CS281B/Stat241B. Statistical Learning Theory. Lecture 5.

Peter Bartlett

1. Concentration inequalities
   (a) Sub-exponential random variables.
   (b) Martingale methods.

Review:Chernoff technique

Theorem: For $t > 0$:

$$P(X - \mathbf{E}X \geq t) \leq \inf_{\lambda > 0} e^{-\lambda t} M_{X-\mu}(\lambda).$$

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

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

Review: Sub-Gaussian, Sub-Exponential Random Variables

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

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

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

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

Review: Sub-Exponential Random Variables

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

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

Example: X with variance σ^2, bounded ($|X - \mu| \leq b$) is sub-exponential with parameters $(2\sigma^2, 2b)$.

Sub-Exponential Random Variables

Theorem: [Bernstein] For X with variance σ^2, bounded ($|X - \mu| \leq b$), and $t > 0$,

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

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:

- $\sigma^2 = \mathbf{E}(X-\mu)^2 \leq b^2$ (and $t < b$), so this bound implies something similar to 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 $b = \max_i b_i$.

Consequences: Fast rates when variance is small

Variance bounded by expectation gives fast rates

Suppose that $\sigma^2 \leq c\mu$. Bernstein's inequality says

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

Set $t = \alpha\mu + \epsilon$.

Then $2(\sigma^2 + bt) \leq c't$, and $t^2/(c't) \geq c''\epsilon$, so

$$P\left(X \geq (1+\alpha)\mu + \epsilon\right) \leq \exp\left(-c''\epsilon\right).$$

Consequences: Fast rates when variance is small

Positive, small expectation: fast rates

For instance, if $X = \sum_{i=1}^{n} Z_i$, $Z_i > 0$, independent, then $\sigma^2 \leq b\mu$, so

$$P\left(\frac{1}{n}\sum Z_i \geq (1+\alpha)\mu + \epsilon\right) \leq \exp\left(-c''n\epsilon\right).$$

Consequences: Fast rates when variance is small

Classification with a margin condition: fast rates

Suppose $|2\eta(X) - 1| \geq c$ a.s. Recall

$$\begin{aligned} R(f) - R(f^*) &= \mathbf{E}1[f(X) \neq f^*(X)]\,|2\eta(X) - 1| \\ &\geq c\mathbf{E}1[f(X) \neq f^*(X)] \\ &= c\mathbf{E}\left(\ell(f(X), Y) - \ell(f^*(X), Y)\right)^2 \\ &\leq c\mathrm{Var}\left(\ell(f(X), Y) - \ell(f^*(X), Y)\right). \end{aligned}$$

Bernstein (for $Z_i = \ell(f(X_i), Y_i) - \ell(f^*(X_i), Y_i)$) implies

$$P\left(\hat{R}(f) - \hat{R}(f^*) \leq (1-\alpha)(R(f) - R(f^*)) - \epsilon\right) \leq \exp\left(-c'n\epsilon\right).$$

Equivalently,

$$P\left(R(f) - R(f^*) \geq \frac{1}{1-\alpha}(\hat{R}(f) - \hat{R}(f^*)) + \frac{\epsilon}{1-\alpha}\right) \leq \exp\left(-c'n\epsilon\right).$$

Consequences: Fast rates when variance is small

Then, for example, for a finite F containing f^*, if $\hat{f}$ is the minimizer of the empirical risk $\hat{R}(f)$,

$$
\begin{aligned}
&P\left(R(\hat{f}) - R(f^*) \geq \frac{\epsilon}{1-\alpha}\right) \\
&\leq P\left(R(\hat{f}) - R(f^*) \geq \frac{1}{1-\alpha}\underbrace{(\hat{R}(\hat{f}) - \hat{R}(f^*))}_{\leq 0} + \frac{\epsilon}{1-\alpha}\right) \\
&\leq P\left(\exists f,\, R(f) - R(f^*) \geq \frac{1}{1-\alpha}(\hat{R}(f) - \hat{R}(f^*)) + \frac{\epsilon}{1-\alpha}\right) \\
&\leq |F| \exp\left(-c' n\epsilon\right).
\end{aligned}
$$

And this is no more than δ for $\epsilon = c'' \dfrac{\log(|F|/\delta)}{n}$.

Consequences: Fast rates when variance is small

Convex regression with a strongly convex loss: fast rates

Consider $\ell(\hat{y}, y) = (\hat{y} - y)^2$. Define $f^* = \arg\min_{f \in F} R(f)$, where F is convex.

$$
\begin{aligned}
& R(f) - R(f^*) \\
& = \mathbf{E}\left((Y - f(X))^2 - (Y - f^*(X))^2\right) \\
& = \mathbf{E}\left((Y - f^*(X) + f^*(X) - f(X))^2 - (Y - f^*(X))^2\right) \\
& = \mathbf{E}\left(\underbrace{2(Y - f^*(X))(f^*(X) - f(X))}_{\geq 0} + (f^*(X) - f(X))^2\right) \\
& \geq \mathbf{E}\left((f^*(X) - f(X))^2\right).
\end{aligned}
$$

Consequences: Fast rates when variance is small

Also, for $|Y|, |f(X)| \leq b$,

$$\begin{aligned} &\mathbf{E}\left((Y - f(X))^2 - (Y - f^*(X))^2\right)^2 \\ &= \mathbf{E}\left(2(Y - f^*(X))(f^*(X) - f(X)) + (f^*(X) - f(X))^2\right)^2 \\ &\leq (6b)^2 \mathbf{E}\left(f^*(X) - f(X)\right)^2 \\ &\leq (6b)^2 (R(f) - R(f^*)). \end{aligned}$$

Again, variance is bounded in terms of expectation. As above, we get fast rates.

Concentration Bounds for Martingale Difference Sequences

Next, we're going to consider concentration of martingale difference sequences. The application is to understand how tails of $f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)$ behave, for some function f.

If we write

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)$$
$$= \sum_{i=1}^{n} \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i] - \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}],$$

then we have represented this deviation as a *martingale difference sequence* (the Doob martingale). We get concentration because of the many (n) independent contributions.

Martingales

Definition: A sequence Y_n of random variables adapted to a filtration $\mathcal{F}_n$ is a **martingale** if, for all n,

$$\mathbf{E}|Y_n| < \infty$$
$$\mathbf{E}[Y_{n+1}|\mathcal{F}_n] = Y_n.$$

$\mathcal{F}_n$ is a **filtration** means these σ-fields are nested: $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.

Y_n is **adapted to** $\mathcal{F}_n$ means that each Y_n is measurable with respect to $\mathcal{F}_n$.

e.g. $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$, the σ-field generated by the first n variables. Then we say Y_n is a martingale sequence.

e.g. $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Then Y_n is a martingale sequence wrt X_n.

Martingale Difference Sequences

Definition: A sequence D_n of random variables adapted to a filtration $\mathcal{F}_n$ is a **martingale difference sequence** if, for all n,

$$\mathbf{E}|D_n| < \infty$$

$$\mathbf{E}[D_{n+1}|\mathcal{F}_n] = 0.$$

e.g., $D_n = Y_n - Y_{n-1}$.

$$\begin{aligned}\mathbf{E}[D_{n+1}|\mathcal{F}_n] &= \mathbf{E}[Y_{n+1}|\mathcal{F}_n] - \mathbf{E}[Y_n|\mathcal{F}_n] \\ &= \mathbf{E}[Y_{n+1}|\mathcal{F}_n] - Y_n = 0\end{aligned}$$

(because Y_n is measurable wrt $\mathcal{F}_n$, and because of the martingale property).
Hence, $Y_n - Y_0 = \sum_{i=1}^n D_i$.