EECS208 Discussion 2

Simon Zhai

Sep. 17, 2021

Reading:

- **Appendix E of *High-Dim Data Analysis with Low-Dim Models*;**
- **Chapter 2 of *High-dimensional statistics: A non-asymptotic viewpoint*, by Martin Wainwright.**

1 Tail Bounds

Reading: High-dimensional statistics: A non-asymptotic viewpoint, Chapter 2.

1.1 Markov bound

Proposition 1.1 (Markov's Inequality) *Given a non-negative random variable x with finite mean, we have*

$$\mathbb{P}[x \geq t] \leq \mathbb{E}[x]/t, \quad \forall t > 0. \tag{1.1}$$

Proof $\forall t > 0$, consider random variable $t\mathbb{1}\{x \geq t\}$, we have

$$t\mathbb{1}\{x \geq t\} \leq x, \quad \forall t > 0, \tag{1.2}$$

taking expectation over both sides of the above inequality, we have

$$t\mathbb{P}[x \geq t] \leq \mathbb{E}x \implies \mathbb{P}[x \geq t] \leq \mathbb{E}x/t. \tag{1.3}$$

■

1.2 Chebyshev bound

Proposition 1.2 (Chebyshev's Inequality) *Given a random variable x with finite mean $\mathbb{E}x = \mu$ and finite variance, we have*

$$\mathbb{P}[|x - \mu| \geq t] \leq \mathsf{var}(x)/t^2, \quad \forall t > 0. \tag{1.4}$$

Proof Consider the random variable $|x-\mu|^2$, we know that $|x-\mu|^2$ is non-negative. Apply Markov's inequality to $|x - \mu|^2$ with t^2, we have

$$\mathbb{P}[|x - \mu|^2 \geq t^2] \leq \mathbb{E}|x - \mu|^2/t^2 \implies \mathbb{P}[|x - \mu| \geq t] \leq \mathsf{var}(x)/t^2. \tag{1.5}$$

■

1.3 Chernoff bound

Definition 1.3 (Definition of MGF from Wikipedia) *Let X be a random variable with cdf F_X. The moment generating function (mgf) of X (or F_X), denoted by $M_X(t)$, is*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] \tag{1.6}$$

provided this expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that for all t in $(-h, h)$, $\mathbb{E}\left[e^{tX}\right]$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

Suppose the random variable x has a moment generating function in a neighborhood of zero, meaning that there is some constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}[\exp(\lambda(x-\mu))]$ exists $\forall \lambda \leq |b|$. In this case, for any $\lambda \in [0, b]$, we can apply Markov's inequality to random variable $Y = \exp(\lambda(X-\mu))$, and obtain the upper bound

$$\mathbb{P}[(x-\mu) \geq t] = \mathbb{P}[\exp(x\lambda(x-\mu)) \geq \exp(\lambda t)] \leq \frac{\mathbb{E}[\exp(\lambda(x-\mu))]}{\exp(\lambda t)}. \tag{1.7}$$

Optimizing $\lambda \in [0, b]$ to obtain the tightest result yields the *Chernoff bound*:

$$\log \mathbb{P}[(x-\mu) \geq t] \leq \inf_{\lambda \in [0,b]} \left\{\log \mathbb{E}\left[\exp\left(\lambda(x-\mu)\right)\right] - \lambda t\right\}. \tag{1.8}$$

1.4 Sub-Gaussian bound

Definition 1.4 (Sub-Gaussian Random Variables) *A random variable X with mean $\mu = \mathbb{E}[X]$ is σ sub-Gaussian if there is a positive number σ such that $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}$, for all $\lambda \in \mathbb{R}$.*

Remark 1.5 *A Gaussian random variable with variance σ is σ sub-Gaussian.*

Applying $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}$, for all $\lambda \in \mathbb{R}$ to the Chernoff bound, we have

$$\mathbb{P}[x-\mu \geq t] \leq \exp[\sigma^2\lambda^2/2 - \lambda t], \tag{1.9}$$

by picking $\lambda = t/\sigma^2$, we have $\mathbb{P}[x-\mu \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$, which is the sub-Gaussian tail bound.

2 Examples of Sub-Gaussian Tail Bounds

Reading:

- High-Dim Data Analysis with Low-Dim Models, Appendix E;
- High-dimensional statistics: A non-asymptotic viewpoint, Chapter 2.

2.1 Hoeffding bound

Suppose that the variables $x_i, i = 1, \ldots, n$ are independent and x_i has μ_i and sub-Gaussian parameter σ_i. Then $\forall t \geq 0$, we have

$$\mathbb{P}\left[\sum_{i=1}^{n}(x_i - \mu_i) \geq t\right] \leq \exp\left[-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right]. \tag{2.1}$$

Another version of the Hoeffding inequality usually appears in for bounded difference inequality, since a bounded random variables in $[a_k, b_k]$ are sub-Gaussian with parameter at most $\sigma = (b_k - a_k)/2$:

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{k=1}^{n} x_i - \mathbb{E}x_i\right| \geq t\right] \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right). \tag{2.2}$$

2.2 Bernstein's inequality (Thm E.2) in High-Dim Data Analysis

Let $x_1, x_2, \dots, x_n$ be independent random variables, with $\mathbb{E}x_i = 0, |x_i| \leq R$ almost surely, and $\mathbb{E}[x_i^2] \leq \sigma^2, \forall i$. Then

$$\mathbb{P}\left[\left|\sum_{i=1^n} x_i\right| > t\right] \leq \exp\left(-\frac{t^2/2}{n\sigma^2 + 3Rt}\right). \tag{2.3}$$

2.3 Gaussian-Lipschitz Concentration

Let $f\mathbb{R}^m \mapsto \mathbb{R}$ be an L-Lipschitz function:

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^m. \tag{2.4}$$

Suppose $g_1, g_2, \dots g_m \sim_{iid} \mathcal{N}(0,1)$, then we have

$$\mathbb{P}\left[|f(g_1, \dots, g_m) - \mathbb{E}[f(g_1, \dots, g_m)]| > t\right] < 2\exp\left(-t^2/2L\right). \tag{2.5}$$

3 A (High-Level) Example of Applying High-Dim Statistics.

Suppose we are given a L-Lipschitz function $f_{\boldsymbol{A}}(\boldsymbol{x})$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n} \in \mathsf{G}$ (G is a matrix group, e.g., the orthogonal group) is a matrix and $\boldsymbol{x}$ is a random vector (e.g., Gaussian vector). Then we can use the following procedures to show that the sampled mean of $\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i)$ is a good approximation of the $\mathbb{E}_{\boldsymbol{x}} f_{\boldsymbol{A}}(\boldsymbol{x})$ *uniformly* for all $\boldsymbol{A} \in \mathsf{G}$:

- **Point-wise convergence:** show that for a given $\boldsymbol{A} \in \mathsf{G}$, applying the high-dimensional statistics concentration bounds we have discussed before, we have some exponential tail bounds like

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x})\right| > t\right) < 2\exp\left(-g(nt)\right), \tag{3.1}$$

  where $g(\cdot)$ is a monotonic increasing function.

- **ε-covering (Lemma 3.25 in High-dim Data Analysis, also refer to lecture note 06/07):** count how many ε-ball we need to cover the whole group G, suppose the number of ε-balls we need is N: meaning that we can find $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_N\}$, such that $\forall \boldsymbol{A} \in \mathsf{G}$, we can find $j \in [N]$, such that $\left\|\boldsymbol{A} - \boldsymbol{A}_j\right\|_\diamond < \varepsilon$.

- **Bound $\left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i) - \mathbb{E} f_{\boldsymbol{A}}\right|$ in a $\varepsilon-$Ball:** we can argue that $\forall \boldsymbol{A} \in \mathbb{B}(\boldsymbol{A}_j, \varepsilon)$, we have

$$\left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i) - \mathbb{E} f_{\boldsymbol{A}}(\boldsymbol{x})\right| < h(\varepsilon, n, L), \tag{3.2}$$

  where h is a function that is monotonic decreasing in ε.

- **Applying Union Bounds:** now we can argue that

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{k=1}^N \boldsymbol{A} \in \mathbb{B}(\boldsymbol{A}_k, \varepsilon), \left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x})\right| > t\right) \\
\leq &\sum_{j=1}^N \mathbb{P}\left(\boldsymbol{A} \in \mathbb{B}(\boldsymbol{A}_j, \varepsilon), \left|\frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{A}}(\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x})\right| > t\right|\right) \\
< &N\exp\left(-l(g(nt), h(\varepsilon, n, L))\right) = \exp\left(-l(g(nt), h(\varepsilon, n, L)) + \log N\right),
\end{aligned} \tag{3.3}$$

  where l is a positive function which is monotonic increasing w.r.t. n, and the sample complexity we are referring to is the order of n (e.g.,$O(n), O(n^2)$, etc.), such that $-l(g(nt), h(\varepsilon, n, L)) + \log N < 0$.