Basic Tail Inequality 1

Wei Xiong

Department of Mathematics
The Hong Kong University of Science and Technology

2022.2.14

1 Concentration

Concentration

- In a variety of settings, it is of interest to obtain bounds on the tails of a random variable that guarantee that it is close to its mean.

$$P(|X - EX| \geq t) < \text{some probability related to t}$$

- Law of large number Let $X_1, X_2, ...$ be i.i.d. random variables with $EX_1 = \mu$. Then, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \mu$$

- Central limit theorem Let $X_1, X_2, ...$ be i.i.d. random variables with $EX_1 = \mu$ and $VarX_1 = \sigma^2$ and let $S_n = \sum_{i=1}^{n} X_i$.Then w.p. 1, we have

$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \rightarrow N(0,1) \;\; \text{in distribution}$$

  - Asymptotic behavior is not satisfactory for our analysis needs.

Tail Probability

Under mild condition, we can have an upper bound for the tail probability.

- Markov's inequality Let X be a non-negative random variable in the sense that $X \geq 0$ w.p. 1. Then,

$$P(X \geq t) \leq \frac{EX}{t}$$

- Markov's inequality is sharp in the sense that we can find some distribution for which the bound is tight;
- $O(\frac{1}{t})$ rate can be improved to $O(\exp(\text{poly(t)}))$ in some cases, e,g., Gaussian:

$$\int_x^\infty \phi(t)\mathrm{d}t = \int_x^\infty \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\mathrm{d}t = \int_x^\infty \frac{1}{t}\frac{1}{\sqrt{2\pi}} t\cdot\exp(-t^2/2)\mathrm{d}t$$
$$= -\frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\Big|_x^\infty - \int_x^\infty \left(-\frac{1}{t^2}\right)\left(-\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\right)\mathrm{d}t$$
$$= \frac{\phi(x)}{x} - \int_x^\infty \frac{\phi(t)}{t^2}\mathrm{d}t \leq \frac{\phi(x)}{x},$$

Chernoff Trick

- Markov's inequality Let X be a non-negative random variable in the sense that $X \geq 0$ w.p. 1. Then,

$$P(X \geq t) \leq \frac{EX}{t}$$

- Chernoff bound If moment generating function $\phi(\lambda) = E[e^{\lambda X}]$ exists,

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}} = \phi(\lambda)e^{-\lambda t}$$

  - This holds for all $\lambda > 0$ and we can take inf on the right side;
  - The deviation probability grows depending on the rate of moment generating function;
  - This trick is frequently used in our derivation.

Sub-Gaussianity

- Sub-Gaussianity A random variable X is σ^2-sub-Gaussian if for all $\lambda \in R$ it holds that

$$E[\exp(\lambda(X - EX))] \leq \exp(\frac{\lambda^2\sigma^2}{2})$$

- Let $X_1, ..., X_n$ be independent σ_i^2-sub-Gaussian random variables. Then $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian:

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right)\right] \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right), \forall \lambda \in R$$

- If X is σ^2-sub-Gaussian, then cX is $c^2\sigma^2$-sub-Gaussian;
- Let X_i be i.i.d. σ^2-sub-Gaussian, then, $\bar{X}_n$ is $\frac{1}{n}\sigma^2$-sub-Gaussian.
- Consider the Gaussian with variance σ_i^2.

Sub-Gaussian: Examples

- The Gaussian is sub-Gaussian where the equality holds.
- Bounded random variables Let X_i be independent bounded random variables supported on $[a_i, b_i]$ respectively. We have the classical Hoeffding bound:

$$P(\sum_{i=1}^{n}(X_i - EX_i) \geq t) \leq \exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}).$$

- In the literature of multi-armed bandit and reinforcement learning, the boundedness assumption is commonly adopted.

Tail bound of Sub-Gaussian random variable.

$$P(X \geq \mathbb{E}X + t) \leq \exp(-\frac{t^2}{2\sigma^2}),$$
$$P(X \leq \mathbb{E}X - t) \leq \exp(-\frac{t^2}{2\sigma^2}). \tag{0.1}$$

Proof.

By the Chernoff trick, we have

$$P(X - \mathbb{E}X > t) = P(\exp(\lambda(X - \mathbb{E}X) > \exp(\lambda t)) \leq \frac{\mathbb{E}\exp(\lambda(X - \mathbb{E}X))}{\exp(\lambda t)}.$$

Combined this with the definition of sub-Gaussian r.v., we have

$$P(X - \mathbb{E}X < t) \leq \exp(\frac{\lambda^2}{2}\sigma^2 - \lambda t).$$

Optimizing RHS w.r.t. $\lambda > 0$, we set $\lambda = \frac{t}{\sigma^2}$. □

Sub-Gaussianity is preserved under linear combination.

Let X_i be **independent** σ_i^2-sub-Gaussian r.v.s. Then, $\sum_{i=1}^n c_i X_i$ is $\sum_{i=1}^n c_i^2 \sigma^2$-sub-Gaussian.

Proof.

$$\begin{aligned}
\mathbb{E}\exp(\lambda(Z - \mathbb{E}Z)) &= \mathbb{E}\left[\prod_{i=1}^n \exp(\lambda(X_i - \mathbb{E}X_i)\right] \\
&= \prod_{i=1}^n \mathbb{E}\exp(\lambda(X_i - \mathbb{E}X_i) \quad \text{Independence} \\
&\leq \exp(\lambda^2 \frac{\sum_{i=1}^n \sigma_i^2}{2}) \quad \text{Sub-Gaussian assumption.}
\end{aligned}$$

$\square$

Hoeffding's inequality

Let X_i be **independent** σ_i^2-sub-Gaussian r.v.s. As a corollary of the previous two slide, we have

$$
\begin{aligned}
P(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t) &\leq \exp(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}) \\
P(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq -t) &\leq \exp(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}) \\
P(|\sum_{i=1}^{n}X_i - \mathbb{E}X_i| \geq t) &\leq 2\exp(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2})
\end{aligned}
\tag{0.2}
$$

where the last inequality follows from a union bound.

Classical Hoeffding's inequality

Let X_i be **independent** bounded random variables on $[a_i, b_i]$

$$
\begin{aligned}
P(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t) &\leq \exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}), \\
P(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq -t)\} &\leq \exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2})
\end{aligned} \tag{0.3}
$$

Or roughly, with $R = \max(b_i - a_i)$, w.p. at least $1 - \delta$, we have

$$
\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \leq \frac{R}{\sqrt{n}}\sqrt{\frac{\log(1/\delta)}{2}} = \tilde{\mathcal{O}}(\frac{R}{\sqrt{n}}) \tag{0.4}
$$

Discussion: Why Chernoff + SubGaussianity is better?

- Why Chernoff trick + Sub-Gaussianity improves the bound?
  - Markov's inequality only employs the information of expectation;
  - Sub-Gaussianity uses polynomial moments information;
  - One has: bound based on polynomial moments is always no worse than the Chernoff trick + Sub-Gaussianity

$$\inf_k \frac{\mathbb{E}|X|^k}{\delta^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}\exp(\lambda X)}{\exp(\lambda\delta)}$$

This is because

$$\mathbb{E}e^{\lambda X} = \sum_{k=0}^{\infty} \mathbb{E}\frac{\lambda^k |X|^k}{k!} \geq \sum_{k=0}^{\infty} \frac{(\lambda\delta)^k}{k!} \inf_k \mathbb{E}\frac{|X|^k}{\delta^k} = e^{\lambda\delta} \inf_k \mathbb{E}\frac{|X|^k}{\delta^k}.$$

Discussion: Tightness of Markov's inequality.

- Markov: $X \in \{0, t\}$ with $t > 0$:

$$\frac{\mathbb{E}X}{t} = P(X = t) = P(X \geq t).$$

- Chebyshev: Let $(X - \mathbb{E}X)^2 \in \{0, a\}$ (because Chebyshev follows from Markov):

$$X = \begin{cases} 2\sqrt{a} & \text{with probability } p \\ \sqrt{a} & \text{with probability } 1 - 2p \\ 0 & \text{with probability } p \end{cases}$$

Discussion: Linear Operation without Independence.

Suppose X_i is σ_i^2-sub-Gaussian.

- $X_1 + X_2$ is sub-Gaussian with at most $\sigma_1 + \sigma_2$;

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X_1+X_2)}\right] &\le \mathbb{E}\left[e^{p\lambda X_1}\right]^{1/p}\mathbb{E}\left[e^{q\lambda X_2}\right]^{1/q}\\ &\le \left(e^{\frac{1}{2}p^2\sigma_1^2\lambda^2}\right)^{1/p}\left(e^{\frac{1}{2}q^2\sigma_2^2\lambda^2}\right)^{1/q} = e^{\frac{1}{2}\left(p\sigma_1^2+q\sigma_2^2\right)\lambda^2}\end{aligned}$$

  where we use Holder's inequality with $1/p + 1/q = 1$. Finally, we note that we can set:

$$(\sigma_1+\sigma_2)^2 = \underbrace{(1+\sigma_2/\sigma_1)}_{p}\sigma_1^2 + \underbrace{(1+\sigma_1/\sigma_2)}_{q}\sigma_2^2.$$

Discussion: Non-Linear Operation.

Let X_i be σ^2-sub-Gaussian (not necessarily independent).

- $\mathbb{E}\max_i X_i \leq \sqrt{2\sigma^2 \log n}$;
- $\mathbb{E}\max_i |X_i| \leq 2\sqrt{\sigma^2 \log n}$

By convexity of $\exp(\cdot)$, we first have $\mathbb{E}\exp(\lambda \max_i X_i) \geq \exp(\lambda \mathbb{E}\max_i X_i)$. We further note that $\exp(\cdot)$ is monotone + non-negative. So it holds that

$$\mathbb{E}\left[\exp\left\{\lambda \max_{i\in[n]} X_i\right\}\right] = \mathbb{E}\left[\max_{i\in[n]} e^{\lambda X_i}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \leq n e^{\frac{\lambda^2\sigma^2}{2}}.$$

We then have $\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \frac{\log n}{\lambda} + \lambda\frac{\sigma^2}{2}$ and optimizing w.r.t. $\lambda > 0$ leads to

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \frac{\sigma}{\sqrt{2}}\sqrt{\log n} + \frac{\sigma}{\sqrt{2}}\sqrt{\log n} = \sqrt{2\sigma^2 \log n}.$$

The 2nd ineq. follows from $\max_i |X_i| = \max\{X_1, \cdots, X_n, -X_1, \cdots, -X_n\}$.

2 Uniform Convergence

Uniform Convergence 1

Union Bound. For any random events A and B, we have

$$P(A \cup B) \leq P(A) + P(B).$$

Suppose that $\{X_i\}_{i=1}^n$ are i.i.d. r.v.s and the support of X_i is a discrete set $\mathcal{S}$ with $|\mathcal{S}| = S$. We consider a class of functions $\mathcal{F} \subset (\mathcal{S} \to [0,1])$.

- f is fixed: then $f(X_i)$ are also i.i.d. r.v.s. and it holds that

$$\mathbb{P}\left(\left|\sum_{i=1}^n \left(f\left(x_i\right) - \mathbb{E} f\left(x_i\right)\right)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n}\right)$$

- $f \in \mathcal{F}$ is random, e.g., $\hat{f} := \mathrm{argmin}_{f \in \mathcal{F}}\, \ell(f, \{X_i\}_{i=1}^n)$ is random because the optimization process depends on the whole dataset:

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^n \left(\widehat{f}\left(x_i\right) - \mathbb{E}\widehat{f}\left(x_i\right)\right)\right| \geq t\right) &\leq \mathbb{P}\left(\exists f \in \mathcal{F}, \left|\sum_{i=1}^n \left(f\left(x_i\right) - \mathbb{E} f\left(x_i\right)\right)\right| \geq t\right) \\ &= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \left|\sum_{i=1}^n \left(f\left(x_i\right) - \mathbb{E} f\left(x_i\right)\right)\right| \geq t\right)\end{aligned}$$

Uniform Convergence 2

- *Finite $\mathcal{F}$.*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\left(\widehat{f}(x_i)-\mathbb{E}\widehat{f}(x_i)\right)\right|\geq t\right)=\mathbb{P}\left(\bigcup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(f(x_i)-\mathbb{E}f(x_i))\right|\geq t\right)$$
$$\leq 2|\mathcal{F}|\exp(-\frac{2t^2}{n}).$$

So,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\mathbb{E}f(x_i))\right|\leq\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{F}|}{\delta}}$$
$$=\mathcal{O}\left(\sqrt{\frac{1}{n}\log\frac{|\mathcal{F}|}{\delta}}\right)$$
$$=\tilde{\mathcal{O}}\left(\sqrt{\frac{\log|\mathcal{F}|}{n}}\right)$$

Uniform Convergence 3

- *Infinite* $\mathcal{F}$. We first find an ϵ-Covering $\mathcal{F}_\epsilon$ s.t. for all $f \in \mathcal{F}$, we can find $f_\epsilon \in \mathcal{F}_\epsilon$ with $\sup_{x\in\mathcal{S}} |f(x) - f_\epsilon(x)| \leq \epsilon$. First:

$$\sup_{f\in\mathcal{F}_\epsilon} \left| \frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - \mathbb{E}f\left(x_i\right)\right) \right| \leq \mathcal{O}\left(\sqrt{\frac{1}{n}(\log|\mathcal{F}_\epsilon| + \log\frac{1}{\delta})} \right)$$

Then, for all $f \in \mathcal{F}$, we find a corresponding f_ϵ and have

$$\left| \frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - \mathbb{E}f\left(x_i\right)\right) \right| = \left| \frac{1}{n}\sum_{i=1}^{n} \left(f_\epsilon\left(x_i\right) - \mathbb{E}of_\epsilon\left(x_i\right) + \left(f - f_\epsilon\right)\left(x_i\right) - \mathbb{E}\left[\left(f - f_\epsilon\right)\left(x_i\right)\right]\right.$$

$$\leq \underbrace{\left| \frac{1}{n}\sum_{i=1}^{n} \left(f_\epsilon\left(x_i\right) - \mathbb{E}f_\epsilon\left(x_i\right)\right) \right|}_{\text{finite set uniform concentration}} + \underbrace{2\epsilon}_{\text{discretization error}}$$

$$\leq \mathcal{O}\left(\epsilon + \sqrt{\frac{1}{n}\left(\log|\mathcal{F}|_\epsilon + \log\frac{1}{\delta}\right)}\right)$$

Discussion

- *Fixed f*:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \mathbb{E} f(x_i) \right) \right| \leq \tilde{\mathcal{O}} \left(\sqrt{\frac{1}{n}} \right)$$

- Finite $\mathcal{F}$:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \mathbb{E} f(x_i) \right) \right| \leq \tilde{\mathcal{O}} \left(\sqrt{\frac{\log |\mathcal{F}|}{n}} \right)$$

- Infinite $\mathcal{F}$:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \mathbb{E} f(x_i) \right) \right| \leq \tilde{\mathcal{O}} \left(\sqrt{\frac{\log |\mathcal{F}_\epsilon|}{n}} + \epsilon \right)$$

- We can tolerate exponentially many functions (or $|\mathcal{F}_\epsilon|$) since the bound depends on it via $\log(|\mathcal{F}|)$;

Establish Generalization Bound via Uniform Convergence

Problem setup: We consider a finite Hypothesis space $\mathcal{H}$ and assume (X, Y) is sampled from some unknown distribution $P(X, Y)$.

- For a fixed $f \in \mathcal{H}$, the *populatin risk* is defined by

$$L(f) = \mathbb{E}_{(X,Y)\sim P}\ell(f(X), Y),$$

- Given a dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, we define the *empirical risk* by

$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i).$$

- We have $\mathbb{E}_P\hat{L}(f) = L(f)$ and $\hat{L}(f) \to L(f)$.
- We assume $\ell(\cdot, \cdot) \in [0, 1]$ for simplicity. (Sub-Gaussian assumption)

Decomposition: What Causes Error

Suppose that we find a $\hat{f} \in \mathcal{H}$ by minimizing $\hat{L}(f)$ (e.g. we may run SGD/Adam) and assume that the minimzer of $L(f)$ is $f^* \in \mathcal{H}$. Then,

$$L(\hat{f}) - L(f^*) = \underbrace{\left(L(\hat{f}) - \hat{L}(\hat{f})\right)}_{A} + \underbrace{\left(\hat{L}(\hat{f}) - \hat{L}(f^*)\right)}_{B} + \underbrace{\left(\hat{L}(f^*) - L(f^*)\right)}_{C}$$

where $B \leq 0$ because $\hat{f}$ minimizes $\hat{L}$. It further holds that

$$\begin{aligned} L(\hat{f}) - L(f^*) &\leq \underbrace{\left(L(\hat{f}) - \hat{L}(\hat{f})\right)}_{A} + \underbrace{\left(\hat{L}(f^*) - L(f^*)\right)}_{C} \\ &\leq \sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)| + \left(\hat{L}(f^*) - L(f^*)\right) \\ &\leq 2 \sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)|. \end{aligned}$$

Applying Concentration 1

- *Finite $\mathcal{H}$.*

$$
\begin{aligned}
&P\left(\sup_{f\in\mathcal{H}}|L(f)-\hat{L}(f)|>\sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}|}}\right)\\
&\le\sum_{f\in\mathcal{H}}P\left(|L(f)-\hat{L}(f)|>\sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}|}}\right)\\
&\le|\mathcal{H}|\times\frac{\delta}{|\mathcal{H}|}=\delta,
\end{aligned}
\tag{0.5}
$$

- *Infinite $\mathcal{H}$.*

$$
\begin{aligned}
|L(f)-\hat{L}(f)|&=|\frac{1}{n}\sum_{i=1}^{n}\left((f_\epsilon-\mathbb{E}f_\epsilon)+(f-f_\epsilon)+\mathbb{E}(f_\epsilon-f)\right)(X_i)|\\
&\le|\frac{1}{n}\sum_{i=1}^{n}(f_\epsilon-\mathbb{E}f_\epsilon)|+|\frac{1}{n}\sum_{i=1}^{n}(f-f_\epsilon)|+|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f_\epsilon-f)|\\
&\le\sqrt{\frac{1}{2n}\log\frac{2}{\delta/|\mathcal{H}_\epsilon|}}+2\epsilon,
\end{aligned}
\tag{0.6}
$$

Applying Concentration 2

Consider $S = \{x \in \mathbb{R}^p : \|X\|_2 \leq B\}$. Then we can find an ϵ-covering of the ℓ_2-ball w.r.t. ℓ_2-norm with at most $(\frac{3B}{\epsilon})^p$ elements. Roughly speaking, if we further assume ℓ is κ-Lipschitz, then we have

- *Finite $\mathcal{H}$.*

$$|L(\hat{f}) - L(f^*)| \leq 2 \sup_{f \in \mathcal{H}} |L(f) - \hat{L}(f)| \leq \tilde{O}(\frac{\log |\mathcal{H}_\epsilon|}{\sqrt{n}})$$

- *Infinite $\mathcal{H}$ with dimension p.*

$$|L(\hat{f}) - L(f^*)| \leq O(\sqrt{\frac{p \max(1, \ln(\kappa B n))}{n}}) = \tilde{O}(\frac{p}{\sqrt{n}}).$$

We can regard the ϵ-covering number as a complexity measure of the hypothesis space.

Discussion

The bound given in the previous slide is not satisfactory: it only employs the information of dimension. But it motivates us to consider the generalization bound in the following form:

$$L(\hat{f}) - \hat{L}(\hat{f}) \leq \tilde{O}\left(\sqrt{\frac{Complexity(\mathcal{H})}{n}}\right).$$

A good idea is to relate the complexity to the distribution P: it measures how easy it is to learn the underlying distribution P. Combing this two ideas, we have the famous Rademacher complexity, which will be introduced later in our reading course.

Next Week

- Bernstein's inequality: Employs the variance information;
- Martingale concentration;
- Concentration of Functions beyond linear combination.