Statistics 210B Lecture 2 Notes

Daniel Raban

January 20, 2022

1 Basic Concentration Inequalities

1.1 Concentration inequalities for sample averages

Suppose we have a random variable $X \sim \mathbb{P}_X$, sampled from the distribution $\mathbb{P}_X$. Let $\mu = \mathbb{E}_{X \sim \mathbb{P}_X}[X]$ be its expectation. In general, $|x - \mu|$ could be very large. However, in many scenarios (especially when X takes a special form), $|x - \mu|$ is very small with high probability.

Example 1.1. Let $X = \frac{1}{n}\sum_{i=1}^n Z_i$, where $Z_i \overset{\text{iid}}{\sim} \mathbb{P}_Z$ with $\mathbb{P}_Z \in \mathcal{P}([0,1])$ (supported in $[0,1]$). Then $\mathbb{E}[X] = \mathbb{E}[Z_i] =: \mu$. We will show in this lecture that

1. For all $t > 0$,

$$\mathbb{P}(|x - \mu| \geq t) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq t\right) \leq \underbrace{2\exp\left(-\frac{nt^2}{2}\right)}_{\xrightarrow{n\to\infty} 0}.$$

2. Equivalently, for any $0 < \delta < 1$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq \sqrt{\frac{2\log(2/\delta)}{n}}\right) \leq \delta.$$

3. Equivalently,

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu < \sqrt{\frac{2\log(2/\delta)}{n}}\right|$$

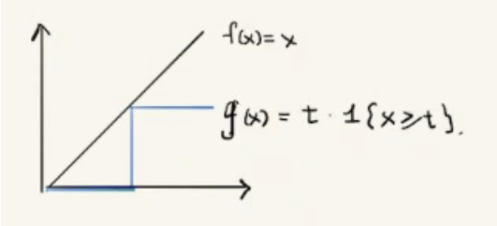
with probability at least $1 - \delta$, or **with high probability**.

1.2 Markov's inequality

Lemma 1.1 (Markov's inequality)**.** *Let X be a nonnegative random variable. Then for all $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Define $f(x) = x$ and $g(x) = t\mathbb{1}_{\{x \geq t\}}$. Then $f(x) \geq g(x)$.



Then

$$\mathbb{E}[X] \geq \mathbb{E}[t\mathbb{1}_{\{X \geq t\}}] = t\mathbb{P}(X \geq t).$$ □

Markov's inequality is important because other concentration inequalities are consequences of Markov's inequality. For our example, we can apply Markov's inequality to $|X - \mu|$ with $X = \frac{1}{n}\sum_{i=1}^n Z_i$ to get

$$\begin{aligned}\mathbb{P}(|X - \mu| \geq t) &\leq \frac{\mathbb{E}[|X - \mu|]}{t} \\ &= \frac{\mathbb{E}[|\frac{1}{n}\sum_{i=1}^n Z_i - \mu|]}{t}\end{aligned}$$

Using Jensen's inequality, we can upper bound this by

$$= \frac{\mathbb{E}[|\frac{1}{n}\sum_{i=1}^n Z_i - \mu|^2]^{1/2}}{t}$$

Observe that $\mathbb{E}[(\frac{1}{n}\sum_{i=1}^n Z_i - \mu)^2] \leq n\,\mathbb{E}[(Z_i - \mu)^2]/n^2 \leq 1/n$. So we get

$$\begin{aligned}&\leq \frac{(1/n)^{1/2}}{t} \\ &= \frac{1}{\sqrt{n}t}.\end{aligned}$$

To rearrange this in terms of a tail probability δ, solve $\frac{1}{\sqrt{n}t} = \delta$:

$$\mathbb{P}\left(|X - \mu| \geq \frac{1}{\sqrt{n}\delta}\right) \leq \delta.$$

That is,

$$|X - \mu| < \frac{1}{\sqrt{n}\delta}$$

with probability at least $1 - \delta$. Here, we have gotten the correct $1/\sqrt{n}$ scaling, but the $1/\delta$ dependence is not optimal yet.

Remark 1.1. Letting $n \to \infty$ gives us a weak law of large numbers. However, if we sum these probabilities in n, we get a divergent sum, so we would need to be more careful if we wanted to use the Borel-Cantelli lemma to prove a strong law of large numbers.

1.3 Chebyshev's inequality

Lemma 1.2. *If* $\mathrm{Var}(X)$ *exists, then or all* $t > 0$*,*

$$\mathbb{P}(X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. Apply Markov's inequality:

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq t) &\leq \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]t^2}{.}\end{aligned}$$

□

For our example, apply Chebyshev's inequality to $X = \frac{1}{n}\sum_{i=1}^n Z_i$ to get

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{\sum_{i=1}^n Z_i - \mu}\right| \geq t\right) &\leq \frac{\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n Z_i)}{t^2} \\ &= \frac{\mathrm{Var}(Z_i)}{nt^2} \\ &\leq \frac{1}{nt^2}.\end{aligned}$$

Solving $\delta = \frac{1}{nt^2}$, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq \frac{1}{\sqrt{n}\sqrt{\delta}}\right) \leq \delta.$$

That is,

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq \frac{1}{\sqrt{n}\sqrt{\delta}}$$

with probability at least $1 - \delta$. In comparison to our application of Markov's inequality, this gives a $1/\sqrt{\delta}$ dependence instead of a $1/\delta$ dependence, which is significant when δ is small.

In general, we have

Lemma 1.3. *For all $t > 0$,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k}{t^k},$$

provided this k-th moment exists.

As an exercise, apply this to our example and carefully bound $\mathbb{E}[|\frac{1}{n}\sum_{i=1}^n Z_i - \mu|^k]$ to show that there is a constant $C_k < \infty$ such that

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \leq \frac{C_k}{\sqrt{n}\delta^{1/k}}$$

with probability at least $1 - \delta$.

As another exercise, derive Cantelli's inequality using the same principle:

Lemma 1.4 (Cantelli's inequality)**.**

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2}.$$

Proof. The events $\{X - \mu \geq t\} = \{f(x - \mu) \geq f(t)$ are teh same, where $f(t) = (t + u)^2$ for some special choice of u. □

1.4 Chernoff's inequality

Lemma 1.5 (Chernoff's inequality)**.** *For all $t > 0$, we have*

$$\begin{aligned}\mathbb{P}(X \geq \mu + t) &\leq \inf_\lambda \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e}^{-\lambda t} \\ &= e^{-h(t)},\end{aligned}$$

where

$$h(t) = \sup_\lambda \lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}].$$

Proof. We will prove the inequality. We can upper bound the tail probability by rewriting this event:

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t})$$

This holds for all λ, so it holds for the inf over all λ. We get

$$\begin{aligned}\mathbb{P}(X - \mu \geq t) &= \inf_\lambda \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \\ &\leq \inf_\lambda \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}},\end{aligned}$$

where we have used Markov's inequality. □

Remark 1.2. To interpret the quantities in the bound, define the **moment generating function** of a random variable Z as

$$M_Z(\lambda) := \mathbb{E}[e^{\lambda Z}].$$

This is called the moment generating function because

$$\frac{d}{d\lambda} M_Z(\lambda)|_{\lambda=0} = \mathbb{E}_Z[Ze^{\lambda Z}]|_{\lambda=0} = \mathbb{E}[Z].$$

In general,

$$\frac{d^k}{d\lambda^k} M_Z(\lambda)|_{\lambda=0} = \mathbb{E}_Z[Z^k e^{\lambda Z}]|_{\lambda=0} = \mathbb{E}[Z^k],$$

the k-th moment.

Define the **cumulant generating function** of Z as

$$K_Z(\lambda) := \log \mathbb{E}[e^{\lambda Z}] = \log M_Z(\lambda).$$

This is called the cumulant generating function because it generates the **cumulants**

$$\kappa_k = \frac{d^k}{d\lambda^k} K_Z(\lambda)|_{\lambda=0}.$$

For example, $\kappa_2 = \mathrm{Var}(Z) \geq 0$. In fact, $K_Z''(\lambda) \geq 0$, so the cumulant generating function is always convex.

Define the Legendre transform f^* of $f : \mathbb{R} \to \mathbb{R}$ as

$$f^*(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - f(\lambda).$$

Then $h(t)$ is the Legendre transform of $K_{X-\mu}(\lambda)$. The Legendre transform can be thought of as a dual[1] in the sense that $f^{**}(\lambda) = (f^*)^*(\lambda) = f(\lambda)$ if f is convex.

For our example, apply Chernoff's inequality to $X = \frac{1}{n}\sum_{i=1}^n Z_i$. Here is a claim we will prove next lecture: If $Z \sim \mathbb{P}_Z \in \mathcal{P}([0,1])$, then

$$\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\lambda^2/2}, \qquad \forall \lambda \in \mathbb{R}.$$

Using this claim, we bound

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mu \geq t\right) &\leq \inf_\lambda \frac{\mathbb{E}[e^{\lambda(\frac{1}{n}\sum_{i=1}^n Z_i - \mu)}]}{e^{\lambda t}} \\
&= \inf_\lambda \frac{\mathbb{E}[\prod_{i=1}^n e^{\lambda \frac{1}{n}(Z_i - \mu)}]}{e^{\lambda t}}
\end{aligned}$$

[1] The Legendre transform is sometimes known as the **Fenchel dual**.

Using independence of the Z_i,

$$\begin{aligned}
&= \inf_\lambda \frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda \frac{1}{n}(Z_i - \mu)}]}{e^{\lambda t}} \\
&= \inf_\lambda \frac{\mathbb{E}[e^{\lambda \frac{1}{n}(Z_i - \mu)}]^n}{e^{\lambda t}} \\
&\leq \inf_\lambda \frac{(e^{(\lambda/n)^2/2})^n}{e^{\lambda t}} \\
&= \inf_\lambda e^{\lambda^2/(2n) - \lambda t}
\end{aligned}$$

This exponent is quadratic in λ, so we can calculate that it is minimized at $\lambda_* = nt$.

$$\begin{aligned}
&= e^{-(nt)^2/(2n) - nt \cdot t} \\
&= e^{-nt^2/2}.
\end{aligned}$$

We will apply this line of reasoning again and again in this course.

Similarly, we have the lower bound

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mu \leq -t\right) \leq e^{-nt^2/2}.$$

Combining these two tail inequalities, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq t\right) \leq 2e^{-nt^2/2}.$$

This is the inequality we presented at the beginning of the lecture. If we solve $\delta = 2e^{-nt^2/2}$, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \geq \sqrt{\frac{2\log(2/\delta)}{n}}\right) \leq \delta.$$

That is,

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| < \sqrt{\frac{2\log(2/\delta)}{n}}$$

with probability at least $1 - \delta$.

1.5 Comparison of inequalities

Here is a table comparing the different inequalities we have seen.

	Markov	Chebyshev	k-th moment	Chernoff
require	First moment	Second moment	k-th moment	Moment generating function
bound	$\frac{1}{\sqrt{n\delta}}$	$\frac{1}{\sqrt{n}\sqrt{\delta}}$	$\frac{1}{\sqrt{n}\delta^{1/k}}$	$\frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}$

Using more moments, we get better bounds; using the MGF is like using all the moments of a random variable. These have the same dependence in n but different dependence in δ. What is the benefit of better dependence in δ? This is useful for the union bound!

1.6 Applying union bounds

Lemma 1.6 (Union bound)**.** *Suppose we have a collection of events* $\{E_s\}_{s\in[d]}$*. If* $\mathbb{P}(E_s^c) \le \frac{\delta}{d}$ *for all* s*, then*

$$\mathbb{P}\left(\bigcup_{s\in[d]} E_s\right) \ge 1-\delta.$$

So if we divide delta by the number of events d, we can use a good δ dependence to get a good union bound.

Remark 1.3. Here is a common mistake that happens in homework, exams, and even ICML and NeurIPS papers. Let $(Z_i^{(s)})_{i\in[n],s\in[d]} \overset{\text{iid}}{\sim} \mathbb{P}_Z \in \mathbb{P}([0,1])$. Suppose someone proves that for all $s \in [d]$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^{(s)} - \mu\right| \le \sqrt{\frac{\log(1/\delta)}{n}}\right) \ge 1-\delta.$$

The common mistake is to claim that

$$\mathbb{P}\left(\forall s\in[d], \left|\frac{1}{n}\sum_{i=1}^n Z_i^{(s)} - \mu\right| \le \sqrt{\frac{\log(1/\delta)}{n}}\right) \ge 1-\delta.$$

This is not true because it ignores the dependence on the dummy variable s. Instead, the correct thing to do is to say

$$\mathbb{P}\left(\forall s\in[d], \left|\frac{1}{n}\sum_{i=1}^n Z_i^{(s)} - \mu\right| \le \sqrt{\frac{\log(d/\delta)}{n}}\right) \ge 1-\delta.$$

This d is usually very large, such as exponential or doubly exponential in n.

So please avoid the following statement:

$$\forall s\in[d], \qquad \left|\frac{1}{n}\sum_{i=1}^n Z_i^{(s)} - \mu\right| \le \varepsilon n, \qquad \text{with probability at least } 1-\delta.$$

This is ambiguous if the probability applies to each individual s or all s at once. Instead, use this statement instead:

For individual bounds, write

(a) $\forall s \in [d]$, $\mathbb{P}(\cdots) \geq 1 - \delta$.

(b) $\forall s \in [d]$, with probability at least $1 - \delta$, the following event happens:

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \varepsilon n.$$

For union bounds use these:

(a) $\mathbb{P}(\forall s, \cdots) \geq 1 - \delta$.

(b) With probability at least $1 - \delta$, the following event happens:

$$\forall s \in [d], \qquad \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \varepsilon n.$$

(c)

$$\sup_{s \in [d]} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \varepsilon n \qquad \text{with probability at least } 1 - \delta.$$

Here are some exercises to do for using union bounds:

Suppose $(Z_i^{(s)})_{i \in [n], s \in [d]} \overset{\text{iid}}{\sim} \mathbb{P}_Z \in \mathcal{P}([0, 1])$.

- Markov's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \qquad \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \frac{d}{\sqrt{n}\delta}.$$

- Chebyshev's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \qquad \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \frac{\sqrt{d}}{\sqrt{n}\sqrt{\delta}}.$$

- Markov's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \qquad \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^{(s)} - \mu \right| \leq \frac{\sqrt{2 \log(2d/\delta)}}{\sqrt{n}}.$$