Lecture 13: Concentration Inequalities

Isabella Zhu

20 March 2025

§1 Setup of Concentration Inequalities

We have a function f of a large number of i.i.d random variables $f(X_1, \ldots, X_n)$ and f does not depend "too much" on any of the X_is. Then, f will be close to its expectation in the sense that

$$\mathbb{P}(|f(X_1, \ldots X_n) - \mathbb{E}[f(X_1, \ldots X_n)]| > t) \leq e^{-t^2/2\sigma_f^2}$$

The canonical example we did already is Hoeffding's inequality, essentially when $f(X_1, \ldots X_n) = \bar{X_n}$. The "not depend too much" refers to when we wiggle X_i from the bottom to top of its bound $[a, b]$, the average changes by at most $\frac{b-a}{n}$, which is small.

We will henceforth refer $\vec{X} = (X_1, \ldots X_n)$. We define

$$D_i f(\vec{X}) = \sup_z f(X_1, \ldots X_{i-1}, z, \ldots) = \inf_z f(X_1, \ldots X_{i-1}, z, \ldots)$$

and we will impose the condition that $\max_i \sup_{\vec{X}} |D_i(\vec{X})| < \epsilon$.

§2 Hoeffding-Azuma Inequality

Definition 2.1. Define Δ_i as

$$\Delta_i = \mathbb{E}[f(\vec{X} \mid X_1, \ldots X_i)] - \mathbb{E}[f(\vec{X} \mid X_1, \ldots X_{i-1})]$$

Note that this telescopes, as

$$\sum_{i=1}^{n} \Delta_i = \mathbb{E}[f(\vec{X}) \mid \vec{X}] - \mathbb{E}[f(\vec{X})] = f(\vec{X}) - \mathbb{E}[f(\vec{X})]$$

Claim 2.2 — Δ_i is bounded and in $[-\epsilon, \epsilon]$ for all i.

We have

$$\Delta_i = \mathbb{E}[\mathbb{E}[f(\vec{X}) \mid X_i] - f(\vec{X}) \mid X_1, \ldots X_{i-1}]$$

which we can bound with

$$\begin{aligned}\Delta_i &= \mathbb{E}[\mathbb{E}[f(\vec{X}) \mid X_i] - f(\vec{X}) \mid X_1, \ldots X_{i-1}] \\ &\leq \mathbb{E}[\sup_z \mathbb{E}[f(\vec{X}) \mid X_i = z] - f(\vec{X}) \mid X_1, \ldots X_{i-1}] \leq \epsilon\end{aligned}$$

Therefore, we have

$$f(\vec{X}) - \mathbb{E}[f(\vec{X})] = \sum_{i=1}^{n} \Delta_i$$

But we can't apply Hoeffding because not independent. Hoeffding's lemma tells us

$$\mathbb{E}[e^{s\Delta_i} \mid X_1, \dots X_{i-1}] \leq e^{s^2\epsilon^2/2}$$

We apply a Chernoff bound

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} \Delta_i > t) &\leq \mathbb{E}[e^{s\sum_{i=1}^{n}\Delta_i}]e^{-st} \\
&= \mathbb{E}[\mathbb{E}[e^{s\sum_{i=1}^{n}\Delta_n} \mid X_1, \dots X_{n-1}]]e^{-st} \\
&= \mathbb{E}[e^{s\sum_{i=1}^{n-1}\Delta_i}]\mathbb{E}[e^{s\Delta_n} \mid X_1, \dots X_{n-1}]e^{-st} \\
&\leq \mathbb{E}[e^{s\sum_{i=1}^{n-1}\Delta_i}] \cdot e^{-s^2\epsilon^2/2} \cdot e^{-st} \\
&\leq e^{ns^2\epsilon^2/2}e^{-st}
\end{aligned}$$

which we get by "peeling" off each X_i one by one. Optimizing over s, we get $e^{-t^2/2n\epsilon^2}$ as our optimal bound.

Theorem 2.3

Let $\{F_n\}_k$ be filtration, $\{\Delta_n\}_k$ be a martingale difference sequence, i.e.

$$\mathbb{E}[\Delta_{k+1} \mid F_k] = 0$$

and $a \leq \Delta_k \leq b$. We have that

$$\mathbb{P}(|\frac{1}{n}\sum_{k=1}^{n}\Delta_k| > t) \leq 2e^{-t^2/2(b-a)^2}$$

§3 McDiarmid's Inequality

Theorem 3.1

Let $X_1, \dots, X_n$ be independent and assume that $\max_i ||D_i f(.)||_\infty < \epsilon$, then

$$\mathbb{P}(|f(\vec{X}) - \mathbb{E}[f(\vec{X})]| > t) \leq 2e^{-t^2/2n\epsilon^2}$$

Remark 3.2. The power of this is that f can be extremely non-linear.

§4 Vector Hoeffding's Inequality

Theorem 4.1

Let $X_1, \dots, X_n$ be independent vectors and assume that $\mathbb{E}[X_i] = 0$, and X_i is in Hilbert space X. Assume that $||X|| \leq R$ almost surely. Then, consider $f(\vec{X}) = ||\bar{X}_n||$. We have

$$||\bar{X}_n|| - \mathbb{E}[||\bar{X}_n||] \leq R\sqrt{2\log(2/\delta)/n}$$

with probability at least $1 - \delta$.

Proof. We have

$$D_i f(\vec{X}) = \frac{1}{n}||\dots + X_i^+|| - \frac{1}{n}||\dots + X_i^-|| \leq \frac{1}{n}||X_i^+ - X_i^-|| \leq \frac{2R}{n}$$

McD tells us that

$$||\bar{X}_n|| - \mathbb{E}[||\bar{X}_n||] \leq R\sqrt{2\log(2/\delta)/n}$$

with probability at least $1 - \delta$. We also have for Hilbert spaces,

$$\mathbb{E}[||\bar{X}_n||] \leq \sqrt{\mathbb{E}[||\bar{X}_n||^2]} = \frac{1}{n}\left(\mathbb{E}\sum_{i,j}\langle X_i, X_j\rangle\right)^{1/2} = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}[||X_i||^2]\right)^{1/2} \leq \frac{R}{\sqrt{n}}$$

§5 Application to SGD

Our goal is to solve the optimization problem

$$\min_{\theta\in\Theta} F(\theta) = \mathbb{E}[f(\theta, \xi)]$$

where $\xi_1, \dots \xi_n$ are observed random variables. We will assume that F and Θ are convex and $f(., \psi)$ is differentiable almost surely.

We can do gradient descent

$$\theta_{k+1} = \theta_k - \alpha_n \nabla F(\theta_k)$$

and then project onto Θ

$$\theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_n \nabla F(\theta_k))$$

but we don't know F, so we will do SGD

$$\theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_n \nabla f(\theta_k, \xi_k))$$

Theorem 5.1

Assume $\Theta \subset R \cdot B_2$, $\theta_0 \in \Theta$ and $|\nabla f(\theta, \xi)|_2 \leq M$ almost surely. Then, we have $\bar{\theta} = \frac{1}{n}\sum_{k=1}^{n} \theta_k$ satisfies

$$F(\bar{\theta}) - \min_{\theta \in \Theta} F(\theta) \leq \frac{2R^2}{n\alpha_n} + \frac{M^2}{2}\bar{\alpha} + 2RM\sqrt{2\log(2/\delta)/n}$$

with probability at least $1 - \delta$. In particular, if we take $\alpha_k = \frac{1}{\sqrt{k}}$, then

$$F(\bar{\theta}) - \min_{\theta \in \Theta} F(\theta) \lesssim \frac{1}{\sqrt{n}}$$

Proof. We have

$$\mathbb{E}[g_k \mid \xi_1, \dots \xi_{k-1}] = \mathbb{E}[\nabla f(\theta_k, \xi_k) \mid \xi_1, \dots \xi_{k-1}] = \nabla F(\theta_k)$$

We have

$$\begin{aligned}
|\theta_{k+1} - \theta^*|_2^2 &= |\Pi_\Theta(\theta_k - \alpha_k g_k) - \theta^*|_2^2 \\
&\leq |\theta_k - \alpha_k g_k - \theta^*|_2^2 \\
&= |\theta_k - \theta^*|_2^2 - 2\alpha_k\langle \theta_k - \theta^*, g_k\rangle + \alpha_k^2 |g_k|_2^2 \\
&\leq |\theta_k - \theta^*|_2^2 + \alpha_k^2 M^2 - 2\alpha_k\langle \theta_k - \theta^*, \nabla F(\theta_k)\rangle + 2\alpha_k\langle \theta_k - \theta^*, \nabla F(\theta_k) - g_k\rangle
\end{aligned}$$

By convexity, $\langle \theta_k - \theta^*, \nabla F(\theta_k)\rangle \geq F(\theta_n) - F(\theta^*)$. We define

$$\Delta_k = \langle \theta_k - \theta^*, \nabla F(\theta_k) - g_k\rangle$$

Chugging along, we have

$$F(\theta_k) - F(\theta^*) \leq \frac{1}{2\alpha_k}[|\theta_k - \theta^*|_2^2 - |\theta_{k+1} - \theta^*|_2^2] + \alpha_k^2 M^2 + \Delta_k$$

Taking the average over all k, we get

$$\frac{1}{n}\sum_{k=1}^{n} F(\theta_k) - F(\theta^*) \leq \frac{1}{2n}\sum_{k=1}^{n} \frac{1}{\alpha_k}[|\theta_k - \theta^*|_2^2 - |\theta_{k+1} - \theta^*|_2^2] + M^2\bar{\alpha} + \bar{\Delta}$$

which we can do some telescoping stuff

$$\begin{aligned}
&\sum_{k=1}^{n} \frac{1}{\alpha_k}[|\theta_k - \theta^*|_2^2 - |\theta_{k+1} - \theta^*|_2^2] \\
&= \frac{1}{\alpha_1}|\theta_1 - \theta^*|_2^2 - \frac{1}{\alpha_{n+1}}|\theta_{n+1} - \theta^*|_2^2 + \sum_{k=1}^{n}\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k+1}}\right)|\theta_{k+1} - \theta^*|_2^2
\end{aligned}$$

and we know $|\theta_{k+1} - \theta^*|_2^2 \leq 4R^2$ so we get the whole thing is less than

$$\frac{4R^2}{2n\alpha_n} + M^2\bar{\alpha} + \bar{\Delta}$$

Then we finish with Cauchy Schwarz as

$$|\Delta_n| \leq |\theta_n - \theta^*|_2 |\nabla F(\theta_n) - g_n)|_2 \leq 4RM$$

and Hoeffding-Azuma implies that $\bar{\Delta} \leq 2RM\sqrt{2\log(1/\delta)/n}$ with probability $1-\delta$.