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.

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

  - 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,

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

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,

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

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

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

- $\lambda_0 \downarrow 0$ とすると

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

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. $\square$

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,

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

  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

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

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

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

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，

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

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.

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

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

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

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,

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

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 $f : \mathbb{R}^n \to \mathbb{R}, k \in \{1, \dots, n\}, x_{\backslash k} = (x_i, i \neq k) \in \mathbb{R}^{n-1}$ に対して, 条件付きエントロピーを
$$\mathbb{H}(e^{\lambda f_k(X_k)} \mid x_{\backslash k}) := \mathbb{H}(e^{\lambda f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n)})$$
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 $X^{\backslash k} \in \mathbb{R}^{n-1}$ に対し, このエントロピー $\mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\backslash k})$ 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,
$$\mathbb{H}(e^{\lambda f(X_1 \dots, X_k)}) \leq \mathbb{E}\left[\sum_{k=1}^n \mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\backslash k})\right] \quad \text{for all } \lambda > 0. \tag{3.21}$$

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）．

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

- 各 $j = 1, \ldots, n$ に対し，$X_j = (X_j \ldots, X_n)$ とする．
- g は $\mathbb{E}[e^{g(X)}] \le 1$ 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．

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

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

- すると定義より，

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

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

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

- Lemma 3.8 より,

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

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

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

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

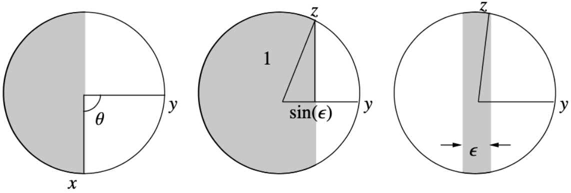
上の一様分布 $\mathbb{P}$ と, geodesic distance $\rho(x, y) := \arccos\langle x, y\rangle$ を考える.

- 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)

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



(a) (b) (c)

- 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, $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$. $\square$

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. $\square$