian metric $\rho(x, y) := \|x - y\|_2$,
  - the discrete cube $\mathcal{X} = \{0, 1\}^n$ with the Hamming metric $\rho(x, y) = \sum_{j=1}^{n} \mathbb{I}[x_j \neq y_j]$

  などがある.

3.2.1 Concentration functions

- 集合 $A \subseteq \mathcal{X}$ と点 $x \in \mathcal{X}$ に対し, それらの距離を以下とする.

$$\rho(x, A) := \inf_{y \in A} \rho(x, y). \tag{3.26}$$

- パラメータ $\epsilon > 0$ に対し, A の ϵ-拡張 (ϵ-enlargement) は以下で与えられる.

$$A^{\epsilon} := \{x \in \mathcal{X} \mid \rho(x, A) < \epsilon\}. \tag{3.27}$$

Definition 3.9

距離測度空間 $(\mathcal{X}, \rho, \mathbb{P})$ の concentration function $\alpha : [0, \infty) \to \mathbb{R}_+$ は,

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) := \sup_{A \subseteq \mathcal{X}} \left\{ 1 - \mathbb{P}[A^{\epsilon}] \;\middle|\; \mathbb{P}[A] \geq \frac{1}{2} \right\} \tag{3.28}$$

で定義される, ただし sup は全ての可測部分集合 A についてとられる.

Example 3.10 (Concentration function for sphere)

- n 次元 Euclidean sphere

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\} \tag{3.29}$$

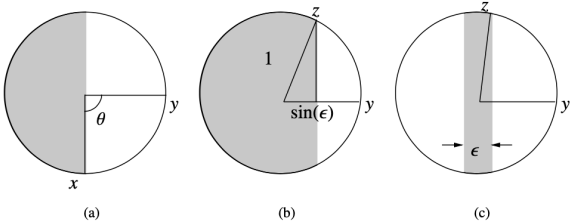
上の一様分布 $\mathbb{P}$ と, geodesic distance $\rho(x, y) := \arccos\langle x, y\rangle$ を考える.

- 各 $y \in \mathbb{S}^{n-1}$ に対し, hemisphere H_y $(\mathbb{P}(H_y) = 1/2)$ を以下で定義する. (a)

$$H_y := \{x \in \mathbb{S}^{n-1} \mid \rho(x, y) \geq \pi/2\} = \{x \in \mathbb{S}^{n-1} \mid \langle x, y\rangle \leq 0\}. \tag{3.30}$$

- この ϵ-拡張は, 以下となる. (b)

$$H_y^\epsilon = \{z \in \mathbb{S}^{n-1} \mid \langle z, y\rangle < \sin(\epsilon)\}. \tag{3.31}$$



- 一様分布であることより, concentration function は,

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) = 1 - \mathbb{P}[H_y^\epsilon]. \tag{3.32}$$

- $\sin(\epsilon) \geq \epsilon/2$ (for all $\epsilon \in (0, \pi/2]$) より, $\mathbb{P}[H_y^\epsilon] \geq \mathbb{P}[\tilde{H}_y^\epsilon]$ である, ただし

$$\tilde{H}_y^\epsilon := \{z \in \mathbb{S}^{n-1} \mid \langle z, y \rangle \leq \epsilon/2\}.$$

- さらに, 幾何的な計算により, 任意の $\epsilon \in (0, \sqrt{2})$ に対して

$$\mathbb{P}[\tilde{H}_y^\epsilon] \geq 1 - \left(1 - \left(\frac{\epsilon}{2}\right)^2\right)^{n/2} \geq 1 - e^{-n\epsilon^2/8} \tag{3.33}$$

  となる (？), ただし最後の不等号は $(1-t) \leq e^{-t}$ より.
- したがって concentration function の上限 $\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq e^{-n\epsilon^2/8}$ が得られる.
- 似たアプローチでさらに注意深く上限をとると, より厳しい上限

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}} e^{-\frac{n\epsilon^2}{2}} \tag{3.34}$$

  が得られる（？）

3.2.2 Connection to Lipschitz functions

- 関数 $f : \mathcal{X} \to \mathbb{R}$ が距離 ρ に関して L-Lipschitz であるとは, 以下が成り立つことをいう.

$$|f(x) - f(y)| \leq L\rho(x, y) \quad \text{for all } x, y \in \mathcal{X}. \tag{3.36}$$

- 確率変数 $X \sim \mathbb{P}$ に対し, m_f は $f(X)$ の任意のメディアンとする, つまり,

$$\mathbb{P}[f(X) \geq m_f] \geq 1/2 \quad \text{and} \quad \mathbb{P}[f(X) \leq m_f] \geq 1/2. \tag{3.37}$$

Proposition 3.11

- 確率変数 $X \sim \mathbb{P}$ と concentration function$\alpha_{\mathbb{P}}$ に対し, $(\mathcal{X}, \rho)$ 上の L-Lipschitz 関数 f は次を満たす.

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha_{\mathbb{P}}(\epsilon/L). \tag{3.39a}$$

- 関数 $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ は, 任意の 1-Lipschitz 関数 f に対し

$$\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \epsilon] \leq \beta(\epsilon) \quad \text{for any } \epsilon \geq 0. \tag{3.39b}$$

を満たすとする. このとき, concentraion function は bound $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\frac{\epsilon}{2})$ をもつ.

Proof of Proposition 3.11

(前半)

- 集合 $A = \{x \in \mathcal{X} \mid f(x) \le m_f\}$ と, その ϵ/L-拡張 $A^{\epsilon/L}$ について考える.
- 任意の $x \in A^{\epsilon/L}$ に対し, $\rho(x, y) < \epsilon/L$ なる $y \in A$ が存在する.
- Lipschitz 性より $|f(x) - f(y)| \le L\rho(x, y) < \epsilon$ なので,

$$A^{\epsilon/L} \subseteq \{x \in \mathcal{X} \mid f(x) < m_f + \epsilon\}. \tag{3.38}$$

- したがって,

$$\mathbb{P}[f(X) \ge m_f + \epsilon] \overset{\text{(i)}}{\le} 1 - \mathbb{P}[A^{\epsilon/L}] \overset{\text{(ii)}}{\le} \alpha_{\mathbb{P}}(\epsilon/L)$$

  が成り立つ, ただし (i) は (3.38), (ii) は $\mathbb{P}(A) \ge 1/2$ と $\alpha_{\mathbb{P}}$ の定義より.
- 同様にして $\mathbb{P}[f(X) \le m_f - \epsilon] \le \alpha_{\mathbb{P}}(\epsilon/L)$ も示されるので, (3.39a) が得られる.

(後半)

- $\epsilon \geq 0$ を固定し, A は $\mathbb{P}[A] \geq 1/2$ を満たす任意の可測集合とする.
- 関数 $f(x) := \min\{\rho(x, A), \epsilon\}$ は 1-Lipschitz（らしい）で $1 - \mathbb{P}[A^\epsilon] = \mathbb{P}[f(X) \geq \epsilon]$.
- また, 定義より $\mathbb{E}[f(X)] \leq (1 - \mathbb{P}[A])\epsilon \leq \epsilon/2$ なので,

$$\mathbb{P}\left[f(X) \geq \epsilon\right] \leq \mathbb{P}\left[f(X) \geq \mathbb{E}[f(X)] + \epsilon/2\right] \leq \beta(\epsilon/2).$$

- したがって, 任意の可測集合 A s.t. $\mathbb{P}[A] \geq 1/2$ に対して

$$1 - \mathbb{P}[A^\epsilon] \leq \beta(\epsilon/2)$$

となるので, $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$. □

Example 3.12 (Lévy concentration on $\mathbb{S}^{n-1}$)

- Example 3.10 の $(\mathbb{S}^{n-1}, \rho, \mathbb{P})$ において, concentration function の上限は以下で得られた.

$$\alpha_{\mathbb{S}^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}} e^{-\frac{n\epsilon^2}{2}}.$$

- したがって, 任意の 1-Lipschitz 関数 $f : \mathbb{S}^{n-1} \to \mathbb{R}$ に対して,

$$\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq \sqrt{2\pi} e^{-\frac{n\epsilon^2}{2}} \tag{3.40}$$

  となる.
- さらに, Exercise 2.14(d) より, 以下も成り立つ.

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq \epsilon] \leq 2\sqrt{2\pi} e^{-\frac{n\epsilon^2}{8}}. \tag{3.41}$$

Example 3.13 (Concentration for Boolean hypercube)

- Boolean hypercube $\mathcal{X} = \{0,1\}^n$ と Hamming metric $\rho_H(x,y) := \sum_{j=1}^n \mathbb{I}[x_j \neq y_j]$ を考える.
- 確率測度 $\mathbb{P}$ は $\mathcal{X}$ 上の一様分布.
- 半径 r, 中心 $x \in \mathcal{X}$ の球を Hamming ball と呼び, 以下で定義する.

$$\mathbb{B}_H(r;x) = \{y \in \{0,1\}^n \mid \rho_H(y,x) \leq r\}.$$

- 非空な部分集合 $A, B \subseteq \mathcal{X}$ に対し, Harper's combinatorial theorem より, 正整数 r_A, r_B と対応する部分集合 $A', B' \subseteq \mathcal{X}$ が存在し, 以下を満たす.
  - $\mathbb{B}_H(r_A - 1; 0) \subseteq A' \subseteq \mathbb{B}_H(r_A; 0) \quad \text{and} \quad \mathbb{B}_H(r_B - 1; 1) \subseteq B' \subseteq \mathbb{B}_H(r_B; 1)$.[2]
  - $\mathrm{card}(A) = \mathrm{card}(A'), \mathrm{card}(B) = \mathrm{card}(B')$.
  - $\rho_H(A', B') \geq \rho_H(A, B)$.
- 上の結果より, concentration function は以下のように抑えられる.

$$\alpha_{\mathbb{P}}(\epsilon) \leq \exp\left(-\frac{2\epsilon^2}{n}\right) \quad \text{for all } n \geq 3. \tag{3.42}$$

[2] $0, 1$ はそれぞれ all-zeros vector, all-ones vector.

- 任意の部分集合 A s.t. $\mathbb{P}[A] = \mathrm{card}(A)/2^n \geq 1/2$ をとる.
- 任意の $\epsilon > 0$ に対し, $B = \mathcal{X} \setminus A^\epsilon$ とする.
- (3.42) にを示すには, $\mathbb{P}[B] \leq \exp(-2\epsilon^2/n)$ を示せば良い.
- $\epsilon \leq 1$ なら $\mathbb{P}[B] \leq 1/2 \leq \exp(-2/n) \leq \exp(-2\epsilon^2/n)$ なので, 以後 $\epsilon > 1$ とする.
- 定義より,

$$\rho_H(A, B) = \min_{a \in A, b \in B} \rho(a, b) \geq \epsilon.$$

- A', B' を Harper の定理によって保証される集合とする.
- $\mathrm{card}(A) = \mathrm{card}(A') \geq 1/2$ より, A' は 1 の数が $n/2$ 以下のベクトルを全て含む.
- $\rho_H(A', B') \geq \rho_H(A, B) \geq \epsilon$ なので, B' は 1 の数が $n/2 + \epsilon$ 以上であるベクトルを全て含む.

- したがって, $\{X_i\}_{i=1}^n$ を i.i.d. Bernoulli 変数とすると, 上の議論と Hoeffding bound より,

$$\mathbb{P}[B] = \mathbb{P}[B'] \leq \mathbb{P}\left[\sum_{i=1}^n X_i \geq \frac{n}{2} + \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{n}\right)$$

  となり, A は $\mathbb{P}[A] \geq 1/2$ の任意の集合で $B = \mathcal{X} \setminus A^\epsilon$ なので, (3.42) が得られる.
- よって Proposition 3.11 より, 任意の 1-Lipschitz 関数に対し

$$\mathbb{P}\left[|f(X) - m_f| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2}{n}\right).$$

3.2.3 From geometry to concentration

- これらの幾何的な面は, 凸幾何学と concentration of measure の関連性を示唆する.
- たとえば, 次の Brunn-Minkowski inequality を考えてみよう:
  - R^n 上の任意の 2 つの compact set $C, D \subset \mathbb{R}^n$ に対し, 以下が成り立つ.

$$\mathrm{vol}(\lambda C + (1-\lambda)D) \geq \left(\lambda \,\mathrm{vol}(C)^{1/n} + (1-\lambda)\,\mathrm{vol}(D)^{1/n}\right)^n \geq \mathrm{vol}(C)^{\lambda}\,\mathrm{vol}(D)^{1-\lambda} \quad \text{for all } \lambda \in [0,1],$$

    ただし, vol は volume, つまり R^n 上の Lebesgue 測度で,

$$\lambda C + (1-\lambda)D = \{\lambda c + (1-\lambda)d \mid c \in C,\ d \in D\}.$$

- Concentration function は $\mathbb{P}[A] \geq 1/2$ のもとでの $1 - \mathbb{P}[A^\epsilon]$ の sup だが, vol を (正規化されていない) 確率測度とみると, Brunn-Minkowski inequality はこの上限の導出に使える.

Example 3.14 (Classical isoperimetric inequality in $\mathbb{R}^n$)

- $\mathbb{B}_2^n$ を $\mathbb{R}^n$ 上の Euclidean sphere とする, つまり $\mathbb{B}_2^n := \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$.
- A は $\mathrm{vol}(A) = \mathrm{vol}(\mathbb{B}_2^n)$ を満たす任意の集合 $A \subset \mathbb{R}^n$ とする.
- このとき, Brunn-Minkowski inequality において λ, C, D を適切に選ぶことで

$$[\mathrm{vol}\,(A^\epsilon)]^{1/n} = [\mathrm{vol}\,(A + \epsilon\mathbb{B}_2^n)]^{1/n} \geq [\mathrm{vol}(A)]^{1/n} + [\mathrm{vol}\,(\epsilon\mathbb{B}_2^n)]^{1/n}$$

  となる (see Exercise 3.10).
- さらに, $\mathrm{vol}(A) = \mathrm{vol}(\mathbb{B}_2^n)$ と $[\mathrm{vol}(\epsilon\mathbb{B}_2^n)]^{1/n} = \epsilon\mathrm{vol}(\mathbb{B}_2^n)^{1/n}$ より [3],

$$\mathrm{vol}(A^\epsilon)^{1/n} \geq (1+\epsilon)\mathrm{vol}(\mathbb{B}_2^n)^{1/n} = [\mathrm{vol}((\mathbb{B}_2^n)^\epsilon)]^{1/n}$$

  となるので, したがって

$$\mathrm{vol}(A^\epsilon) \geq \mathrm{vol}([\mathbb{B}_2^n]^\epsilon). \tag{3.44}$$

♣

[3] ここテキストでは $[\mathrm{vol}(\epsilon\mathbb{B}_2^n)]^{1/n} = \epsilon\mathrm{vol}(\mathbb{B}_2^n)$ となっているがおそらくタイポ.

- Brunn-Minkowski inequality は, 次の functional-analytic generalization が知られている.

Theorem 3.15 (Prékopa-Leindler inequality)

u, v, w は非負可積分関数で, ある $\lambda \in [0,1]$ が存在して,

$$w(\lambda x + (1-\lambda)y) \geq u(x)^{\lambda} v(y)^{1-\lambda} \quad \text{for all } x, y \in \mathbb{R}^n. \tag{3.46}$$

を満たすものとする. このとき,

$$\int w(x)dx \geq \left(\int u(x) \right)^{\lambda} \left(\int v(x)dx \right)^{1-\lambda}. \tag{3.47}$$

- この Prékopa-Leibler inequality は, strongly log-concave な分布のもとでの Lipschitz 関数の集中不等式の導出に使われる.

- $\mathbb{R}^n$ 上の分布 $\mathbb{P}$ with density p が strongly log-concave distribution であるとは, $\log p$ が strongly concave であることをいう.
  $\Leftrightarrow$ density が $p(x) = \exp(-\psi(x))$ とかける, ただし $\psi : \mathbb{R}^n \to \mathbb{R}$ は storongly convex, つまりある $\gamma > 0$ が存在して

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|x-y\|_2^2 \tag{3.48}$$

  を任意の $\lambda \in [0,1],\ x, y \in \mathbb{R}^n$ に対して満たす.

- n 次元標準 Gaussian 分布はパラメータ $\gamma = 1$ の strongly log-concave である.
- 一般に, covariance matrix $\Sigma \succ 0$ [4] の Gauusian 分布は $\gamma = \gamma_{\min}(\Sigma^{-1}) = (\gamma_{\max}(\Sigma))^{-1}$ で strognly log-cocave.
- non-Gaussian でも storongly log-concave な分布は多数ある（らしい）.

[4] つまり, Σ が正定値.

Theorem 3.16

$\mathbb{P}$ はパラメータ $\gamma > 0$ の strongly log-concave distribution とする. 任意の L-Lipschitz 関数 $f : \mathbb{R}^n \to \mathbb{R}$ に対し, 以下が成り立つ.

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{\gamma t^2}{4L^2}}. \tag{3.49}$$

Proof.

- h を $\mathbb{E}[h(X)] = 0$ なる任意の L-Lipschitz 関数とする.
- $\mathbb{E}[e^{h(X)}] \leq e^{L^2/\gamma}$ を示せば十分である.
  - もしこれが成り立つなら, 任意の K-Lipschitz 関数 f と $\lambda \in \mathbb{R}$ に対し, $L = \lambda K$-Lipschitz 関数 $h := \lambda(f - \mathbb{E}[f(X)])$ を考えれば,

$$\mathbb{E}[e^{\lambda(f(X) - \mathbb{E}[f(X)])}] \leq e^{\frac{\lambda^2 K^2}{\gamma}} \quad \text{for all } \lambda \in \mathbb{R}$$

  となり, $f(X) - \mathbb{E}[f(X)]$ が sub-Gaussian となるので, (3.49) の tail bound が得られる.

- h に対し, g を以下で定義する.

$$g(y) := \inf_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{\gamma}{4} \|x - y\|_2^2 \right\}.$$

- Prékopa-Leindler inequality を $\lambda = 1/2$ として使うために, 関数 u, v, w を,

$$\begin{aligned} w(z) &= p(z) = \exp(-\psi(z)), \\ u(x) &= \exp(-h(x) - \psi(x)), \\ v(x) &= \exp(g(y) - \psi(y)), \end{aligned}$$

とする, ただし p は $\mathbb{P}$ の density で, log-concavity より ψ は storongly convex.

- (3.46) の RHS の log をとったものを R とすると,

$$R = \frac{1}{2}\{g(y) - h(x)\} - \frac{1}{2}\psi(x) - \frac{1}{2}\psi(y) = \frac{1}{2}\{g(y) - h(x) - 2E(x, y)\} - \psi(x/2 + y/2),$$

where $E(x, y) := \frac{1}{2}\psi(x) + \frac{1}{2}\psi(y) - \psi(x/2 + y/2)$.

- ψ が strongly convex なので, $2E(x,y) \geq \frac{\gamma}{4}\|x-y\|_2^2$.
- これを代入すると,

$$R \leq \frac{1}{2}\left\{g(y) - h(x) - \frac{\gamma}{4}\|x-y\|_2^2\right\} - \psi(x/2+y/2) \leq -\psi(x/2+y/2)$$

  となる (ただし最後の不等号は g の定義より) ので, (3.46) は $\lambda = 1/2$ で満たされる.
- Prékopa-Leindler inequality と $\int w(x)dx = \int p(x)dx = 1$ より, (3.47) の log をとって

$$0 \geq \frac{1}{2}\log \int e^{-h(x)-\psi(x)}dx + \frac{1}{2}\log \int e^{g(y)-\psi(y)}dy.$$

- これを書き換えると,

$$\mathbb{E}[e^{g(Y)}] \leq \frac{1}{\mathbb{E}\left[e^{-h(X)}\right]} \overset{\text{(i)}}{\leq} \frac{1}{e^{\mathbb{E}[-h(X)]}} \overset{\text{(ii)}}{=} 1$$

  となる, ただし (i) は Jensen, (ii) は $\mathbb{E}[h(X)] = 0$ より.

- また, h の Lipschitz 性 $|h(x) - h(y)| \leq L\|x - y\|_2$ より,

$$g(y) = \inf_{x \in \mathbb{R}^n} \left\{ h(x) + \frac{\gamma}{4}\|x - y\|_2^2 \right\} \geq h(y) + \inf_{x \in \mathbb{R}^n} \left\{ -L\|x - y\|_2 + \frac{\gamma}{4}\|x - y\|_2^2 \right\}$$
$$= h(y) - \frac{L^2}{\gamma}.$$

- したがって, $\mathbb{E}[e^h(X)] \leq \exp(L^2/\gamma)$ となる. □

3.3 Wasserstein distances and information inequalities

- Wasserstein distances と information inequalities (transportation cost inequalities) について

3.3.1 Wasserstein distances

- 距離空間 $(\mathcal{X}, \rho)$ は所与とし, 関数 $f : \mathcal{X} \to \mathbb{R}$ のノルム $\|f\|_{\mathrm{Lip}}$ を

$$\|f\|_{\mathrm{Lip}} := \inf\{L \geq 0 : |f(x) - f(x')| \leq L\rho(x, x') \ \forall x, x' \in \mathcal{X}\}$$

  で定義する. つまり $\|f\|_{\mathrm{Lip}}$ は f が L-Lipschitz であるような最小の L.
- $\mathcal{X}$ 上の確率分布 $\mathbb{Q}, \mathbb{P},$ の距離 $W_\rho(\mathbb{Q})$ を

$$W_\rho(\mathbb{Q}, \mathbb{P}) = \sup_{f \text{ s.t. } \|f\|_{\mathrm{Lip}} \leq 1} \left[\int f d\mathbb{Q} - \int f d\mathbb{P} \right] \tag{3.52}$$

  で定義し, これを Wasserstein metric induced by ρ とよぶ.
  - 任意の ρ に対しこれが確率分布空間上の距離になることは確かめられる.

Example 3.17 (Hamming metric and total variation distance)

- Hamming metric $\rho(x, x') = \mathbb{I}[x \neq x']$ を考える.
- このとき Wasserstein distance は, 以下で定義される total variation distance と等しい.

$$\|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}} := \sup_{A \subseteq \mathcal{X}} |\mathbb{Q}(A) - \mathbb{P}(A)| \tag{3.53}$$

- Hamming 距離のもとでは, $\|f\|_{\mathrm{Lip}} \leq 1$ は $|f(x) - f'(x)| \leq 1$ $(\forall x, x')$ と同値.
- (3.52) において f の定数移動は影響を与えないので $f(x) \in [0, 1]$ に限定してよい, よって

$$W_{\mathrm{Ham}}(\mathbb{Q}, \mathbb{P}) = \sup_{f: \mathcal{X} \to [0,1]} \int f \, (d\mathbb{Q} - d\mathbb{P}) \overset{\text{(i)}}{=} \|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}}$$

  となる, ただし (i) は Exercise 3.13 参照.
- それぞれ ν を base measure としたときの density p, q が存在するとすると,[5]

$$W_{\mathrm{Ham}}(\mathbb{Q}, \mathbb{P}) = \|\mathbb{Q} - \mathbb{P}\|_{\mathrm{TV}} = \frac{1}{2} \int |p(x) - q(x)| \nu(dx)$$

  となり, $L^1(\nu)$-norm の $1/2$ にも一致する (see Exercise 3.13). ♣

[5] $\nu = \frac{1}{2}(\mathbb{P} + \mathbb{Q})$ とすると $\mathbb{P}, \mathbb{Q}$ は density をもつので, これは一般性を失わない.

- Wasserstein distance には coupling を用いた同値な定義がある.
- Product space $\mathcal{X} \otimes \mathcal{X}$ 上の分布 $\mathbb{M}$ がペア $(\mathbb{Q}, \mathbb{P})$ の coupling であるとは, その marginal distribution が $\mathbb{Q}, \mathbb{P}$ に一致するときをいう.
- $f : \mathcal{X} \to \mathbb{R}$ を 1-Lipschitz 関数とすると,

$$\int \rho\left(x, x'\right) d\mathbb{M}\left(x, x'\right) \overset{\text{(i)}}{\geq} \int \left(f(x) - f\left(x'\right)\right) d\mathbb{M}\left(x, x'\right) \overset{\text{(ii)}}{=} \int f(d\mathbb{P} - d\mathbb{Q}) \tag{3.54}$$

が成り立つ, ただし (i) は Lipschitz 性, (ii) は $\mathbb{M}$ が coupling であることより.

- *Kantorovich-Rubinstein duality* によると, coupling について minimum をとると次の同値性が成り立つ:

$$\underbrace{\sup_{\|f\|_{\text{Lip}} \leq 1} \int f(d\mathbb{Q} - d\mathbb{P})}_{W_\rho(\mathbb{P}, \mathbb{Q})} = \inf_{\mathbb{M}} \int_{\mathcal{X} \times \mathcal{X}} \rho\left(x, x'\right) d\mathbb{M}\left(x, x'\right) = \inf_{\mathbb{M}} \mathbb{E}_{\mathbb{M}}\left[\rho\left(X, X'\right)\right] \tag{3.55}$$

ただし, infimum はペア $(\mathbb{P}, \mathbb{Q})$ のすべての coupling についてとる.

- Wasserstein distance はこの coupling による表現から “transportation cost” とも呼ばれる.
- $\mathbb{P}, \mathbb{Q}$ は $\mathcal{X}$ 上の Lebesgue measure を base として density p, q を持つとし, $p(x), q(x)$ は点 $x \in \mathcal{X}$ に堆積した土の量と解釈する.
- そして土を移動させることで, 土の堆積状態を p から q へ変化させること考える.
- ただし点 x から点 x' へ土を 1 単位移動させるにはコスト $\rho(x, x')$ がかかる.
- joint distribution $m(x, x')$ を x から x' へ移す土の量とみると, m は p から q への transportation plan と解釈できる.
- この transportation plan にかかるコストは

$$\int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')m(x, x')dxdx'$$

となるので, Wasserstein distance(3.55) は最小の transportation cost と考えられる.

3.3.2 Transportation cost and concentration inequalities

- Transportation cost inequality とその concentration inequalities への応用の議論に移る.
- *Kullback-Leibler(KL) divergence*: 2 つの分布 $\mathbb{Q}, \mathbb{P}$ に対し, それらの KL-divergence は以下で与えられる.

$$D(\mathbb{Q}\|\mathbb{P}) := \begin{cases} \mathbb{E}_{\mathbb{Q}}\left[\log \frac{d\mathbb{Q}}{d\mathbb{P}}\right] & \text{when } \mathbb{Q} \text{ は } \mathbb{P} \text{ に関して絶対連続,} \\ +\infty & \text{otherwise.} \end{cases} \tag{3.56}$$

- もし $\mathbb{Q}, \mathbb{P}$ が base measure ν に関して density q, p をもつなら, KL-divergence は以下でかける.

$$D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \nu(dx). \tag{3.57}$$

- KL-divergence は分布の差異を表すものと解釈できるが, 対称でないので metric ではない.

Definition 3.18 (information inequalities)

確率測度 $\mathbb{P}$ が ρ-transportation cost inequality with parameter $\gamma > 0$ を満たすとは, 任意の確率測度 $\mathbb{Q}$ に対して

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

を満たすことをいう.

- 例 *Pinsker–Csiszár–Kullback inequality*: 任意の確率分布 $\mathbb{P}, \mathbb{Q}$ に対し以下が成り立つ.

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})} \tag{3.59}$$

  Example 3.17 より, Hamming metric $\rho(x, x') = \mathbb{I}[x \neq x']$ のもとで (3.59) は $\gamma = 1/4$ の information inequality に対応する.

- 次の定理で information inequality は Lipschitz 関数の concentration bound を導く.

Theorem 3.19 (From transportation cost to concentration)

距離測度空間 $(\mathbb{P}, \mathcal{X}, \rho)$ において, $\mathbb{P}$ は ρ-transportation cost inequality (3.58) をパラメータ $\gamma > 0$ で満たすとする. このとき, concentraion function は以下の bound をもつ.

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(t) \leq 2\exp\left(-\frac{t^2}{2\gamma}\right) \tag{3.60}$$

さらに, 確率変数 $X \sim \mathbb{P}$ と L-Lipschitz 関数 $f : \mathcal{X} \to \mathbb{R}$ に対し, 次の concentration inequality が成り立つ.

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right). \tag{3.61}$$

Remarks:

- Proposition 3.11 より, (3.60) は f のメディアン m_f 周りの以下の concentration inequality も導く.

$$\mathbb{P}\left[|f(X) - m_f| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right). \tag{3.62}$$

Proof.
(前半)

- $A \subset \mathcal{X}$ with $\mathbb{P}[A] \geq 1/2$ を任意にとり, $\epsilon > 0$ に対し次の集合 B を考える.

$$B := (A^\epsilon)^c = \{y \in \mathcal{X} \mid \rho(x, y) \geq \epsilon \ \forall x \in A\}.$$

- まず定義より $\rho(A, B) := \inf_{x \in A} \inf_{y \in B} \rho(x, y) \geq \epsilon$.
- $\mathbb{P}_A, \mathbb{P}_B$ は, $\mathbb{P}$ を A, B に条件づけた分布とする.
- $(\mathbb{P}_A, \mathbb{P}_B)$ の任意の coupling $\mathbb{M}$ に対し, $\rho(A, B) \leq \int \rho(x, x') d\mathbb{M}(x, x')$.
- $\mathbb{M}$ について inf をとると, $\epsilon \leq \rho(A, B) \leq W_\rho(\mathbb{P}_A, \mathbb{P}_B)$.
- よって,

$$\epsilon \overset{\text{(i)}}{\leq} W_\rho\left(\mathbb{P}_A, \mathbb{P}_B\right) \leq W_\rho\left(\mathbb{P}, \mathbb{P}_A\right) + W_\rho\left(\mathbb{P}, \mathbb{P}_B\right) \overset{\text{(ii)}}{\leq} \sqrt{\gamma D\left(\mathbb{P}_A \| \mathbb{P}\right)} + \sqrt{\gamma D\left(\mathbb{P}_B \| \mathbb{P}\right)}$$
$$\overset{\text{(iii)}}{\leq} \sqrt{2\gamma}\left\{D\left(\mathbb{P}_A \| \mathbb{P}\right) + D\left(\mathbb{P}_B \| \mathbb{P}\right)\right\}^{1/2}$$

となる, ただし (i) は三角不等式, (ii) は transportation cost inequality の仮定, (iii) は $(a+b)^2 \leq 2a^2 + 2b^2$ より.

- また, KL-divergence は $D(\mathbb{P}_A\|\mathbb{P}) = \log\frac{1}{\mathbb{P}(A)}, D(\mathbb{P}_B\|\mathbb{P}) = \log\frac{1}{\mathbb{P}(B)}$ となる.
  - なぜなら, 適当な base measure のもとで $\mathbb{P}, \mathbb{P}_A$ は density p, p_A をもち,

$$\frac{d\mathbb{P}_A}{d\mathbb{P}}(x) = \frac{p_A(x)}{p(x)} = \frac{(\mathbb{I}[x \in A] \cdot p(x)/\mathbb{P}(A))}{p(x)} = \frac{\mathbb{I}[x \in A]}{\mathbb{P}(A)}$$

  より,

$$D(\mathbb{P}_A\|\mathbb{P}) = \mathbb{E}_{\mathbb{P}_A}\left[\log\frac{d\mathbb{P}_A}{d\mathbb{P}}\right] = \int_{x\in\mathcal{X}} \log\frac{\mathbb{I}[x \in A]}{\mathbb{P}(A)}\mathbb{P}_A(dx) = \int_{x\in A} \log\frac{1}{\mathbb{P}(A)}\mathbb{P}_A(dx) = \log\frac{1}{\mathbb{P}(A)}.$$

- よって,

$$\epsilon^2 \le 2\gamma\{\log(1/\mathbb{P}(A)) + \log(1/\mathbb{P}(B))\} = 2\gamma\log\left(\frac{1}{\mathbb{P}(A)\mathbb{P}(B)}\right)$$

- したがって, $\mathbb{P}(B) \le \exp(-\epsilon^2/2\gamma)/\mathbb{P}(A) \le 2\exp(-\epsilon^2/2\gamma)$ となる ($\mathbb{P}(A) \ge 1/2$ より).
- $A \subset \mathcal{X}$ は $\mathbb{P}(A) \ge 1/2$ なる任意の集合で, $B = (A^\epsilon)^c$ だったので, (3.60) が成り立つ.

(後半)

- $f : \mathcal{X} \to \mathbb{R}$ は L-Lipschitz 関数, $\mathbb{Q}$ は任意の分布とする.
- f の Lipschitz 性, Wasserstein の定義と information inequality より,

$$\int f(d\mathbb{Q} - d\mathbb{P}) \leq LW_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2L^2\gamma D(\mathbb{Q}\|\mathbb{P})}$$

- 任意の $u, v, \lambda > 0$ に対し $\sqrt{2uv} \leq \frac{u}{2}\lambda + \frac{v}{\lambda}$ なので, $u = L^2\gamma, v = D(\mathbb{Q}\|\mathbb{P})$ とすると

$$\int f(d\mathbb{Q} - d\mathbb{P}) \leq \frac{\lambda\gamma L^2}{2} + \frac{D(\mathbb{Q}\|\mathbb{P})}{\lambda} \quad \text{for all } \lambda > 0. \tag{3.63}$$

- 分布 $\mathbb{Q}$ として, Radon-Nikodym derivative が $\frac{d\mathbb{Q}}{d\mathbb{P}}(x) = e^{g(x)}/\mathbb{E}_{\mathbb{P}}[e^{g(X)}]$, where $g(x) = \lambda(f(x) - \mathbb{E}_{\mathbb{P}}[f(X)]) - L^2\gamma\lambda^2/2$ となるものを考えると,

$$D(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log\left(\frac{e^{g(X)}}{\mathbb{E}_{\mathbb{P}}[e^{g(X)}]}\right)\right] = \lambda \underbrace{\{\mathbb{E}_{\mathbb{Q}}[f(X)] - \mathbb{E}_{\mathbb{P}}[f(X)]\}}_{\int f(d\mathbb{Q}-d\mathbb{P})} - \frac{\gamma L^2\lambda^2}{2} - \log \mathbb{E}_{\mathbb{P}}[e^{g(X)}].$$

  となるので, (3.63) にいれると $\log \mathbb{E}_{\mathbb{P}}[e^{g(X)}] \leq 0$, つまり

$$\mathbb{E}_{\mathbb{P}}[e^{\lambda(f(X)-\mathbb{E}_{\mathbb{P}}[f(X')])}] \leq e^{\frac{\lambda^2\gamma L^2}{2}}.$$

- よって Chernoff bound より upper tail bound が得られる.
- 同じ議論を $-f$ に対して行えば lower tail bound も得られる. □

3.3.3 Tensorization for transportation cost

- Transportation cost inequality を満たす確率分布の product は Transportation cost inequality を満たす.

Proposition 3.20

各 $k = 1, 2, \ldots, n$ に対し, 1 変数分布 $\mathbb{P}_k$ は距離 ρ_k のもとでパラメータ γ_k で ρ_k-transportation cost inequality を満たすとする. Product distribution $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$ は transportation cost inequality

$$W_\rho(\mathbb{Q}, \mathbb{P}) \le \sqrt{2 \left(\sum_{k=1}^n \gamma_k \right) D(\mathbb{Q} \| \mathbb{P})} \quad \text{for all distributions } \mathbb{Q} \tag{3.64}$$

を満たす, ただし距離は $\rho(x, y) := \sum_{k=1}^n \rho_k(x_k, y_k)$.

Example 3.21 (Bounded differences inequality)

- f は bounded differences inequality を各 coordinate k についてパラメータ L_k で満たす.
  - つまり, $f : \mathbb{R}^n \to \mathbb{R}$ は各 $k = 1, \dots, n$ について

$$|f(x_1, \dots, x_{k-1}, x_k, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots, x_n)| \le L_k. \quad \text{for all } x, x'_k$$

- このとき, f は rescaled Hamming metric

$$\rho(x, y) := \sum_{k=1}^{n} \rho_k(x_k, y_k), \quad \text{where } \rho_k(x_k, y_k) := L_k \mathbb{I}[x_k \neq y_k]$$

  に関して 1-Lipschitz であることが確かめられる.
- Pinsker-Csiszár-Kullback inequality (3.59) より, 各 $\mathbb{P}_k$ はパラメータ $\gamma_k = L_k^2/4$ で ρ_k-transportation cost inequality を満たす.
- よって Proposition 3.20 より $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$ はパラメータ $\gamma := \frac{1}{4} \sum_{k=1}^n L_k^2$ で ρ-transportation cost inequality を満たすので, f の Lipschitz 性と Theorem 3.19 より

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \ge t] \le 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2} \right) \tag{3.65}$$

Proof of Proposition 3.20

- $\mathbb{Q}$ を $\mathcal{X}^n$ 上の任意の分布とする.
- ペア $(\mathbb{P}, \mathbb{Q})$ のカップリング $\mathbb{M}$ に対し, $\mathbb{M}_1^j$ を $(X_1^j, Y_1^j) = (X_1, \ldots, X_j, Y_1, \ldots, Y_j)$ の joint distribution, $\mathbb{M}_{j|j-1}$ を (X_1^{j-1}, Y_1^{j-1}) に条件づけた (X_j, Y_j) の条件付き分布とする.
- Wasserstein metric の coupling 表現 (3.55) より,

$$W_\rho(\mathbb{Q}, \mathbb{P}) \le \mathbb{E}_{\mathbb{M}}[\rho(X, Y)] = \mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1, Y_1)] + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{M}_1^{j-1}} \left[\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)] \right].$$

- Coupling $\mathbb{M}$ を次のように構成していく:
- $\mathbb{M}_1$ をペア $(\mathbb{P}_1, \mathbb{Q}_1)$ に対する optimal-coupling とすると,

$$\mathbb{E}_{M_1}\left[\rho_1\left(X_1, Y_1\right)\right] \overset{\text{(i)}}{=} W_{\rho_1}\left(Q_1, \mathbb{P}_1\right) \overset{\text{(ii)}}{\leq} \sqrt{2\gamma_1 D\left(\mathbb{Q}_1 \| \mathbb{P}_1\right)}$$

  ただし (i) は coupling の optimality, (ii) は transportation-cost inequality の仮定より.
- 次に (X_1^{j-1}, Y_1^{j-1}) の joint distribution $\mathbb{M}_1^{j-1}$ が定義されてる時に, (X_j, Y_j) の条件付き分布 $\mathbb{M}_{j|j-1}(\cdot \mid x_1^{j-1}, y_1^{j-1})$ を, ペア $(\mathbb{P}_j, \mathbb{Q}_{j|j-1}(\cdot \mid y_1^{j-1}))$ の optimal coupling で定義する.
- すると, 各 y_1^{j-1} に対して

$$\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j, Y_j)] = W_{\rho_j}\left(\mathbb{Q}_{j|j-1}, \mathbb{P}_j\right) \leq \sqrt{2\gamma_j D(\mathbb{Q}_{j|j-1} \| \mathbb{P}_j)}.$$

- $Y_1^j \sim \mathbb{Q}_1^{j-1}$ について期待値をとると, $\sqrt{\cdot}$ の concavity と Jensen の不等式から,

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

- したがって,

$$W_\rho(\mathbb{Q},\mathbb{P}) \le \sqrt{2\gamma_1 D\left(\mathbb{Q}_1\|\mathbb{P}_1\right)} + \sum_{j=2}^{n} \sqrt{2\gamma_j \mathbb{E}_{Q_1^{j-1}}\left[D\left(\mathbb{Q}_{j|j-1}\left(\cdot \mid Y_1^{j-1}\right)\|\mathbb{P}_j\right)\right]}$$

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

$$\overset{\text{(ii)}}{=} \sqrt{2\left(\sum_{j=1}^{n}\gamma_j\right) D(\mathbb{Q}\|\mathbb{P})},$$

となる, ただし (i) は Cauchy-Schwarz, (ii) は KL-divergence の chain rule より (see Exercise 3.2). □

3.3.4 Transportation cost inequalities for Markov chains

- ここでは dependent random variables の Lipschitz 関数への応用, 特に Markov chain について考える.
- $(X_1, \dots, X_n)$ は Markov chain から発生する random vector とする, ただし $X_i \in \mathcal{X}$ で $\mathcal{X}$ は countable.
- その分布 $\mathbb{P}$ over $\mathcal{X}^n$ は, initial distribution $X_1 \sim \mathbb{P}_1$ と, transition kernels

$$\mathbb{K}_{i+1}(x_{i+1} \mid x_i) = \mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i) \tag{3.66}$$

  によって定義される.
- discrete state Markov chain が β-contractive であるとは, $\beta \in [0, 1)$ に対して

$$\max_{i=1,\dots,n-1} \sup_{x_i, x_i'} \|\mathbb{K}_{i+1}(\cdot \mid x_i) - \mathbb{K}_{i+1}(\cdot \mid x_i')\|_{\mathrm{TV}} \leq \beta \tag{3.67}$$

  となることをいう.

Theorem 3.22

$\mathbb{P}$ は, discrete space $\mathcal{X}^n$ 上の β-cntractive な Markov chain の分布とする. このとき, 任意の分布 $\mathbb{Q}$ over $\mathcal{X}^n$ に対して,

$$W_\rho(\mathbb{Q}, \mathbb{P}) \leq \frac{1}{1-\beta}\sqrt{\frac{n}{2}D(\mathbb{Q}\|\mathbb{P})} \tag{3.68}$$

が成り立つ, ただし ρ は Hamming metric $\rho(x, y) = \sum_{i=1}^n \mathbb{I}[x_i \neq y_i]$.

Remark:

- 証明は省略 (see Marton (1996b)).
- このとき, Theorem 3.19 より, Hamming metric に関して L-Lipschitz な関数 $f : \mathcal{X}^n \to \mathbb{R}$ について,

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq t\right] \leq 2\exp\left(\frac{2(1-\beta)^2 t^2}{nL^2}\right). \tag{3.69}$$

- independent random variables の bounded differences inequality を満たす関数についての結果 (Example 3.21, $\beta = 0$ のときに対応) の一般化とみなせる.

Example 3.23 (Parameter estimation for a binary Markov chain)

- 二値変数 $X_i \in \{0,1\}$ の Markov chain で, 初期分布は $\mathbb{P}_1$ は uniform, transition kernel は

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

  なるものを考える, ただし $\delta \in [0,1]$ は "stickness" parameter とよぶ.
- 長さ n のベクトル $(X_1, \ldots, X_n)$ からこのパラメータ δ を推定する.
- $\frac{1+\delta}{2}$ の不偏推定量は次の関数である:

$$f(X_1, \ldots, X_n) := \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{I}[X_i = X_{i+1}].$$

- このとき, この Markov chain は $\beta = \delta$ で β-cntractive であり, f は Hamming metric に関して $\frac{2}{n-1}$-Lipschitz なので, Theorem 3.22 から f の concentration inequality

$$\mathbb{P}\left[\left|f(X) - \frac{1+\delta}{2}\right| \geq t\right] \leq 2\exp\left(-\frac{(n-1)^2(1-\delta)^2t^2}{2n}\right) \leq 2\exp\left(-\frac{(n-1)(1-\delta)^2t^2}{4}\right) \tag{3.70}$$

  が得られる.

3.3.5 Asymmetric coupling cost

- 次の (非対称な) 確率分布の距離を考える:

$$\mathcal{C}(\mathbb{Q},\mathbb{P}) := \inf_{\text{coupling } \mathbb{M}} \sqrt{\int \sum_{j=1}^{n} \left(\mathbb{M}[Y_i \neq x_i \mid X_i = x_i]\right)^2 d\mathbb{P}(x)} \tag{3.71}$$

- これは, product distribution に関する Pinsker-type inequality を満たす:
  任意の n 変数の product distribution $\mathbb{P}$ に対して,

$$\max\{\mathcal{C}(\mathbb{Q},\mathbb{P}), \mathcal{C}(\mathbb{P},\mathbb{Q})\} \leq \sqrt{2D(\mathbb{Q}\|\mathbb{P})} \tag{3.73}$$

  が全ての n-次元分布 $\mathbb{Q}$ について成り立つ (see Samson (2000)).
- この距離を用いて, convex かつ Lipschitz な関数の次の concentration inequality が示される.

Theorem 3.24

独立な確率変数列 $(X_1, \ldots, X_n)$ with $X_i \in [0,1]$ を考える. $f : \mathbb{R}^n \to \mathbb{R}$ は convex かつ Euclidean norm に関して L-Lipschitz とする. このとき, 任意の $t \geq 0$ に対して,

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{3.74}$$

Remark:

- 同じ bound が concave かつ Lipschitz 関数についても成り立つ.
- Theorem 3.4 で弱い条件 (separate convexity) のもとで upper tail bound を求めたが, ここではより強い条件 (jointly convex or concave) のもとで両側 bound を求める.

Example 3.25 (Rademacher revisited)

- Example 3.5 でみたように, Rademacher complexity of $\mathcal{A} \subseteq \mathbb{R}^n$

$$Z \equiv Z(\epsilon_1, \ldots, \epsilon_n) := \sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k \epsilon_k$$

  を考える, ただし $\epsilon_k \in \{1, -1\}$ は i.i.d. な Rademacher 変数列.
- Example 3.5 より, Z は jointly convex かつ Eucledean norm に関してパラメータ $\mathcal{W}(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$ で Lipschitz.
- よって Theorem 3.24 より,

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\mathcal{W}^2(\mathcal{A})}\right) \tag{3.75}$$

  が成り立ち, この bound は Example 3.5 で求めたもの (3.17) より強い. ♣

Proof of Theorem 3.24

- f は convex なので $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ であり, また $x_j, y_j \in [0,1]$ より $|x_j - y_j| \leq \mathbb{I}[x_j \neq y_j]$ なので,

$$f(y) - f(x) \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}[x_j \neq y_j].$$

- ペア $(\mathbb{P}, \mathbb{Q})$ の任意の coupling $\mathbb{M}$ について, $(X, Y) \sim \mathbb{M}$ として上の期待値をとると

$$\begin{aligned}
\int f(y) d\mathbb{Q}(y) - \int f(x) d\mathbb{P}(x) &\leq \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}\left[x_j \neq y_j\right] d\mathbb{M}(x, y) \\
&= \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{M}\left[X_j \neq y_j \mid Y_j = y_j\right] d\mathbb{Q}(y) \\
&\leq \int \|\nabla f(y)\|_2 \sqrt{\sum_{j=1}^{n} \mathbb{M}^2\left[X_j \neq y_j \mid Y_j = y_j\right]} d\mathbb{Q}(y),
\end{aligned}$$

となる, ただし最後の不等号は Cauchy-Schwarz より.

- Lipschitz 性と convexity より $\|\nabla f(y)\|_2 \leq L$ almost everywhere, よって

$$\begin{aligned}
\int f(y)d\mathbb{Q}(y) - \int f(x)d\mathbb{P}(x) &\leq L \int \left\{ \sum_{j=1}^{n} \mathbb{M}^2 \left[X_j \neq y_j \mid Y_j = y_j \right] \right\}^{1/2} d\mathbb{Q}(y) \\
&\leq L \left[\int \sum_{j=1}^{n} \mathbb{M}^2 \left[X_j \neq y_j \mid Y_j = y_j \right] d\mathbb{Q}(y) \right]^{1/2} \\
&= L\mathcal{C}(\mathbb{P}, \mathbb{Q}) \\
&\leq L\sqrt{2D(\mathbb{Q}\|\mathbb{P})}
\end{aligned}$$

  となる, ただし最後の不等号は (3.73) より.
- 以下, Theorem 3.19 の後半の証明と同様にすれば, f の upper tail bound が得られる.

- convex 関数の lower tail bound を求めるには, concave 関数の upper bound を求められればよい.
- $g : [0,1]^n \to \mathbb{R}$ は concave かつ L-Lipschitz とする.
- このとき,

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle \leq g(x) + \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I}[x_j \neq y_j]$$

より, 先ほどと同様にして

$$\int g d\mathbb{Q}(y) - \int g d\mathbb{P}(x) \leq \int \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I}[x_j \neq y_j] d\mathbb{M}(x,y) \leq L\mathcal{C}(\mathbb{Q}, \mathbb{P})$$

となり, g の upper bound が得られる. □

- Theorem 3.24 は independent random variables について述べたが, 独立性は information inequality (3.73) を用いるときにしか使っていない.
- しかし, (3.73) は必ずしも独立でない多くの確率変数列について成り立つ.
- よって, n-次元分布 $\mathbb{P}$ についてある $\gamma > 0$ が存在して任意の分布 $\mathbb{Q}$ について

$$\max\{\mathcal{C}(\mathbb{Q}, \mathbb{P}), \mathcal{C}(\mathbb{P}, \mathbb{Q})\} \leq \sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})} \tag{3.76}$$

が成り立つなら, 同じ証明により任意の convex Lipschitz 関数について以下が成り立つ.

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right). \tag{3.77}$$

- 例えば, β-contractive な Markov chain は (3.76) を $\gamma = \left(\frac{1}{1-\sqrt{\beta}}\right)^2$ で満たすので,

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{(1-\sqrt{\beta})^2 t^2}{2L^2}\right). \tag{3.78}$$

が成り立ち, これは (3.76) と異なり dimension-independent となっている.

3.4 Tail bounds for empirical processes

- この章で紹介した concentration inequality の empirical processes への応用を紹介する.
- この節の例は, Chapter4 の議論から statistical motivation が得られる.

- $\mathcal{F}$ は, 関数 $f : \mathcal{X} \to \mathbb{R}$ の集合.
- $(X_1, \ldots, X_n)$ は product distribution $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$ からひかれる.
- $\mathbb{P}_i$ のサポートは $\mathcal{X}_i (\subseteq \mathcal{X})$ に含まれるとする.
- 次の確率変数 Z を考える.

$$Z = \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) \right\}. \tag{3.79}$$

- 目標は, tail event $\{Z \geq \mathbb{E}[Z] + \delta\}$ の upper bound を求めこと.

- Remark: $\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^n f(X_i)|$ について考えたいなら, (3.79) で $\tilde{\mathcal{F}} = \mathcal{F} \cup \{-\mathcal{F}\}$ とすれば良い.

3.4.1 A functional Hoeffding inequality

- Hoeffding type bound の議論から始める.

Theorem 3.26 (Functional Hoeffding theorem)

各 $f \in \mathcal{F}$ と $i = 1, \ldots, n$ について, ある実数 $a_{i,f} \leq b_{i,f}$ が存在して, $f(x) \in [a_{i,f}, b_{i,f}]$ $\forall x \in \mathcal{X}_i$ が成り立つとする. このとき, 任意の $\delta \geq 0$ に対して以下が成り立つ.

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

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

Remark:

- $\mathcal{F} = \{f\}$ with $f(x) = x$ なら $Z = \frac{1}{n}\sum X_i$ となり, $x_i \in [a_i, b_i]$ に対し Theorem 3.26 は

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq \delta\right] \leq \exp\left(-\frac{n\delta^2}{4L^2}\right) \quad \text{where } L^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$$

  と classical Hoeffding theorem(ただし定数項は optimal でない) を導く.

- 関数 $(x_1, \ldots, x_n) \mapsto Z(x_1, \ldots, x_n)$ はパラメータ $L_i := \sup_{f\in\mathcal{F}} |b_{i,f} - a_{i,f}|$ (for each i) の bounded difference inequality を満たすので, Corollary 2.21 から sub-Gaussian tail bound を得られるが, そのときのパラメータは

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

  各 i について別々に f の sup をとっているので, Theorem 3.26 の L^2 の方がずっと小さい.

Proof of Theorem 3.26

- 有限な $\mathcal{F}$ について示ばよい.
  - General result は有限集合の増加列の limit をとれば示される.
- また, まずは Z の unrescaled version $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ について考える.
- 各 $j = 1, \dots, n$ について, random function Z_j を以下で定義する.

$$x_j \mapsto Z_j(x_j) = Z(X_1, \dots, X_{j-1}, x_j, X_{j+1}, \dots, X_n).$$

- Entropy の tensorization Lemma 3.8 と Lemma 3.7 の bound より,

$$\mathbb{H}(e^{\lambda Z(X)}) \le \lambda^2 \mathbb{E}\left[\sum_{j=1}^n \mathbb{E}\left[(Z_j(X_j) - Z_j(Y_j))^2 \mathbb{I}[Z_j(X_j) \ge Z_j(Y_j)] e^{\lambda Z(X)} \mid X^{\setminus j}\right]\right]. \tag{3.81}$$

- 各 $f \in \mathcal{F}$ について, $\mathcal{A}(f) := \{(x_1, \dots, x_n) \in \mathbb{R} \mid Z = \sum_{i=1}^n f(x_i)\}$ とする.
  - つまり, Z の maximum の部分を f が達成するような x の集合.
  - tie がある場合には, 適当に決めて $\mathcal{A}(f)$ が disjoint になるようにする.

- 各 x について, ある f が存在して $x \in \mathcal{A}(f)$ で,

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

- $Z_j(x_j) \geq Z_j(y_j)$ である限り, 両辺の 2 乗についても同じ不等号が成り立つ.
- $\mathcal{A}(f)$ は $\mathcal{X}$ の disjoint な分割となっているので,

$$\left(Z_j(x_j) - Z_j(y_j)\right)^2 \mathbb{I}[Z_j(x_j) \geq Z_j(y_j)] \leq \sum_{f \in \mathcal{F}} \mathbb{I}[x \in \mathcal{A}(f)] \left(f(x_j) - f(y_j)\right)^2. \quad (3.82)$$

- $(f(x_j) - f(y_j))^2 \leq (b_{j,f} - a_{j,f})^2$ より, j について足しあげると

$$\begin{aligned}\sum_{j=1}^{n} \left(Z_j\left(x_j\right) - Z_j\left(y_j\right)\right)^2 \mathbb{I}\left[Z_j\left(x_j\right) \geq Z_j\left(y_j\right)\right] e^{\lambda Z(x)} &\leq \sum_{h \in \mathcal{F}} \mathbb{I}[x \in \mathcal{A}(h)] \sum_{k=1}^{n} \left(b_{k,h} - a_{k,h}\right)^2 e^{\lambda Z(x)} \\ &\leq \sup_{f \in \mathcal{F}} \sum_{j=1}^{n} \left(b_{j,f} - a_{j,f}\right)^2 e^{\lambda Z(x)} \\ &= nL^2 e^{\lambda Z(x)}.\end{aligned}$$

- (3.81) に入れると,

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

- よって Proposition 3.2 より,

$$\mathbb{P}[Z \geq \mathbb{E}Z + t] \leq \exp\left(-\frac{t^2}{4nL^2}\right).$$

- (3.80) の rescaled version の Z については, $t = n\delta$ とすれば良い. □

3.4.2 A functional Bernstein inequality

- 次は Bernstein type bound.
- ここでは $\mathcal{F}$ は countable かつ uniformly bounded by b (i.e., $\forall f \in \mathcal{F},\ \forall x \in \mathcal{X},\ |f(x)| \leq b$) とする.

Theorem 3.27 (Talagrand concentration for empirical processes)

$\mathcal{F}$ は countable かつ uniformly bound by b とする. このとき, 任意の $\delta > 0$ について,

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

が成り立つ, ただし $\Sigma^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f^2(X_i)$.

Remark:

- $\mathbb{E}[\Sigma^2]$ の upper bound から, より simple な bound が得られる.
- Chapter 4 の symmetrization techniques より, 以下が示される.

$$\mathbb{E}[\Sigma^2] \leq \sigma^2 + 2b\mathbb{E}[Z], \tag{3.84}$$

  ただし, $\sigma^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X)]$ ($\leftarrow$ これなに？).
- これを使うと, 以下を満たすある正定数 (c_0, c_1) があることが示される.

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + c_0\gamma\sqrt{t} + c_1 bt] \leq e^{-nt} \quad \text{for all } t > 0, \tag{3.85}$$

  ただし, $\gamma^2 = \sigma^2 + 2b\mathbb{E}[Z]$ (see Exercise 3.16).
- 今知られている best は, $c_0 = \sqrt{2}, c_1 = 1/3$.

Proof of Theorem 3.27

- 一般の b については rescaling できるので, $b = 1$ として良い.
- また, unrescaled version の $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$ について考える.
- Theorem 3.26 の証明中の (3.81) と (3.82) より,

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

- ここで,

$$\begin{aligned}\sum_{i=1}^n \sum_{f \in \mathcal{F}} \mathbb{I}[X \in \mathcal{A}(f)]\left(f\left(X_j\right) - f\left(Y_j\right)\right)^2 &\leq 2 \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2\left(X_i\right) + 2 \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2\left(Y_i\right) \\ &= 2\{\Gamma(X) + \Gamma(Y)\},\end{aligned}$$

  となる, ただし $\Gamma(X) := \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i)$ は Σ^2 の unrescaled version.

- この 2 式を合わせると,

$$\mathbb{H}(e^{\lambda Z}) \le 2\lambda^2\{\mathbb{E}[\Gamma e^{\lambda Z}] + \mathbb{E}[\Gamma]\mathbb{E}[e^{\lambda Z}]\}. \tag{3.87}$$

- $\mathbb{H}(e^{\lambda(Z+c)}) = e^{\lambda c}\mathbb{H}(e^{\lambda Z})$ なので (see Exercise 3.4), 同じ式が $\widetilde{Z} = Z - \mathbb{E}[Z]$ でも成り立つ.
- ここで次の lemma を使う（証明はあとで）.

Lemma 3.28 (Controlling the random variance)

任意の $\lambda > 0$ に対し,

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

- (3.88) と (3.87) for $\widetilde{Z}$ より,

$$\mathbb{H}[e^{\lambda \tilde{Z}}] \le \lambda^2\{2e\mathbb{E}[\Gamma]\varphi(\lambda) + 2\varphi'(\lambda)\} \quad \text{for all } \lambda > 0,$$

ただし $\varphi(\lambda) := \mathbb{E}[e^{\lambda \tilde{Z}}]$ の moment generating function.

- $\mathbb{E}[\widetilde{Z}] = 0$ なので, これは $b = 2,\ \sigma^2 = 2e\mathbb{E}[\Gamma]$ の Bernstein form の entropy bound (3.10) となっているので, Proposition 3.3 より

$$\mathbb{P}\left[\widetilde{Z} \geq \mathbb{E}[\widetilde{Z}] + \delta\right] \leq 2\exp\left(-\frac{\delta^2}{8e\mathbb{E}[\Gamma] + 4\delta}\right) \quad \text{for all } \delta \geq 0.$$

- Γ の定義に注意して $1/n$ で rescaling すれば, theorem の statement が示される. □

Proof of Lemma 3.28

- 関数 $g(t) = e^t$ を考えると, conjugate dual は $g^*(s) = s\log s - s$ (for $s > 0$).
  - 関数 $h : \mathbb{R}^k \to \mathbb{R}$ の conjugate dual は, $h^*(y) = \sup_x (\langle y, x\rangle - h(x))$.
- すると定義より, 任意の $s > 0,\ t \in \mathbb{R}$ について $st \leq s\log s - s + e^t$.
- $s = e^{\lambda\widetilde{Z}}, t = \Gamma - (e-1)\mathbb{E}[\Gamma]$ として期待値をとると,

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

- $\mathbb{E}[e^{\Gamma-(e-1)\mathbb{E}[\Gamma]}] \leq 1$ (by Exercise 3.15) と $\mathbb{E}[e^{\lambda\widetilde{Z}}] \geq e^{\lambda\mathbb{E}[\widetilde{Z}]} = 1$ (by Jensen's inequality) を入れると, (3.88) を得る？ (右辺 $\lambda\mathbb{E}[\widetilde{Z}e^{\lambda\widetilde{Z}}]$ の λ が一致しなくない？) □