Statistics 210B Lecture 5 Notes

Daniel Raban

February 1, 2022

1 Martingale Concentration Inequalities

1.1 Motivation and overview

Our goal is to get a tail bound for $X_1 + \cdots + X_n$, where the X_i are independent. Here is our solution so far:

(a) Chernoff inequality bounded by MGF.

(b) Bound MGF using sub-Gaussian and sub-exponential properties.

(c) Many commonly used random variables are sub-Gaussian or sub-exponential.

What about more complicated structure?

1. Sometimes, we want to show concentration of $S_n = f(X_1, \dots, X_n) =: f(X_{1:n})$.
2. Sometimes, we want to show concentration of $S_n = \sum_{t=1}^T X_t$, where $\{X_t\}_{t \geq 1}$ is correlated. We can deal with this if it is a Martingale difference sequence.

This lecture, we will take the approach of a Martingale concentration inequaltiy. We will use Markov's inequality on $e^{\lambda S_n}$ along with a conditional MGF bound and optimizing over λ. We will see

(a) Doob's Martingale representation

(b) Azuma-Hoeffding, Azuma- Bernstein, and bounded difference inequalities

(c) Applications

(d) Variants: Freedman's inequality and Doob's maximal inequality

Example 1.1. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}([a,b])$. We want to estimate $\theta = \mathbb{E}_{X,X' \overset{\text{iid}}{\sim} P_X}[g(X,X')]$, where we assume that $g : \mathbb{R}^2 \to \mathbb{R}$ is symmetric (such as $g(x,x') = |x - x'|$ or $g(x,x') = \frac{1}{2}(x-x')^2$. In the latter case, $\theta = \operatorname{Var}(X)$.

Hoeffding introduced U**-statistics** for estimating these parameters θ:

$$U(X_{1:n}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} g(X_i, X_j).$$

If we let

$$\widehat{\mathbb{P}}_{X,X'} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \delta_{(X_i, X_j)}$$

be the empirical distribution, then $U(X_{1:n}) = \widehat{E}_{(X,X')}[g(X,X')]$. The U statistic is an unbiased estimator of θ because

$$\mathbb{E}[U(X_{1:n})] = \mathbb{E}[g(X_i, X_j)] = \theta.$$

This has the smallest variance among all unbiased estimators.

Today, we will show the concentration bound

$$\mathbb{P}(|U - \theta| \ge t) \le 2 \exp\left(-\frac{nt^2}{2\|g\|_\infty}\right).$$

This is significant because U is not a sum of independent random variables, so our previous technology does not work here.

1.2 Doob's martingale representation of $f(X_1, \dots, X_n)$

Now return to the setting where we are dealing with $f(X_1, \dots, X_n)$, where the X_i are independent. Define

$$Y_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] \qquad k \ge 0.$$

We can think of conditioning on $X_{1:k}$ as conditioning on the σ-algebra $\mathcal{F}_k = \sigma(X_{1:k})$

Example 1.2. Here is the example to keep in mind: Let $f(X_{1:n}) = X_1 + \cdots + X_n$ with independent X_i. Then

$$Y_k = X_1 + \cdots + X_k + \mathbb{E}[X_{k+1}] + \cdots + \mathbb{E}[X_n].$$

Further define the difference

$$D_k = Y_k - Y_{k-1}.$$

In the previous example, $D_k = X_k - \mathbb{E}[X_k]$. We can in general write

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^n (Y_k - Y_{k-1}) = \sum_{k=1}^n D_k.$$

We call $\{Y_k\}$ a **martingale sequence** and $\{D_k\}$ a **martingale difference sequence**.

Let us recall what a martingale is.

Definition 1.1. A **filtration** is an increasing nested sequence of σ-algebras

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \cdots .$$

Often, we take $\mathcal{F}_k = \sigma(X_{1:k})$. If the filtration is not defined properly, the result you get may not be true.

Definition 1.2. If we have $\{Y_k\}_{k=1}^{\infty}$, where Y_k is $\mathcal{F}_k$-measurable, then we way that $\{Y_k\}$ is **$\{\mathcal{F}_k\}$-adapted**.

Definition 1.3. $\{(Y_k, \mathcal{F}_k)\}_{k\geq 1}$ is a **martingale sequence** if

1. $\{Y_k\}$ is adapted to $\{\mathcal{F}_k\}$.
2. $\mathbb{E}[|Y_k|] < \infty$,
3. $\mathbb{E}[Y_k \mid \mathcal{F}_{k-1}] = Y_{k-1}$.

Martingales are often used to model gambling problems where your strategy can depend on the outcomes of the past. If you don't have a martingale, you can sometimes subtract the mean to get one.

Definition 1.4. $\{D_k\}_{k\geq 1}$ is a **martingale difference sequence** if $\{\sum_{k=1}^n D_k\}_{n\geq 1}$ is a martingale with respect to $\{\mathcal{F}_k\}_{k\geq 1}$.

Example 1.3. Let $\{X_i\}_{i\geq 1} \overset{\text{iid}}{\sim} P_X$, where $\mathbb{E}[|X|] < \infty$. Denote $\mu = \mathbb{E}_X[X]$ and $S_k = \sum_{s=1}^k X_s$. Then $\{(X_k - k\mu, \sigma(X_{1:k}))\}_{k\geq 1}$ is a martingale.

Proof. We only need to check the third property:

$$\begin{aligned}\mathbb{E}[S_k - k\mu \mid X_{1:k-1}] &= S_{k-1} - (k-1)\mu \\ &= Y_{k-1}.\end{aligned}$$

□

Example 1.4 (Doob's martingale)**.** Let $\{X_i\}_{i\geq 1}$ be independent[1] and $\mathbb{E}[|f(X_1, \ldots, X_n)|] < \infty$. Then $\{(Y_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}], \sigma(X_{1:k}))\}_{k\geq 1}$ is a martingale sequence.

Proof. Again, we only check the third property:

$$\begin{aligned}\mathbb{E}[Y_{k+1} \mid \sigma(X_{1:k})] &= \mathbb{E}[\mathbb{E}[f(X_{1:n}) \mid X_{1:n+1}]X_{1:k}] \\ &= \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] \\ &= Y_k\end{aligned}$$

The second equality is by the tower property of conditional expectation. □

[1] In class, we had this assumption, but I don't think it is actually needed.

1.3 Martingale concentration

Most inequalities for an iid sum have a martingale version. Here is a martingale version of Bernstein's inequality.[2]

Theorem 1.1. *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence. If*

$$\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2} \qquad a.s.\ \forall \lambda \leq \frac{1}{\alpha_k},$$

then

1. $\sum_{k=1}^n D_k$ *is* $\mathrm{sE}(\sqrt{\sum_{k=1}^n \nu_+ k^2}, \max_{k \leq n} \alpha_k)$.
2. $$\mathbb{P}\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t^2}{2\sum_{k=1}^n \nu_k^2}, \frac{t}{2\alpha_*}\right\}\right)$$

This condition is that a random variable given by the MGF is bounded. We will see later how to check this condition.

Proof. We can start with the Chernoff bound

$$\mathbb{P}\left(\sum_{k=1}^n D_k\right) \geq t \leq \inf_\lambda \frac{\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}]}{e^{\lambda t}}.$$

Then we can bound the moment generating function by using the tower property of conditional expectation

$$\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}] = \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}\, \mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]$$

Using $\lambda \leq \frac{1}{\alpha_n}$,

$$\leq \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k} e^{\lambda^2 \nu_k^2/2}]$$
$$= \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}] e^{\lambda^2 \nu_k^2/2}$$

Iterating this argument, we get

$$\leq e^{\lambda^2 (\sum_{k=1}^n \nu_k^2)/2}$$

for all $\lambda \leq \frac{1}{\max_{k \leq n} \alpha_k}$. □

Remark 1.1. In this theorem, the ν_k are deterministic. In the case where the ν_k are $\mathcal{F}_{k-1}$-measurable, we will get a related but different bound.

Here is a corollary which is sometimes easier to use than the previous theorem.

[2]This inequality does not have a formal name, but you may call it an Azuma-Bernstein inequality.

Corollary 1.1 (Azuma-Hoeffding inequality)**.** *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence. Suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s., where b_k, a_k are $\mathcal{F}_{k-1}$-measurable and $|b_k - a_k| \leq L_k$. Then*

1. $\sum_{k=1}^n D_k$ *is* $\mathrm{sG}(\sqrt{\sum_{k=1}^n L_k^2}/2)$.
2. $$\mathbb{P}\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right).$$

Proof. We have $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 (b_k - a_k)^2/8}$. Use the same proof as before. $\square$

Now specialize to Doob's martingale

$$D_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}].$$

Definition 1.5. $f(x_1, \dots, x_n)$ is a **bounded difference function** if for all $k \in [n], x_{1:n}, x_k'$,

$$|f(x_{1:k-1}, x_k, x_{k+1:n}) - f(x_{1:k-1}, x_k', x_{k+1:n})| \leq L_k.$$

This is a condition on how much the function changes if we change 1 coordinate. Here is a corollary of the Azuma-Hoeffding inequality

Corollary 1.2. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $L_{1:n}$ bounded and $X_{1:n}$ has independent components. Then for all $t \geq 0$,*

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

Proof. This is Azuma-Hoeffding with $\sum_{k=1}^n D_k = f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$. Here, there exist $A_k \leq D_k \leq B_k$, where $|B_k - A_k| \leq L_k$ because we can let

$$B_k = \sup_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}],$$
$$A_k = \inf_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}].$$

$\square$

1.4 Applications

Example 1.5 (U-statistics)**.** Here is how we can get a cncentration inequality for U-statistics: Recall that

$$U(X_{1:n}) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} |X_i - X_j|, \qquad X_i \sim P_X \in \mathcal{P}([-b, b]).$$

Then

$$
\begin{aligned}
|U(X_{1:k-1}, X_k, X_{k+1:n}) - U(X_{1:k-1}, X_k', X_{k+1:n}) &= \frac{1}{\binom{n}{2}} \left| \sum_{s \neq k} |X_s - X_k| - |X_s - X_k'| \right| \\
&\leq \frac{1}{\binom{n}{2}} \sum_{s \neq k} |X_k - X_k'| \\
&\leq \frac{2}{n(n-1)} \cdot (n-1) \cdot 2b \\
&\leq \frac{4b}{n}.
\end{aligned}
$$

So U is $(\frac{4b}{n}, \frac{4b}{n}, \dots, \frac{4b}{n})$-bounded difference. This gives the tail bound

$$
\mathbb{P}(|U(X_{1:n}) - \theta| \geq t) \leq 2 \exp\left(\frac{2t^2}{n \frac{16}{n^2}} \right) = 2 \exp\left(-\frac{nt^2}{16} \right).
$$

That is,

$$
|U(X_{1:n} - \theta| \lesssim b \sqrt{\frac{\log(2/\delta)}{n}} \qquad \text{with probability } 1 - \delta.
$$

Example 1.6 (Supremum of empirical process)**.** Suppose we have samples $(Z_i)_{i \in [n]} \overset{\text{iid}}{\sim} P_Z$, where $Z_i = (X_i, Y_i)$. We can define the **loss function** $\ell : Z \times \Theta \to [0,1]$ and the **empirical risk**

$$
\widehat{R}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(Z_i; \theta).
$$

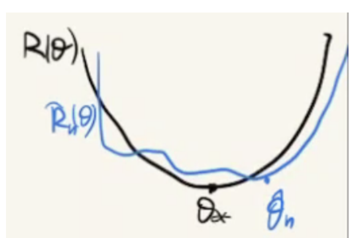
Correspondingly, we have the **population risk**

$$
R(\theta) = \mathbb{E}[\widehat{R}_n \mid \theta] = \mathbb{E}[\ell(Z; \theta)]
$$

In statistical learning theory, we are often concerned with the **excess risk**

$$
\mathcal{E}[Z_{1:n}] := \sup_{\theta \in \Theta} R(\theta - \widehat{R}_n(\theta).
$$

We can use an **empirical risk minimizer** $\widehat{\theta}_n$, and we want to upper bound $R(\widehat{\theta}_n) \leq \widehat{R}_n(\widehat{\theta}_n) + \mathcal{E}(Z_{1:n})$.



We claim that $\mathcal{E}(Z_{1:n})$ is $(1/n, \ldots, 1/n)$-bounded difference. Then

$$|\mathcal{E}(Z_{1:n}) - \mathbb{E}[\mathcal{E}(Z_{1:n})]| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \qquad \text{with probability } 1-\delta.$$

Proof. Fix $Z_{1:n}$, and let $\theta_* = \arg\max_{\theta \in \Theta}(R(\theta) - \widehat{R}_n(\theta))$. Then $\mathcal{E}(Z_{1:n}) = R(\theta_*) - \widehat{R}_n(\theta_*)$. We want to look at

$$\begin{aligned}
|\mathcal{E}(Z_{1:n}) - \mathcal{E}(Z_{1:k-1}, Z_k', Z_{k+1:n})| &= \frac{1}{n}\sum_{i=1}^n (\ell(Z_i;\theta_*) - \mathbb{E}[\ell(Z_i;\theta_*)]) \\
&\qquad - \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i\neq k}(\ell(Z_i;\theta) - \mathbb{E}[\ell(Z_i;\theta)) \\
&\qquad - \frac{1}{n}(\ell(Z_k';\theta) - \mathbb{E}[\ell(Z_k';\theta)]) \\
&\leq \frac{1}{n}\sum_{i=1}^n (\ell(Z_i;\theta_*) - \mathbb{E}[\ell(Z_i;\theta_*)]) \\
&\qquad - \frac{1}{n}\sum_{i\neq k}(\ell(Z_i;\theta_*) - \mathbb{E}[\ell(Z_i;\theta_*)) \\
&\qquad - \frac{1}{n}(\ell(Z_k';\theta_*) - \mathbb{E}[\ell(Z_k';\theta_*)]) \\
&= \frac{1}{n}(\ell(Z_k;\theta_*) - \ell(Z_k';\theta_*)) \\
&\leq \frac{1}{n}.
\end{aligned}$$

□

Remark 1.2. This doesn't say anything about

$$\mathbb{E}\left[\sup_\theta \widehat{R}_n(\theta) - R(\theta)\right].$$

1.5 Freedman's inequality

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 instead is $\mathcal{F}_{k-1}$ measurable.

Theorem 1.2 (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. $D_k \le b$ *a.s.*

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

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

This is useful in bandit and reinforcement learning research.[3]

1.6 Maximal Azuma-Hoeffding inequality

Recall Doob's maximal inequality for sub-martingales.

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

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

then for all $u > 0$,

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

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

Theorem 1.3 (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*

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

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