CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 15**

Concentration Inequalities

Lecturer: Peter Bartlett *Scribe: Lara Stoll*

1 Concentration Inequalities

Recall that last time, we reviewed both Markov's inequality, a bound on tails of nonnegative random variables, and Chebyshev's inequality, which is just a specific application of Markov's. Both of these are, in general, loose upper bounds.

Consider the sum $S_n = \sum_{i=1}^n X_i$ of independent random variables, X_1, ..., X_n. Then, we have, for $s > 0$,

$$\begin{aligned} P\{S_n - \mathbb{E}S_n \geq t\} &= P\{\exp(s(S_n - \mathbb{E}S_n) \geq \mathrm{e}^{st}\} \\ &\leq \mathrm{e}^{-st}\mathbb{E}\exp(s(S_n - \mathbb{E}S_n)) \\ &= \mathrm{e}^{-st}\prod_{i=1}^n \mathbb{E}\,\mathrm{e}^{s(X_i - \mathbb{E}X_i)} \qquad (1) \end{aligned}$$

where the inequality is true through application of Markov's inequality, and the second equality follows because of the independence of X_i. Note that $\mathbb{E}\,\mathrm{e}^{s(X_i - \mathbb{E}X_i)}$ is the moment-generating function of X_i.

Lemma 1.1. (Hoeffding) For a random variable X with $\mathbb{E}X = 0$ and $a \leq X \leq b$ then for $s > 0$,

$$\mathbb{E}\,\mathrm{e}^{sX} \leq \mathrm{e}^{s^2(b-a)^2/8}$$

Note that this exponential has the same length of interval that the random variable X does.

Theorem 1.2. Hoeffding's Inequality

For bounded random variables, $X_i \in [a_i, b_i]$, where $X_1, ..., X_n$ are independent, then

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

and

$$P(S_n - \mathbb{E}S_n \leq -t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. By combining (1) and Lemma 1.1, we have

$$\begin{aligned} P(S_n - \mathbb{E}S_n \geq t) &\leq \inf_{s>0} \mathrm{e}^{-st} \prod_{i=1}^{n} \exp\left(\frac{s^2(b_i-a_i)^2}{8}\right) \\ &= \inf_{s>0} \exp\left(-st + \frac{s^2 \sum_{i=1}^{n}(b_i-a_i)^2}{8}\right) \end{aligned} \tag{2}$$

Minimize (2) by setting $s = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$ to obtain

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

□

Example. Consider random variables $X_i \in [0,1]$ with $\mathbb{E}[X_i] = p$. Then,

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \varepsilon\right\} \leq \exp\left(\frac{-2(\varepsilon n)^2}{n}\right) = \exp(-2\varepsilon^2 n) \tag{3}$$

For this binomial distribution, consider then, that for all f, $P(\hat{L}(f) - L(f) \geq \varepsilon) \leq e^{-2\varepsilon^2 n}$, and by symmetry we have,

$$P\left(|\hat{L}(f) - L(f)| \geq \varepsilon\right) \leq 2\mathrm{e}^{-2\varepsilon^2 n} \tag{4}$$

Then, for some fixed probability threshold δ, what do deviations of the random variable look like? (4) is $\leq \delta$ for $2\varepsilon^2 n \geq \ln\frac{2}{\delta}$, or $\varepsilon \geq \sqrt{\frac{2\ln\frac{2}{\delta}}{n}}$. This shows that sums of random variables are concentrated. A similar conclusion can be made when we consider a large class of functions of random variables.

Theorem 1.3. Bounded Differences (or Hoeffding/Azuma or McDiarmid's) Inequality

Consider independent random variables $X_1, ..., X_n \in \mathcal{X}$ and a mapping $f : \mathcal{X}^n \to \mathbb{R}$. If, $\forall i \in \{1, ..., n\}$, $\forall x_1, ..., x_n,\ x_i' \in \mathcal{X}$, f satisfies

$$|f(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n)| \leq c_i \tag{5}$$

then,

$$P(f(x_1, ..., x_n) - \mathbb{E}f \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

and

$$P(f(x_1, ..., x_n) - \mathbb{E}f \leq -t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

That is, changing any one of the values of x_i can only change f by a bounded amount, and then we get a concentration of f around its expectation. Clearly, Hoeffding's Inequality is just an example of the Bounded Differences Inequality, where f is a sum.

Proof. Define the random variable $V_i = \mathbb{E}[f|X_1, ..., X_i] - \mathbb{E}[f|X_1, ..., X_{i-1}]$ for $i = 1, ..., n$. Then,

$\sum_{i=1}^n V_i = f - \mathbb{E}f$. Also, $\mathbb{E}[V_i|X_1,...,X_{i-1}] = 0$. We have

$$\begin{aligned}
P(f - \mathbb{E}f \geq t) &= P(\sum_{i=1}^n V_i \geq t) \\
&\leq \inf_{s>0} \mathrm{e}^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^n V_i\right)\right] \\
&= \inf_{s>0} \mathrm{e}^{-st}\mathbb{E}\,\mathbb{E}\left[\exp\left(s\sum_{i=1}^n V_i\right)|X_1,...,X_{n-1}\right] \\
&= \inf_{s>0} \mathrm{e}^{-st}\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n-1} V_i\right)\mathbb{E}\left[\mathrm{e}^{sV_n}|X_1,...,X_{n-1}\right]\right]
\end{aligned} \tag{6}$$

where the last step is justified because $V_1, \ldots, V_{n-1}$ are functions only of $X_1, \ldots, X_{n-1}$. Now define

$$\begin{aligned}
L_i &= \inf_x \mathbb{E}\left[f|X_1,...,X_{i-1},x\right] - \mathbb{E}\left[f|X_1,...,X_{i-1}\right] \\
U_i &= \sup_x \mathbb{E}\left[f|X_1,...,X_{i-1},x\right] - \mathbb{E}\left[f|X_1,...,X_{i-1}\right]
\end{aligned}$$

So, we have almost surely $L_i \leq V_i \leq U_i$ and

$$\begin{aligned}
U_i - L_i &= \sup_x \mathbb{E}\left[f|X_1,...,X_{i-1},x\right] - \inf_y \mathbb{E}\left[f|X_1,...,X_{i-1},y\right] \\
&= \sup_{x,y} \int \left(f(x_1,...,x_{i-1},x,x_{i+1},...,x_n) - f(x_1,...,x_{i-1},y,x_{i+1},...,x_n)\right) dP(x_{i+1},...,x_n) \\
&\leq c_i
\end{aligned}$$

where we use the independence of X_i to obtain the second equality. Applying lemma 1.1 to (6) gives

$$P\left(f - \mathbb{E}f \geq t\right) \leq \inf_{s>0} e^{-st}\,\mathbb{E}\left[\exp\left(s\sum_{i=1}^{n-1} V_i\right)\exp\left(\frac{s^2 c_i^2}{8}\right)\right]$$

Repeating the argument $n-1$ more times, we get

$$P\left(f - \mathbb{E}f \geq t\right) \leq \inf_{s>0} \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n c_i^2\right) \tag{7}$$

Finally, choose the optimal value of $s = 4t/\sum_{i=1}^n c_i^2$ and apply to (7) to obtain

$$P\left(f(x_1,...,x_n) - \mathbb{E}f \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Replacing f with $-f$ yields

$$P\left(f(x_1,...,x_n) - \mathbb{E}f \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right)$$

$\square$