Concentration Inequalities II: Bernstein, Freedman, Martingale Methods and Applications

Presenter: Seth Strimas-Mackey
Scribe: Leo Huang

1st March 2019

Topics Covered

- Sub-Gaussian/Sub-Exponential Random Variables
- Bernstein's Inequality (3 types)
- Johnson-Lindenstrauss (JL) Lemma

1 Sub-Gaussian Random Variables

Definition 1 (Sub-Gaussian Random Variable). *A random variable X, with $\mathbb{E}[X] = 0$, is Sub-Gaussian with variance proxy σ^2, i.e., $X \sim subG(\sigma^2)$ if*

$$\mathbb{E}e^{sX} \leq e^{s^2\sigma^2/2} \quad \forall s \in \mathbb{R}$$

.

Theorem 1 (Sub-Gaussian Hoeffding). *Let $X_1, ...X_n \sim subG(\sigma^2)$ be independent, with $\mathbb{E}[X] = 0$. Then*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t) \leq e^{-nt^2/2\sigma^2}$$

In the bounded Case: $|X_i| \leq k, \sigma^2 \leq k^2$

Proof. Let $\bar(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$\begin{aligned} P(\overline{X} > t) &\leq e^{-st}\mathbb{E}[e^s\overline{X}] \\ &= e^{-st}(\mathbb{E}[e^{sX_i}/n])^n \\ &\leq e^{-st}(e^{s^2\sigma^2/2n^2})^n \\ &= e^{-st+s^2\sigma^2/2n} \end{aligned}$$

Setting $s = \frac{nt}{\sigma^2}$ minimizes the exponent, so that we have

$$P(\overline{X} > t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

□

As an example, one can observe that for $n = 1$, $P(X \geq t) \leq e^{-t^2/2\sigma^2}$. Sub-Gaussian random variables have Gaussian like tails.

2 Sub-exponential Random Variables

This class of random variables is similar to sub-Gaussian random variables, but have heavier exponential tails.

Example Consider Laplace(1), with $f_X(x) = \frac{1}{2}e^{-|x|}$ for $x \in \mathbb{R}$. Then, $P(|x| \geq t) = e^{-t}$, and clearly, X is not subG(σ^2) for any σ.

But for small s, i.e., $|s| < \frac{1}{2}$, $\mathbb{E}[e^{sX}] = \frac{1}{1-s^2} \leq e^{2s^2}$, which is bounded by a sub-Gaussian moment generating function. Thus, the random variable behaves like sub-gaussian for small s, but not as s gets larger. It turns out that this is more general.

Lemma 1. $\mathbb{E}[X] = 0, \mathbb{P}(|x| > t) \leq 2e^{-t/\lambda}, \lambda > 0 \implies \mathbb{E}[|x|^k] \leq 2\lambda^k k!$ and $\mathbb{E}[e^{sX}] \leq e^{2s^2\lambda^2}$.

This lemma is used to prove bound on all moments of X. The first step is to write $\mathbb{E}[|X|^k] = \int_0^\infty P(|x|^k > t)\, dt$. The second step is to Taylor expand and use the bound on $\mathbb{E}[|X|^k]$.

This motivates an equivalent definition using moment generating functions.

Definition 2. *(Sub-Exponential Random Variables) A random variable X, with $\mathbb{E}[X] = 0$, is Sub-Exponential with parameter λ, i.e., $X \in subE(\lambda)$, if $\mathbb{E}[e^{sX}] \leq e^{s^2\lambda^2/2}, \quad \forall\ |s| \leq \frac{1}{\lambda}$.*

3 Bernstein's Inequality

3.1 Bernstein's Inequality I

Theorem 2. *Let $X_1, ...X_n \sim subE(\lambda)$ be independent random variables with $\mathbb{E}[X] = 0$. Then,*

$$P(\frac{1}{n}\sum_{i=1}^n X_i > t) \leq \exp\left(-\frac{n}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda}\right)\right)$$

Proof.

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \leq e^{-snt}\prod_{i=1}^n \mathbb{E}[e^{sX_i}] \leq e^{-snt}e^{ns^2\lambda^2/2} = \exp(-snt + ns^2\lambda^2/2) = \exp\left(-\frac{n}{2}\left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda}\right)\right)$$

If, $|t| \leq \lambda^2$, optimizer s, otherwise set $S = \frac{1}{\lambda}$.

□

Lemma 2. Let $X \sim \text{subG}(\sigma^2)$. Consider $Z = X^2 - \mathbb{E}[X^2]$. Then $Z \sim \text{subE}[16\sigma^2]$.

Proof. (informal)

$$P(|x| > t) \leq 2e^{-ct^2}$$

implies,

$$P(X^2 > t) \leq 2e^{-ct^2}$$

implies,

$$P(X^2 > t) \leq 2e^{-ct}.$$

□

For bounded RV, we can get stronger version of Bernstein's Inequality, with smooth transition from regime to regime.

3.2 Bernstein's Inequality II

Theorem Consider $X_1, ..., X_n$ with $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma^2$ and $|X_i| \le K$. Then

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > t\right) \le \exp\left(\frac{-nt^2/2}{\sigma^2 + Kt/3}\right)$$

Proof. Using the standard Chernoff method,

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > t\right) &\le \frac{\mathbb{E}\left[e^{s\frac{\sum_{i=1}^{n} X_i}{n}}\right]}{e^{st}} \\ &= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s\frac{X_i}{n}}\right] \end{aligned} \tag{1}$$

Note that if $|s| < \frac{3n}{k}$, then $|s\frac{X_i}{n}| \le 3$, and thus using lemma 3

$$\begin{aligned} \mathbb{E}[e^{s\frac{X_i}{n}}] &\le 1 + \frac{s}{n}\mathbb{E}[X_i] + \mathbb{E}\left[\frac{\frac{s^2X_i^2}{2n^2}}{1 - \frac{|s||X_i|}{3n}}\right] \\ &\le 1 + \mathbb{E}\left[\frac{\frac{s^2X_i^2}{2n^2}}{1 - \frac{|s|K}{3n}}\right] \\ &\le 1 + \frac{\frac{s^2\sigma^2}{2n^2}}{1 - \frac{|s|K}{3n}} \\ &\le \exp\left(\frac{\frac{s^2\sigma^2}{2n^2}}{1 - \frac{|s|K}{3n}}\right) \end{aligned} \tag{2}$$

Using this back in Equation 1, we get:

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i > t\right) \le \exp\left(\frac{\frac{s^2\sigma^2}{2n}}{1 - \frac{|s|K}{3n}} - st\right)$$

Choosing $s = \frac{nt}{\sigma^2 + tK/3}$

$$\le \exp\left(\frac{-nt^2}{2(\sigma^2 + \frac{tK}{3})}\right)$$

□

Comparison to Hoeffding's inequality for Bounded Random variables For the same failure probability δ, Bernstein's inequality allows with probability at-least $1 - \delta$,

$$\frac{\sum_{i=1}^{n} X_i}{n} \le \mathbb{E}[X] + O\left(\frac{\sigma}{\sqrt{n}} + \frac{K}{n}\right)$$

whereas, Hoeffding's inequality gives us

$$\frac{\sum_{i=1}^{n} X_i}{n} \le \mathbb{E}[X] + O\left(\frac{K}{\sqrt{n}}\right).$$

Thus, for random variables for which $\sigma << K$, Bernstein gives a tigher rate.

Lemma 3. For all $z \in [-3, 3]$,

$$e^z \le 1 + z + \frac{z^2/2}{1 - |z|/3}$$

Proof. Proof by picture. Compare the two graphs. $\square$

3.2.1 An Application: Johnson Lindenstrauss Lemma

For any two vectors $a, a' \in \mathbb{R}^p$, define the distance to be $\|a - a'\|_2^2$. The Johnson Lindenstrauss (JL) lemma deals with the following question.

Question Given a finite set of n points, $A = \{a_1, ...a_n\} \subset \mathbb{R}^D$, with D large. Can one find a $d < D$ such that there exists a linear mapping $f : \mathbb{R}^D \mapsto \mathbb{R}^d$ that preserves distance upto an error ϵ, i.e., f is an $\epsilon-$isometry on $A \subset \mathbb{R}^D$, or,

$$(1 - \epsilon)\|a - a'\|^2 \le \|f(a) - f(a')\|^2 \le (1 + \epsilon)\|a - a'\|^2 \quad \forall a, a' \in A$$

We first show that such a mapping exists using the probabilistic method. Consider a random matrix $X \in \mathbb{R}^{d\times D}$ such that for all $i \in \{1, ...d\}, j \in \{1, ...D\}$, X_{ij} is an independent random variable with $\mathbb{E}[X_{ij}] = 0$ and $var(X_{ij}) = 1$ (for example normal gaussians). Define the function $f : \mathbb{R}^D \mapsto \mathbb{R}^d$ as $f(a) := \frac{1}{\sqrt{d}} Xa$, or for any $k \in [d]$, $f_k(a) = \sum_{j=1}^{d} X_{ij} a_j$. Thus, we have:

$$\mathbb{E}[f_k^2(a)] = \frac{1}{d} \sum_{j=1}^{D} a_j^2 \mathbb{E}[X_{ij}^2] = \|a\|^2$$

$$\mathbb{E}[\|f(a)\|^2] = \frac{1}{d} \mathbb{E} \sum_{k=1}^{d} f_k^2(a) = \|a\|^2$$

This shows that $\|\alpha\|^2$ is preserved by f in expectation. The following theorem gives a condition on d such that this also holds in high probability.

Theorem 3 (JL lemma)**.** *Let $A \subset \mathbb{R}^D$ such that $|A| = n$. Consider a matrix $X : \mathbb{R}^D \mapsto \mathbb{R}^d$ such that for all $i \in [d], j \in [D]$, $\;X_{ij} \in subG(\sigma^2)$ and is sampled independently. Then for any $\epsilon, \delta \in (0, 1)$, if $d \ge 100 \frac{\sigma^4}{\epsilon^2} \log(n/\sqrt{\delta})$, then with probability at-least $1 - \delta$, the linear map defined by $f(a) := \frac{1}{\sqrt{d}} Xa$ is an $\epsilon-$isometry on A, or more specifically,*

$$(1 - \epsilon)\|a - a'\|^2 \le \|f(a) - f(a')\|^2 \le (1 + \epsilon)\|a - a'\|^2 \quad \forall a, a' \in A$$

Before we provide a proof for the JL lemma, observe that the projection dimension d is independent of the dimension D of our feature vectors.

Proof. Define $T = \{\frac{a-a'}{\|a-a'\|} : a, a' \in A, a \ne a'\}$. We first state the following fact, which is quite easy to prove,

Fact 1. f is a linear $\epsilon-$isometry of A iff

$$\left|\|f(\alpha)\|^2 - 1\right| \le \epsilon \quad \forall \alpha \in T$$

We will thus show that with probabilty at-least $1-\delta$, $\left|\|f(\alpha)\|^2-1\right| \le \epsilon \quad \forall \alpha \in T$. Note that $|T| \le \binom{n}{2} \le \frac{n^2}{2}$, and $\mathbb{E}[f(\alpha)] = 1 \ \forall \alpha \in T$. First observe that $f_i(\alpha)$ is σ^2 sub-gaussian.

$$\begin{aligned}
\mathbb{E}[e^{sf_i(\alpha)}] &= \mathbb{E}\exp\left(s\sum_{j=1}^{D}\alpha_j X_{ij}\right) \\
&= \prod_{j=1}^{D}\exp(s\alpha_j X_{ij}) \\
&\le \exp(\frac{s^2\sigma^2}{2}\sum_{j=1}^{D}\alpha_j^2/\|\alpha\|^2) \\
&= e^{s^2\sigma^2/2}
\end{aligned}$$

Using Lemma 2, this implies that $f_i^2(\alpha) \sim \text{subE}(16\sigma^2)$. Thus, applying Bernstein's inequality, with $\mathbb{E}[f_i(\alpha)] = 1$ and taking a union bound over all $\alpha \in T$,

Using the Bernstein's inequality in the regime $|t| \le 16\sigma^2$,

$$P(\sup_{\alpha\in T}\left|\sum_{i=1}^{d}\frac{f_i^2(\alpha)}{d}-1\right| \ge \epsilon) \le 2\times\frac{n^2}{2}\exp(\frac{-dt^2}{2\times 16^2\sigma^4}),$$

Setting $t = \sqrt{\frac{512\sigma^4\log(n^2/\delta)}{d}}$ $(\le 16\sigma^2)$, we get,

$$P\left(\sup_{\alpha\in T}\left|\|f(\alpha)\|^2-1\right| \ge \sqrt{\frac{512\sigma^4\log(n^2/d)}{d}}\right) \ge \delta.$$

Thus, our choice of d suffices for ϵ-isometry. $\square$

3.3 Berstein's Inequality III: Martingales

Theorem 4 (Freedman's Inequality)**.** *Let $X_1,\ldots,X_n$ be a bounded martingale difference sequence, i.e., $\mathbb{E}[X_i|X_{i-1}] = 0$ and $|X_i| \le K$. Define the martingale $S_i = \sum_{j=1}^{i}X_i$. Additionally, define $\mathbb{E}_{i-1}[S_1]$ to be the expectation w.r.t. X_i while the random variables $X_1,\ldots,X_i$ are fixed. Similarly, define $V_n = \sum_{i=1}^{n}\mathbb{E}_{i-1}[X_i^2]$. Then,*

$$P\big(S_n > t \text{ and } V_n \le \sigma^2\big) \le \exp\left(\frac{-t^2/2}{\sigma^2+Kt/3}\right)$$

$E_{i-1}[S_i] = E_{i-1}[X_i] + S_{i-1} = S_{i-1}$. Let $V_n = \sum_{i=1}^{n}\mathbb{E}[X_i^2]$. Then

Proof. (informal) Previously, we saw $\mathbb{E}[e^{\lambda X_i}] \le \exp(\mathbb{E}[X_i^2\psi(\lambda)])$, $\psi(s) = \frac{\lambda^2/2}{1-|\lambda|K/3}$ (see (2)). Repeat argument using $\mathbb{E}_{i-1}$ instead of $\mathbb{E}$ to show that $\mathbb{E}_{i-1}e^{\lambda X_i} \le \exp(\mathbb{E}_{i-1}X_i^2\psi(\lambda))$. Then

$$
\begin{aligned}
P(S_n \geq t, V_n \leq \sigma^2) &= \mathbb{E}1(e^{\lambda S_n} \geq e^{\lambda t})1(V_n \leq \sigma^2) \\
&\leq e^{-\lambda t}\mathbb{E}\big[e^{\lambda S_n}1(V_n \leq \sigma^2)\big] \\
&= e^{-\lambda t}\mathbb{E}\big[e^{\lambda S_n - V_n\psi(\lambda)}e^{V_n\psi(\lambda)}1(V_n \leq \sigma^2)\big] \\
&\leq e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_n - V_n\psi(\lambda)}1(V_n \leq \sigma^2)\big] \\
&\leq e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_n - V_n\psi(\lambda)}\big] \\
&= e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_{n-1} - V_{n-1}\psi(\lambda) - \mathbb{E}_{n-1}[X_n]^2\psi(\lambda)} \times e^{\lambda X_n}\big] \\
&= e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_{n-1} - V_{n-1}\psi(\lambda) - \mathbb{E}_{n-1}[X_n]^2\psi(\lambda)} \times \mathbb{E}_{n-1}\big[e^{\lambda X_n}\big]\big] \\
&= \big[e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_{n-1} - V_{n-1}\psi(\lambda) - \mathbb{E}_{n-1}[X_n^2]\psi(\lambda)} \times e^{\mathbb{E}_{n-1}[X_n^2]\psi(\lambda)}\big] \\
&= e^{-\lambda t+\sigma^2\psi(\lambda)}\mathbb{E}\big[e^{\lambda S_{n-1} - V_{n-1}\psi(\lambda)}\big] \leq \ldots \leq e^{-\lambda t+\sigma^2\psi(\lambda)}
\end{aligned}
$$

Optimizing over λ gives the required tail bound. $\square$