Statistics 210B Lecture 7 Notes

Daniel Raban

February 8, 2022

1 Concentration Inequalities for Convex Functions

1.1 Overview

Let $X_1, X_2, \ldots, X_n$ be independent, and let $Z = f(X_{1:n})$, were $f : \mathbb{R}^n \to \mathbb{R}$. We have been asking the question: "when is there a high probability bound for $|Z - \mathbb{E}[Z]|$.

Earlier, we had a solution in terms of the bounded differences inequality:

Theorem 1.1 (Bounded differences inequality)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $L_{1:n}$ bounded, i.e.*

$$|f(X_{1:n}) - f(x_{1:k-1}, x_k', x_{k+1:n})| \le L_k \qquad \forall x_{1:n}, x_k,$$

and $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$. Then for all $t \ge 0$,

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

This martingale concentration method let us control U-statistics and the supremum of an empirical process.

Last lecture, we had the Gaussian concentration inequality:

Theorem 1.2 (Gaussian concentration)**.** *Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$ and $f : \mathbb{R}^n \to \mathbb{R}$ such that f is L-**Lipschitz** in $\|\cdot\|_2$, i.e.*

$$|f(x) - f(y)| \le L\|x - y\|_2 \qquad \forall x, y \in \mathbb{R}^n.$$

Then

1. *$f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ is $\mathrm{sG}(L)$.*
2. $$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \ge t) \le 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

This let us discuss the singular values of a Gaussian random matrix and Gaussian complexity. To generalize this, the intuition is that we need 2 components:

1. We require the function f to not change much under perturbation of x.
2. We require the measure of X to behave sufficiently nicely.

1.2 Concentration of separately convex, Lipschitz functions

Theorem 1.3 (Concentration of separately convex, Lipschitz functions)**.** *Suppose that*

1. *f is L-Lipschitz and coordinatewise convex:*

$$\partial_k^2 f(x_{1:n}) \geq 0 \qquad \text{if } \partial_k^2 f \text{ exists}$$

2. *$(X_i)_{i\in[n]}$ independent with $X_i \in [a,b]$ a.s.*

Then

$$\mathbb{P}(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \geq t) \leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

This is a one-sided inequality; we don't have a lower tail bound here. To derive this result, we use the entropy method and the Herbst argument. This is covered in chapter 3.1 in Wainwright's textbook.

Remark 1.1. This has a stronger assumption than the bounded difference inequality, but it gives a stronger result.

1.3 Concentration of convex Lipschitz functions

Theorem 1.4 (Concentration of convex Lipschitz functions)**.** *Suppose that*

1. *f is L-Lipschitz and convex:*

$$\nabla^2 f(x) \succ 0 \qquad \text{if } \nabla^2 f \text{ exists}$$

2. *$(X_i)_{i\in[n]}$ independent with $X_i \in [a,b]$ a.s.*

Then $f(X_{1:n}) - \mathbb{E}[f]$ is $\mathrm{sG}(L(b-a))$, so

$$\mathbb{P}(f(|X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \geq t) \leq 2\exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Remark 1.2. Unlike the previous inequality, this one gives us an upper and lower tail bound. This has a stronger assumption than separate convexity, but it gives a stronger result.

To derive this result, Wainwright's book use a transportation approach. This is in chapter 3.6.

1.4 Applications

1.4.1 Rachemacher complexity

If $A \subseteq \mathbb{R}^n$, how do we measure its size? We previously defined the **Gaussian complexity**

$$\mathcal{G}(A) : - \mathbb{E}_{W \overset{\text{iid}}{\sim} N(0,1)} \left[\sup_{a \in A} \langle W, a \rangle \right].$$

Definition 1.1. The **Rademacher complexity** is

$$\mathcal{R}(A) : - \mathbb{E}_{\varepsilon_i \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left[\sup_{a \in A} \langle \varepsilon, a \rangle \right].$$

These notions are related, but they are useful in different situations.

Example 1.1. For all $1 < p < \infty$,

$$\mathcal{R}(B_p(r)) = \mathbb{E}_\varepsilon \left[\sup_{\|a\|_p \le r} \langle a, \varepsilon \rangle \right] = r \, \mathbb{E}_\varepsilon [\|\varepsilon\|_q] = r n^{1/q},$$
$$\mathcal{G}(B_p(r)) = r c_q n^{1/q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

If $p = 1$, then

$$\mathcal{R}(B_1(r)) = r \, \mathbb{E}_\varepsilon [\|\varepsilon\|_\infty] = r,$$
$$\mathcal{G}(B_1(r)) = \mathbb{E}_W \left[\sup_{i \in [n]} |W_i| \right] \approx r \sqrt{2 \log n} + O(1).$$

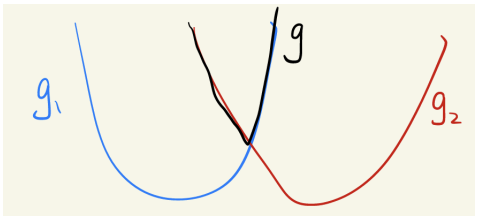
Here is an exercise from Wainwright's book.

Proposition 1.1. *There exist universal constants c, C such that for all $A \subseteq \mathbb{R}^n$,*

$$c \mathcal{R}(A) \le \mathcal{G}(A) \le C \sqrt{\log n} \mathcal{R}(A).$$

If we want to talk about concentration of Rachemacher random variables, we can use the above concentration inequalities. Define $f(\varepsilon) = \sup_{a \in A} \langle \varepsilon, a \rangle$. Then $f(\varepsilon)$ is $D(A)$-Lipschitz, where $D(A) = \sup_{a \in A} \|a\|_2$.

Lemma 1.1. *Let g_1, g_2 be convex functions. Then $g(x) = \max\{g_1(x), g_2(x)\}$ is convex.*



This implies that $f(\varepsilon)$ is convex. So we get that $f(\varepsilon) - \mathbb{E}[f(\varepsilon)]$ is $\mathrm{sG}(2D(A))$. This tells us that

$$f(\varepsilon) \approx \mathcal{R}(A) + O(D(A)).$$

1.4.2 Operator norm

Let

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix} \in \mathbb{R}^{n \times d},$$

where $X_{i,j} \in [-1, 1]$ a.s. Then, if we let $f(x) = \|x\|_{\mathrm{op}}$, then f is 1-Lipschitz and convex. So $f(X) = \mathbb{E}[f(X)]$ is $\mathrm{sG}(2)$, which tells us that

$$\|X\|_{\mathrm{op}} \simeq \mathbb{E}[\|X\|_{\mathrm{op}}] + O(1).$$

1.5 Proof techniques: the Herbst argument and transportation

Here is how we can prove the above concentration inequalities.

1. Entropy method and the Herbst argument

   Definition 1.2. The **Herbst** argument is that a sufficient condition for X to be $\mathrm{sG}(\sigma)$ is to show that

   $$\mathbb{H}(e^{\lambda X}) \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}],$$

   where $\mathbb{H}$ is the entropy.

   Why do we want to look at $\mathbb{H}(e^{\lambda X})$? This is because it has a good **tensorization property** when X_i are independent:

   $$\mathbb{H}(e^{\lambda f(X_{1:n})}) \leq \mathbb{E}\left[\sum_{i=1}^{n} \underbrace{\mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\setminus k})}_{\substack{\text{easy to handle when} \\ f_k \text{ Lip., } X_k \text{ bdd.}}}\right]$$

   For this, see chapter 3.1 of Wainwright's textbook or chapter 3 of van Handel's textbook.

2. Transportation approach:

   Lemma 1.2 (Bobkov-Gotze)**.** *Given a measure $\mu \in \mathcal{P}(\mathbb{R}^n)$,*

   $X \sim \mu, \forall f$ *1-Lipschitz,* $f(X)$ *is* $\mathrm{sG}(\sigma) \iff W_1(\nu, \mu) \leq \sqrt{2\sigma^2 \,\mathrm{KL}(\nu \mid\mid \mu)} \forall \nu \in \mathcal{P}(\mathbb{R}^n)$,

   where W_1 is the transportation distance and KL *is the relative entropy.*

   This property on the right also tensorizes in some way. For more on this, see chapter 3.3 in Wainwright's book or chapter 4 in van Handel's book.

1.6 Concentration of Lipschitz functions of log-concave random variables

Definition 1.3. A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is **r-strongly convex** if $\nabla^2 \psi(x) \succeq r I_n$, if this exists.

Definition 1.4. If $\mu \in \mathcal{P}(\mathbb{R}^n)$, we say that μ is **r-strongly log-concave** if $\mu(x) = \exp(-\psi(x))$, where ψ is r-strongly convex.

Example 1.2. Let $p_\theta(x) = \frac{1}{Z_(\theta)} \exp(\langle \theta, T(x) \rangle)$ be an exponential family. Suppose we have the prior

$$\pi(\theta) \sim N(0, I_n)$$

and the posterior

$$p(\theta \mid x) \propto p_\theta(x)\pi(\theta) = \frac{1}{\widetilde{Z}(x)} \exp(\langle \theta, T(x) \rangle - \log Z(\theta) - \tfrac{1}{2}\|\theta\|_2^2)$$

So we may let

$$\psi(\theta) = -\langle \theta, T(x) \rangle + \log Z(\theta) + \frac{1}{2}\|\theta_2^2 + \log \widetilde{Z}(x).$$

Note that

$$(\log Z(\theta))'' = \mathrm{Cov}_\theta(T(X), T(X)) \geq 0.$$

Theorem 1.5 (Concentration of Lipschitz functions of log-concave random variables)**.** *Suppose that*

1. *f is L-Lipschitz,*
2. *$X \sim \mu \in \mathcal{P}(\mathbb{R}^n)$, where μ is r-log-concave.*

Then $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ is $\mathrm{sG}(L/\sqrt{r})$.

1.7 Proof technique: the isoperimetric inequality

The isoperimetric inequality is a geometric property in $\mathbb{R}^n$ with Lebesgue measure. If $A \subseteq \mathbb{R}^n$ has fixed volume and we want to minimize the perimeter, then the solution is when A is a ball. This generalizes to other measures:

$X \sim \mu =$	$N(0, I_n)$	$S^{n-1}(\sqrt{n})$	$\mathrm{Unif}(\{\pm 1\}^n)$
	Half space	Spherical cap	Hamming ball

The isoperimetric inequality implies that $f(X)$ concentrates when f is Lipschitz. Suppose that $\mathbb{P}(A) = 1/2$, and take μ to be, for example, the Gaussian measure. Then define $A_\varepsilon = \{a : \exists b \in A \text{ s.t. } \|a - b\| \leq \varepsilon\}$. In this sitaution, perimeter is defined as

$$\lim_{\varepsilon \to 0} \frac{f(A_\varepsilon) - f(A)}{\varepsilon}.$$

Then, using the fact that $\mathbb{P}(\{x \in \mathbb{R}^n : x_1 \leq 0\}) = 1/2$. the isoperimetric inequality tells us that for all small enough ε,

$$\mathbb{P}(A_\varepsilon) \geq \mathbb{P}(\{x \in \mathbb{R}^n : x_1 \leq \varepsilon\}) = 1 - \Phi(\varepsilon) \geq 1 - \exp\left(-\frac{t^2}{2}\right).$$

For more on this, see chapter 3.2 of Wainwright's book and also see Chapter 7 of the book by Lugosi, Massart, and Boucheron.