Statistics 210B Lecture 6 Notes

Daniel Raban

February 3, 2022

1 Gaussian Concentration

1.1 Freedman's inequality

Last time, we generalized the Hoeffding and Bernstein inequalities for independent random variables to Azuma-Hoeffding and "Azuma Bernstein inequalities for martingales."

Our "Azuma-Bernstein" inequality says that if $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2}$, then

$$\left| \frac{1}{n} \sum_{k=1}^n D_k \right| \leq \max \left\{ \sqrt{\frac{\frac{2}{n} \sum_{k=1}^n \nu_k^2}{n} \log\left(\frac{2}{\delta}\right)}, \frac{2\alpha_* \log\left(\frac{2}{\delta}\right)}{n} \right\} \qquad \text{with probability } 1 - \delta.$$

However, sometimes ν_k^2 is not deterministic and $\nu_k^2 = \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}]$ instead is $\mathcal{F}_{k-1}$ measurable.

Theorem 1.1 (Freedman's inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence such that*

1. $\mathbb{E}[D_k \mid \mathcal{F}_{k=1}] = 0$.
2. $D_k \leq b$ *a.s.*

Then for all $\lambda \in (0, 1/b)$ and $\delta \in (0,1)$,

$$\mathbb{P}\left(\sum_{t=1}^T X_t \leq \lambda \sum_{t=1}^T \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}] + \frac{\log(1/\delta)}{\lambda} \right) \geq 1 - \delta.$$

This is useful in bandit and reinforcement learning research.[1]

[1] For example, see Theorem 1 in Beygelzimer, Langford, et. al. 2010.

1.2 Maximal Azuma-Hoeffding inequality

Recall Doob's maximal inequality for sub-martingales.

Lemma 1.1 (Doob's maximal inequality)**.** *If $\{X_s\}_{s\geq 0}$ is a sub-martingale, i.e.*

$$X_s \leq \mathbb{E}[X_t \mid \mathcal{F}_s] \qquad \forall s < t,$$

then for all $u > 0$,

$$\mathbb{P}\left(\sup_{0\leq t\leq T} X_t \geq u\right) \leq \frac{\mathbb{E}[\max\{X_T, 0\}]}{u}.$$

This gives rise to a maximal version of the Azuma-Hoeffding inequality:

Theorem 1.2 (Maximal Azuma-Hoeffding inequality)**.** *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence, and suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s. Then*

$$\mathbb{P}\left(\sup_{0\leq k\leq n} \sum_{s=1}^{k} D_k \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right).$$

If we used the usual Azuma-Hoeffding inequality instead, we would need to use a union bound, which would give a factor of n in the bound. We can write this conclusion as

$$\sup_{0\leq k\leq n} \sum_{s=1}^{k} D_k \leq \sqrt{\frac{C\log(1/\delta)}{n}}.$$

If we have the extra factor of n, we get an n/δ instead, which can sometimes be not a big deal for our bound since we are taking a log.

1.3 Gaussian concentration

Lemma 1.2. *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.*

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

Then

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

Remark 1.1. We need f to be Lipschitz as a whole function! It's not just sufficient for the function to be coordinate-wise Lipschitz.

Remark 1.2. If the X_is are non-Gaussian, this doesn't always hold with only Lipschitzness.

There are many different proofs of this lemma, but none are very simple.

Proof 1: Gaussian interpolation method

Proof 2: Gaussian isoperimetric inequality

Proof 3: Gaussian log-Sobolev inequality + Herbst argument

Today, we will present a proof using the Gaussian interpolation method, which is useful in research. However, this is a technique where you need to develop some intuition to understand it.

1.4 Examples of Gaussian concentration

Example 1.1 (Order statistics)**.** Let $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} N(0,1)$. The order statistics are the random variables arranged in increasing order: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Let $f_k(X_{1:n}) = X_{(k)}$. This is Lipschitz:

$$\begin{aligned} |f_k(X_{1:n}) - f_k(Y_{1:n})| &= |X_{(k)} - Y_{(k)}| \\ &\leq \sqrt{\sum_{k=1}^{n} |X_{(k)} - Y_{(k)}|^2} \end{aligned}$$

The **rearrangement inequality** says that if you sort the terms, the distance is greater than the distance of with unsorted terms.

$$\begin{aligned} &\leq \sqrt{\sum_{k=1}^{n} |X_k - Y_k|^2} \\ &= \|X - Y\|_2. \end{aligned}$$

This means that $L = 1$, so $X_{(k)} - \mathbb{E}[X_{(k)}]$ is sG(1). Therefore,

$$|X_{(k)} - \mathbb{E}[X_{(k)}]| \leq \sqrt{\log(2/\delta)} \qquad \text{with probability } 1 - \delta.$$

If we apply this to $k = n$, we get

$$\left| \max_{i\in[n]} X_i - \underbrace{\mathbb{E}\left[\max_{i\in[n]} X_i\right]}_{\sqrt{2\log n}} \right| = O_p(1).$$

Example 1.2 (Singular value of Gaussian random matrices)**.** Let

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix} \in \mathbb{R}^{n\times d}, \qquad X_{i,j} \overset{\text{iid}}{\sim} N(0,1).$$

Let $f_k(X) = \sigma_k(X)$ be the k-th largest singular value of X. For example, $f_1(X) = \|X\|_{\text{op}}$. It can be shown that $\mathbb{E}[\|X\|_{\text{op}}] \approx \sqrt{n} + \sqrt{d}$. We can show that f_k is Lipschitz; what is the norm we want to be using for a matrix? Define the vectorized version of the matrix as $\text{vec}(X) := (X_{1,1}, X_{1,2}, \dots, X_{1,d}, X_{2,1}, \dots, X_{2,d}, \dots, X_{n,d})$. Then

$$\|\text{vec}(X) - \text{vec}(Y)\|_2 = \|X - Y\|_F = \sqrt{\sum_{i,j} (X_{i,j} - Y_{i,j})^2},$$

where $\|\cdot\|_F$ is the **Frobenius norm**. Now we have

$$|f_k(X) - f_k(Y)| \le |\sigma_k(X) - \sigma_k(Y)|$$

Weyl's inequality, a deterministic linear algebra result, says that

$$\begin{aligned} &\le \|X - Y\|_{\text{op}} \\ &\le \|X - Y\|_F, \end{aligned}$$

so $L = 1$. Weyl's inequality can be proven by using the variational representation of singular values.

This calculation tells us that $f_k(X) - \mathbb{E}[f_k(X)]$ is sG(1), so

$$f_k(X) - \mathbb{E}[f_k(X)] \le \sqrt{\log(2/\delta)} \qquad \text{with probability } 1 - \delta.$$

Applying this to $k = 1$ gives

$$|\|X\|_{\text{op}} - \underbrace{\mathbb{E}[\|X\|_{\text{op}}]}_{\sqrt{n}+\sqrt{d}}| = O(1).$$

1.5 Gaussian complexity

Gaussian complexity is a very important notion in compressed sensing. Suppose we have a set $A \subseteq \mathbb{R}^n$. How do we measure its "size"? A reasonable size function S should at least satisfy $S(A) \le S(B)$ if $A \subseteq B$. Here are some reasonable size functions:

1. Euclidean width: $D(A) = \max_{a\in A} \|a\|_2$.

2. Dimension: A line has dimension 1, and a plane has dimension 2.

Definition 1.1. Given a set A, let $W = (W_1, \dots, W_n)^\top \in \mathbb{R}^n$ with $W_i \overset{\text{iid}}{\sim} N(0,1)$. The **Gaussian complexity** or "**statistical dimension**" of A is

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

Note that if we don't take the supreumum in the expectation, the quantity would be 0. This quantity is always nonnegative.

Example 1.3. Let $B_p(r) = \{x \in |R^n : \|x\|_p \le r\}$. Then

$$\mathcal{G}(B_p(r)) = \mathbb{E}\left[\sup_{\|x\|_p\le r}\langle x, W\rangle\right]$$

If q is the conjugate exponent of p, so $\frac{1}{p} + \frac{1}{q} = 1$, this is the variational representation of the $\|\cdot\|_q$ norm:

$$\begin{aligned} &r\,\mathbb{E}[\|W\|_q] \\ &\approx rn^{1/q}. \end{aligned}$$

Note that if $p_1 \le p_2$, then $q_1 \ge q_2$, so $\mathcal{G}(B_{p_1}(r)) \le \mathcal{G}(B_{p_2}(r))$.

We want to show that $f(W) := \sup_{a\in A}\langle a, W\rangle$ concentrates. Fix $w, w' \in \mathbb{R}^n$. Then

$$f(w) - f(w') = \sup_{a\in A}\langle a, w\rangle - \sup_{a\in A}\langle a, w'\rangle$$

Denote $a^* = \arg\max_a \langle a, w\rangle$

$$\begin{aligned} &= \langle a^*, w\rangle - \sup_{a\in A}\langle a, w'\rangle \\ &= \inf_{a\in A}\langle a^*, w\rangle - \langle a, w'\rangle \\ &\le \langle a^* w - w'\rangle \\ &\le \|a_*\|\|w - w'\|_2 \\ &\le D(A)\|w - w'\|_2. \end{aligned}$$

The other side can be proven similarly, so f is $D(A)$-Lipschitz. Concentration says that $f(W)$ is sG($D(A)$).

Example 1.4. If we let $A = B_2(R)$, then

$$\mathbb{E}[f(W)] = \mathcal{G}(B_2(r)) = r\sqrt{n},$$

since $D(A) = r$.

1.6 Proof of the Gaussian concentration inequality (interpolation method)

Lemma 1.3. *For all convex $\phi : \mathbb{R} \to \mathbb{R}$ and differentiable $f : \mathbb{R}^n \to \mathbb{R}$,*

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(Y)])] \leq \mathbb{E}[\phi(\tfrac{\pi}{2}\langle \nabla f(X), Y\rangle],$$

where $X, Y \overset{\text{iid}}{\sim} N(0, I_n)$.

First, assume this lemma holds, and prove Gaussian concentration:

Proof. Take $\phi = \exp(\lambda \cdot)$. THen

$$\mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(Y)]))] \leq \mathbb{E}[\exp(\lambda \tfrac{\pi}{2}\langle \nabla f(X), Y\rangle)]$$

Observe that $\frac{\pi}{2}\langle \nabla f(X), Y\rangle$ is $N(0, \frac{\pi^2}{4}\|\nabla f(X)\|_2^2$ given X.

$$\begin{aligned} &= \mathbb{E}_X[\exp(\tfrac{\lambda^2}{2}\frac{\pi^2}{4}\|\nabla f(X)\|_2^2)] \\ &\leq \exp\left(\frac{\lambda^2}{2}\frac{\pi^2}{4}L^2\right). \end{aligned}$$

This says that $f(X) - \mathbb{E}[f(X)]$ is $\mathrm{sG}(\frac{\pi}{2}L)$. □

The above proof gives a worse constant, but the constant can be improved with different methods. Here is the proof of the lemma:

Proof. First, use conditioning and Jensen's inequality to say that.

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(Y)])] \leq \mathbb{E}_{X,Y}[\phi(f(X) - f(Y))]$$

The idea is to use the integral representation of the Taylor expansion to interpolate between X and Y. Observe that if $Z(\theta) = X\cos\theta + Y\sin\theta$, then for every θ, $Z(\theta) \overset{d}{=} X \overset{d}{=} Y$ and $Z'(\theta) \overset{d}{=} X \overset{d}{=} Y$. Another important property is that $Z(\theta) \perp Z'(\theta)$; this is because $Z(\theta), Z'(\theta)$ are Gaussians with 0 covariance. Now

$$f(X) - f(Y) = \int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta)\rangle \, d\theta,$$

so we can write

$$\mathbb{E}[\phi(f(X) - f(Y))] = \mathbb{E}\left[\phi\left(\int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta)\rangle \, d\theta\right)\right]$$

Using Jensen's inequality, $\phi(\int \cdot \, d\theta) \leq \int \phi(\cdots)\, d\theta$ when $\int \cdots \, d\theta = 1$.

$$\begin{aligned} &\leq \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}[\phi(\tfrac{\pi}{2}\langle \nabla f(Z(\theta)), Z'(\theta)\rangle)]\, d\theta \\ &= \mathbb{E}[\phi(\tfrac{\pi}{2}\langle \nabla f(X), Y\rangle)]. \end{aligned}$$

□

This proof is very delicate, and the construction looks ad hoc, but it is actually very useful in a variety of situations.

1.7 Other methods for establishing concentration

1. Matrix concentration: If $(X_i)_{i\in[n]} \subseteq \mathbb{R}^{m\times d}$ with $X_i \overset{\text{iid}}{\sim} X$, can we find a bound for

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \right\|_{\text{op}} ?$$

   The answer is yes; there is a matrix Bernstein inequality, Rudelson's inequality, and a matrix Freedman inequality. These involve the matrix MGF and Lieb's inequality. For more, see *An Introduction to Matrix Inequalities*, Tropp 2015, and *Introduction to Non-asymptotic analysis of random matrices*, Vershynin 2010.

2. Entropy method and the Herbst argument

   Definition 1.2. The **Herbst** argument is that a sufficient condition for X to be sG(σ) 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:

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

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

3. Isoperimetric inequality: This 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:

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

   The isoperimetric inequality implies that $f(X)$ concentrates when f is Lipschitz. For this, see chapter 3.2 of Wainwright's book and also see Chapter 7 of the book by Lugosi, Massart, and Boucheron.

4. Transportation approach:

Lemma 1.4 (Bobkov-Gotze)**.** *Given a measure* $\mu \in \mathcal{P}(\mathbb{R}^n)$*,*

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

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.