THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 7: Talagrand's inequality

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin

Convex Lipschitz functions of bounded random variables

Theorem

Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with Lipschitz constant L. Also consider n iid random variables $X_1, \ldots, X_n \in \{-1, 1\}$. We have for $t > 0$

$$P(|f(X) - M_f| \geq t) \leq 4 \exp\left(-\frac{t^2}{16L^2}\right),$$

where M_f is the median of f.

- Often the median can be replaced by the mean with a little give in the t.

From convex Lipschitz functions to sets

- Consider a metric space (X, d).
- Define $A = \{x : f(x) \leq M_f\}$
- Define $d(x, A) = \inf_{y \in A} d(x, y)$
- Define $A_t = \{x : d(x, A) \leq t\}$
- Since f is 1 Lipschitz (WLOG), $x \in A_t \Rightarrow f(x) \leq M_f + t$
- So $P(x \in A_t) \leq P(f(x) \leq M_f + t)$
- All we need is to upper bound $P(x \notin A_t)$
- Since f is convex, A is a convex set.

Talagrand's inequality: original statement

Theorem

Let $A \subset \mathbb{R}^n$ be a convex set. Then,

$$P(X \in A)P(X \notin A_t) \leq 4e^{-t^2/16}.$$

- This is basically saying that if A is convex and and $P(x \in A)$ is large then A_t takes up most of the space in the unit hypercube for $t \gg 1$.

Is convexity needed?

Example

Let $A := \{x \in \{-1,1\}^n : \sum_{i=1}^{n} 1(x_i = 1) \leq n/2\}$. Then $|f(x) - f(y)| \leq \sum_i |1(x_i = 1) - 1(y_i = 1)| \leq \sum_i |x_i - y_i| = \|x - y\|_1$. Then $P(x \in A)$ is large. But $P(x \notin A_t) \geq P(\sum_{i=1}^{n} 1(x_i = 1) \geq n/2 + t)$, which is large for $t \approx \sqrt{\log n}$, contrary to the result of Talagrand.

- Note that A is not convex.
- What if we define A as a subset of $\mathbb{R}^n$?

How about Azuma Hoeffding or McDiarmid?

- Let f is convex and one Lipschitz. Also, say $E[f(X)]$ was equal to the median.
- Note that in our setting, $|f(x) - f(y)| \leq 2$ when x, y differ in one coordinate.
- So using McDiarmid's inequality gives
  $P(|f(X) - E[f(X)]| \geq t) \leq 2\exp\left(-\frac{2t^2}{4n}\right)$,
  - i.e. it gives concentration when $t \gg \sqrt{n}$.
- But Talagrand's inequality gives
  $P(|f(X) - E[f(X)]| \geq t) \leq 4\exp\left(-\frac{t^2}{16}\right)$
  - i.e. it gives concentration when $t \gg 1$. ($X \gg 1$ implies X has factors logarithmic in n)

Going from median to expectation

- First note that $E[(f(X) - M_f)^2] \leq CL^2$ by using Talagrand's inequality. (How?)
- Now note that $\mathrm{var}(f(X)) \leq E[(f(X) - M_f)^2] \leq CL^2$
- Finally $E[|f(X) - E[f(X)]| \geq 2\sqrt{\mathrm{var}(f(X))}] \leq 1/4$.
- So we must have $M_f \in [E[f(X)] \pm cL]$
- So, $P(|f(X) - E[f(X)]| \geq cL + t) \leq 4e^{-t^2/16L^2}$

Operator norm of random matrices

Example

Consider a random matrix $M = [X_{ij}] \in [a, b]^{n \times m}$ where X_{ij} are independent random variables.

$$P(\|M\|_{op} \geq E[\|M\|_{op}] + c\sqrt{\log n}) = o(1)$$

- For $E[X_{ij}] = 0$ and $\mathrm{var}(X_{ij}) = \sigma^2$, it can be shown that $E[\|M\|_{op}] \leq 2\sigma\sqrt{n}$.
- $\|M\|_{op}$ is 1 Lipschitz and convex. (how?)

Operator norm of random matrices

Example

Consider a random matrix $M = [X_{ij}] \in [a, b]^{n \times m}$ where X_{ij} are independent random variables.

$$P(\|M\|_{op} \geq E[\|M\|_{op}] + c\sqrt{\log n}) = o(1)$$

- For $E[X_{ij}] = 0$ and $\text{var}(X_{ij}) = \sigma^2$, it can be shown that $E[\|M\|_{op}] \leq 2\sigma\sqrt{n}$.
- $\|M\|_{op}$ is 1 Lipschitz and convex. (how?)

Complexity

Example

Consider a iid sequence $X = \{X_i\}_{i=1}^n$. We will bound $f(X) := \sup_{a \in \mathcal{A}} a^T X$ where $\mathcal{A}$ is a compact subset of $\mathbb{R}^n$ such that $\mathcal{W} = \sup_{a \in \mathcal{A}} \|a\|_2 < \infty$.

- Why cant we just use Chernoff?
- First let us check if $f(X)$ is Lipschitz. Let a_* and a'_* be the maximizers of $f(X)$ and $f(X')$.
  - $$\begin{aligned} f(X) - f(X') = a_*^T X - a_*'^T X' &\le a_*^T (X - X') \\ &\le \sup_{a \in \mathcal{A}} a^T (X - X') \le \mathcal{W} \|X - X'\|_2 \end{aligned}$$
- How about convex? Consider the set $S_c = \{x : f(x) \le c\}$.
  - consider $x, y \in S_c$. Then

$$f(\lambda x + (1 - \lambda)y) \le f(\lambda x) + f((1 - \lambda)y) \le c$$

Complexity

Example

Consider a iid sequence $X = \{X_i\}_{i=1}^n$. We will bound $f(X) := \sup_{a \in \mathcal{A}} a^T X$ where $\mathcal{A}$ is a compact subset of $\mathbb{R}^n$ such that $\mathcal{W} = \sup_{a \in \mathcal{A}} \|a\|_2 < \infty$.

- If $X_i \sim N(0,1)$ using Gaussian+Lipschtz
  $P(|f(X) - E[f(X)]| \geq t) \leq 2e^{-\frac{t^2}{2\mathcal{W}^2}}$
- If X_i are bounded, then Talagrand gives us the same thing (modulo constants).
- How about McDiarmid?

Rademacher Complexity

Example

Consider a iid Rademacher sequence $X = \{X_i\}_{i=1}^n$. We will bound $f(X) := \sup\limits_{a \in \mathcal{A}} a^T X$ where $\mathcal{A}$ is a compact subset of $\mathbb{R}^n$ such that $\mathcal{W} = \sup\limits_{a \in \mathcal{A}} \|a\|_2 < \infty$.

- Consider X and X' differing in the k-th coordinate,
- $$f(X) - f(X') = a_*^T X - a_*'^T X' \le a_*^T (X - X')$$ $$\le \sup_{a \in \mathcal{A}} a_k (X(k) - X'(k)) \le \sup_{a \in \mathcal{A}} |a_k|$$
- So McDiarmid gives:

$$P(|f(X) - E[f(X)]| \ge t) \le 2\exp\left(-\frac{t^2}{2\sum_i \sup_{a \in \mathcal{A}} |a_i|^2}\right)$$