Theoretical Statistics. Lecture 3.

Peter Bartlett

1. Concentration inequalities.

Review. Markov/Chebyshev Inequalities

Theorem: [Markov] For $X \geq 0$ a.s., $\mathbf{E}X < \infty$, $t > 0$:

$$P(X \geq t) \leq \frac{\mathbf{E}X}{t}.$$

Theorem: Chebyshev's inequality:

$$P(|X - \mathbf{E}X| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

Review. Chernoff technique

Theorem: For $t > 0$:

$$P(X - \mathbf{E}X \geq t) = P\left(\exp(\lambda(X - \mathbf{E}X)) \geq \exp(\lambda t)\right) \leq \frac{\mathbf{E}\exp(\lambda(X - \mathbf{E}X))}{\exp(\lambda t)} = e^{-\lambda t} M_{X-\mu}(\lambda).$$

Hence,

$$\log P(X - \mu \geq t) \leq -\sup_{\lambda > 0} \left(\lambda t - \Gamma(\lambda)\right),$$

where $\Gamma(\lambda) = \log M_{X-\mu}(\lambda)$ is the **cumulant generating function** of $X - \mu$.

Example: Gaussian

For $X \sim N(\mu, \sigma^2)$, $M_{X-\mu}(\lambda)$ is

$$\begin{aligned}
\mathbf{E}\exp(\lambda(X-\mu)) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp(\lambda x - x^2/(2\sigma^2))\,dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left(\lambda^2\sigma^2/2 - (x/\sigma - \lambda\sigma)^2/2\right)\,dx \\
&= \exp(\lambda^2\sigma^2/2)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-(y-\lambda\sigma)^2/2\right)\,dy \\
&= \exp(\lambda^2\sigma^2/2),
\end{aligned}$$

for the change of variable $y = x/\sigma$. Thus,

$$\begin{aligned}
\log P(X - \mu \geq t) &\leq -\sup_{\lambda > 0} \left(\lambda t - \log M_{X-\mu}(\lambda)\right) \\
&= -\sup_{\lambda > 0} \left(\lambda t - \frac{\lambda^2 \sigma^2}{2}\right) \\
&= -\frac{t^2}{2\sigma^2},
\end{aligned}$$

using the optimal choice $\lambda = t/\sigma^2 > 0$.

Example: Gaussian

For $X \sim N(\mu, \sigma^2)$, it's easy to check that

$$P(X - \mu \geq t) \leq 0.5 \exp\left(-\frac{t^2}{2\sigma^2}\right) \leq P(X - \mu \geq t - \sigma).$$

Hence, for $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$,

$$0.5 \exp\left(\frac{-n(t + \sigma/\sqrt{n})^2}{2\sigma^2}\right) \leq P(\bar{X}_n - \mu \geq t) \leq 0.5 \exp\left(-\frac{nt^2}{2\sigma^2}\right),$$

and so the Chernoff bound is tight:

$$\lim_{n\to\infty} \frac{1}{n} \ln P(\bar{X}_n - \mu \geq t) = -\frac{t^2}{2\sigma^2}.$$

Example: Bounded Support

Theorem: [Hoeffding's Inequality] For a random variable $X \in [a, b]$ with $\mathbf{E}X = \mu$ and $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

Note the resemblance to a Gaussian: $\lambda^2\sigma^2/2$ vs $\lambda^2(b-a)^2/8$. (And since P has support in $[a, b]$, $\mathrm{Var}X \leq (b-a)^2/4$.)

Example: Hoeffding's Inequality Proof

Define

$$A(\lambda) = \log\left(\mathbf{E}e^{\lambda X}\right) = \log\left(\int e^{\lambda x}\, dP(x)\right),$$

where $X \sim P$. Then A is the log normalization of the exponential family random variable X_λ with reference measure P and sufficient statistic x. Since P has bounded support, $A(\lambda) < \infty$ for all λ, and we know that

$$A'(\lambda) = \mathbf{E}(X_\lambda), \qquad\qquad A''(\lambda) = Var(X_\lambda).$$

Since P has support in $[a, b]$, $Var(X_\lambda) \leq (b-a)^2/4$. Then a Taylor expansion about $\lambda = 0$ (at this value of λ, X_λ has the same distribution as X, hence the same expectation) gives

$$A(\lambda) \leq \lambda \mathbf{E}X + \frac{\lambda^2}{2}\frac{(b-a)^2}{4}.$$

Sub-Gaussian Random Variables

Definition: X is **sub-Gaussian** with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}.$$

Note:

- Gaussian is sub-Gaussian.
- X sub-Gaussian iff $-X$ sub-Gaussian.

Sub-Gaussian Random Variables

Note:

- X sub-Gaussian implies

$$P(X - \mu \geq t) \leq \exp(-t^2/(2\sigma^2)),$$
$$P(X - \mu \leq -t) \leq \exp(-t^2/(2\sigma^2)),$$
$$P(|X - \mu| \geq t) \leq 2\exp(-t^2/(2\sigma^2)).$$

Sub-Gaussian Random Variables

Note:

- X_1, X_2 independent, sub-Gaussian with parameters σ_1^2, σ_2^2, implies $X_1 + X_2$ sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$.

Indeed, for independent X_1, X_2,

$$\begin{aligned} M_{X_1+X_2} &= \mathbf{E} \exp\left(\lambda(X_1 + X_2)\right) \\ &= \mathbf{E} \exp\left(\lambda X_1\right) \mathbf{E} \exp\left(\lambda X_2\right) \\ &= M_{X_1} M_{X_2}. \end{aligned}$$

So $\ln M_{X_1+X_2-\mu} = \ln M_{X_1-\mu_1} + \ln M_{X_2-\mu_2} \leq \lambda^2(\sigma_1^2 + \sigma_2^2)/2$.

Hoeffding Bound

Theorem: For $X_1, \dots, X_n$ independent, $\mathbf{E}X_i = \mu_i$, X_i sub-Gaussian with parameter σ_i^2, then for all $t > 0$,

$$P\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

e.g., for $\mathbf{E}X_i = 0$, $X_i \in [a, b]$, we have $\sigma_i^2 = (b-a)^2/4$ so

$$P\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Sub-Exponential Random Variables

Definition: X is **sub-exponential** with parameters (σ^2, b) if, for all $|\lambda| < 1/b$,

$$\ln M_{X-\mu}(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}.$$

Examples:

- Sub-Gaussian X with parameter σ^2 is sub-exponential with parameters (σ^2, b) for all $b > 0$.

Sub-Exponential Random Variables

Theorem: For X sub-exponential with parameters (σ^2, b),

$$P\left(X \geq \mu + t\right) \leq \begin{cases} \exp\left(-\frac{t^2}{2\sigma^2}\right) & \text{if } 0 \leq t \leq \sigma^2/b, \\ \exp\left(-\frac{t}{2b}\right) & \text{if } t > \sigma^2/b. \end{cases}$$

Sub-Exponential Random Variables

Proof: Assume $\mu = 0$. As before,

$$P(X \geq t) \leq \exp(-\lambda t)\mathbf{E}\exp(\lambda X) \\ \leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right)$$

provided $0 \leq \lambda < 1/b$. As before, we optimize the choice of λ. But now, it is constrained to $[0, 1/b)$. Without this constraint, the minimum occurs at $\lambda^* = t/\sigma^2$. So if

$$t/\sigma^2 < 1/b \iff t < \sigma^2/b,$$

we have

$$P(X \geq t) \leq \exp(-\lambda^* t + {\lambda^*}^2\sigma^2/2) = \exp(-t^2/(2\sigma^2)).$$

Sub-Exponential Random Variables

If t is larger, the minimum occurs at $\lambda = 1/b$ (since the function $t \mapsto -\lambda t + \frac{\lambda^2\sigma^2}{2}$ is monotonically decreasing in $[0, \lambda^*]$, which contains $[0, 1/b]$). Substituting this λ gives

$$P(X \geq t) \leq \exp(-t/b + \sigma^2/(2b^2)) \leq \exp(-t/(2b)),$$

where the second inequality follows from $t \geq \sigma^2/b$.

Sub-Exponential Random Variables

Examples:

- $X \sim \chi_1^2$ has

$$\mathbf{E}\exp(\lambda(X-1)) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(\lambda(z^2-1))\exp(-z^2/2)\,dz$$
$$= \frac{1}{\sqrt{1-2\lambda}}\exp(-\lambda)$$

for $|\lambda| < 1/2$. And for $|\lambda| \geq 1/2$, $M_X(\lambda)$ does not exist, so X is not sub-Gaussian.

But it is easy to check that

$$\frac{1}{\sqrt{1-2\lambda}}\exp(-\lambda) \leq \exp(2\lambda^2)$$

for $|\lambda| < 1/4$. Thus, X is sub-exponential with parameters $(4, 4)$.

Sub-Exponential Random Variables

Example: X variance σ^2, bounded: $|X-\mu| \le b$.

$$
\begin{aligned}
\mathbf{E}\exp(\lambda(X-\mu)) &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbf{E}(X-\mu)^k}{k!} \\
&\le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2}.
\end{aligned}
$$

And for $|\lambda| < 1/b$, this is no more than

$$
\mathbf{E}\exp(\lambda(X-\mu)) \le 1 + \frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right).
$$

Sub-Exponential Random Variables

So if $|\lambda| < 1/(2b)$, $1 - b|\lambda| > 1/2$ and

$$\mathbf{E}\exp(\lambda(X-\mu)) \leq \exp\left(\lambda^2\sigma^2\right).$$

Thus, X is sub-exponential with parameters $(2\sigma^2, 2b)$.

Sub-Exponential Random Variables

Theorem: [Bernstein] For X bounded as above and all $t > 0$,

$$P\left(X \geq \mu + t\right) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right).$$

Proof:

We saw above that

$$\mathbf{E}\exp(\lambda(X-\mu)) \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right)$$

for $|\lambda| < 1/b$. Setting $\lambda = t/(bt + \sigma^2) < 1/b$ gives the result.

Sub-Exponential Random Variables

Note:

- $\sigma^2 = \mathbf{E}(X-\mu)^2 \leq b^2$, so this bound is always at least as good as Hoeffding's inequality. If the variance is small ($\sigma^2 \ll b^2$), then it can be a large improvement. We'll see examples where this improvement is necessary to get optimal rates.

Sub-Exponential Random Variables

Note:

- For independent X_i, sub-exponential with parameters (σ_i^2, b_i), the sum $X = X_1 + \cdots + X_n$ is sub-exponential with parameters $\left(\sum_i \sigma_i^2, \max_i b_i\right)$.

Indeed, for $\mathbf{E}X_i = 0$,

$$\begin{aligned} M_X(\lambda) &= \prod_i \mathbf{E} \exp(\lambda X_i) \\ &\leq \prod_i \exp(\lambda^2 \sigma_i^2 / 2) = \exp\left(\lambda^2 \sum_i \sigma_i^2 / 2\right), \end{aligned}$$

where the inequality holds provided $|\lambda| < 1/b_i$ for all i.

Sub-Exponential Random Variables

Hence,

Theorem: For independent X_i, sub-exponential with parameters (σ_i^2, b_i), with mean μ_i,

$$P\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \begin{cases} \exp(-nt^2/(2\sigma^2)) & \text{for } 0 \leq t \leq \sigma^2/b, \\ \exp(-nt/(2b)) & \text{for } t > \sigma^2/b, \end{cases}$$

where $\sigma^2 = \sum_i \sigma_i^2$ and $b = \max_i b_i$.