Chernoff, Hoeffding's and Bennett's Inequalities

Jimmy Jin, James Wilson and Andrew Nobel
UNC-Chapel Hill

Last updated: 1/8/14

1 Hoeffding's inequality

Hoeffding's inequality is one of the most important inequalities in the machine learning literature. It gives a **exponential** bound for sums of random variables where the increments are bounded. Before we state this theorem, we state and prove a related and simple exponential bound known as the Chernoff bound. You should ask yourself: why is it good that the bound is exponential?

Theorem 1.1. (Chernoff MGF bound)

For any r.v. X and $s > 0$,

$$P(X > t) \leq \inf_{s>0} \left[e^{-st} E(e^{sx}) \right]$$

Proof. Apply Markov's to $P(e^{sx} > e^{st})$ and then take infimum over s. □

Theorem 1.2. (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent r.v.'s with $a_i \leq X_i \leq b_i$ a.s. Then $\forall\, t > 0$,

$$P(S_n - ES_n \geq t) \leq \exp\left\{ \frac{-2t^2}{\sum (b_i - a_i)^2} \right\}$$

$$P(S_n - ES_n \leq -t) \leq \exp\left\{ \frac{-2t^2}{\sum (b_i - a_i)^2} \right\}$$

And the two-sided bound:

$$P(\left|S_n - ES_n\right| \geq t) \leq 2 \exp\left\{ \frac{-2t^2}{\sum (b_i - a_i)^2} \right\}$$

Not only is the inequality of great use, but its proof also draws on many of the concepts of the previous sections. The main idea of the proof is actually quite simple. First, we bound the MGFs of the increments using Taylor expansion. Then we plug these bounds into a Chernoff bound for the overall sum.

Lemma 1.3. (MGF bound for Hoeffding)

Let X be an r.v. with $EX = 0$ and $a \le X \le b$. Then $\forall s > 0$,

$$E(e^{sX}) \le \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$$

Proof. Fix some $s > 0$ and consider the function $f(x) = e^{sx}$. Then for every $x \in [a, b]$, Jensen's inequality gives us:

$$e^{sx} \le \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

Taking expectations (note $EX = 0$) and defining $p = -a/(b-a)$, we have:

$$Ee^{sX} \le \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = \left[1 - p + pe^{s(b-a)}\right]e^{-ps(b-a)} = e^{\phi(u)}$$

Where $u = s(b-a)$ and $\phi(u) = -pu + \log(1-p+pe^u)$. Therefore it is sufficient to show that $\phi(u) \le s^2(b-a)^2/8$.

Since ϕ is sufficiently smooth, the 1st order Taylor expansion about $u = 0$ with Lagrange remainder is:

$$\phi(u) = \phi(0) + \phi'(0) \cdot u + \frac{u^2}{2}\phi''(c), \quad \text{some } c \in [0, u]$$

Now some calculate show that

$$\begin{aligned}
\phi'(u) &= -p + \frac{p}{p + (1-p)e^{-u}} & \Rightarrow \quad & \phi'(0) = 0 \\
\phi''(u) &= \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} & = \quad & \frac{\alpha\beta}{(\alpha+\beta)^2} \le \frac{1}{4}
\end{aligned}$$

Plugging this into the Taylor expansion above shows that

$$\phi(u) \le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

□

Now the proof of the main result:

Proof. Applying Markov's to $P(S_n - ES_n \ge t) = P(e^{s(S_n - ES_n)}) \ge e^{st})$, we have:

$$\begin{aligned}
P(S_n - ES_n \ge t) &\le e^{-st}E\left\{\exp\left(s\sum_{i=1}^n (X_i - EX_i)\right)\right\} \\
&= e^{-st}E\left\{\prod_{i=1}^n e^{s(X_i - EX_i)}\right\} \\
&= e^{-st}\prod_{i=1}^n E\Big\{e^{s(X_i - EX_i)}\Big\} \quad \text{(by independence)}
\end{aligned}$$

Now applying our lemma to the RHS, we have

$$\begin{aligned} P(S_n - ES_n \geq t) &\leq e^{-st} \cdot \prod_{i=1}^{n} \exp\left\{\frac{s^2(b_i - a_i)^2}{8}\right\} \\ &= e^{-st} \cdot \exp\left\{\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right\} \end{aligned}$$

Choosing $s = 4t/\sum(b_i - a_i)^2$ completes the proof.

□

1.1 Bennett's Inequality

Like Hoeffding's inequality, both Bennett's and Bernstein's inequality give exponential bounds for a sum of random variables whose increments are bounded. However, the key difference is that Bennett's and Bernstein's inequality take into account the **variances** of the increments. Here, we discuss Bennett's inequality but recommend the interested reader Bernstein's inequality.

Theorem 1.4. (Bennett's inequality) Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \leq c$ for all i. Then for $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left\{\frac{-n\sigma^2}{c^2} \cdot h\left(\frac{ct}{n\sigma^2}\right)\right\}$$

where $h(u) = (1+u)\log(1+u) - u$, for $u \geq 0$.

The proof of this inequality is actually quite similar to the proof of Hoeffding's inequality. Again we first bound the MGFs of the increments using Taylor expansion, and plug those into a Chernoff-type bound for the overall sum.

Lemma 1.5. (MGF bound for Bennett)

Let X be an r.v. with $\mathbb{E}X = 0, \mathbb{E}X^2 = \sigma^2$, and $|X| \leq c$. Then:

$$\mathbb{E}(e^{sX}) \leq \exp\left\{\frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\right\}, \quad \text{for all } s > 0$$

Proof. By Taylor series for e^x we have for any $s > 0$,

$$\mathbb{E}(e^{sX}) = \mathbb{E}\left\{1 + sX + \sum_{r=2}^{\infty} \frac{s^r X^r}{r!}\right\} = 1 + \sum_{r=2}^{\infty} \frac{s^r \mathbb{E}(X^r)}{r!}$$

Now by Holder's inequality,

$$\mathbb{E}X^r \;\leq\; \mathbb{E}|X|^r \;\leq\; \mathbb{E}|X|^2|X|^{r-2} \;\leq\; \sigma^2 \cdot c^{r-2}$$

Therefore plugging this into our Taylor expansion and summing, we obtain:

$$\begin{aligned}
\mathbb{E}(e^{sX}) &\le 1 + \sum_{r=2}^{\infty} \frac{s^r \sigma^2 c^{r-2}}{r!} \\
&= 1 + \frac{\sigma^2}{c^2} \sum_{r=2}^{\infty} \frac{(sc)^r}{r!} \\
&= 1 + \frac{\sigma^2}{c^2} (e^{sc} - 1 - sc)
\end{aligned}$$

Then apply the identity $1 + x \le e^x$.

□

Now we plug into a basic Chernoff-type bound to prove the main result:

Proof. (of Bennett's inequality)

Recall that we have $X_1, \dots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \le c$ for all i.

Define $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i^2$. Then by the basic Chernoff bound for $s > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t \right) \le e^{-st} \prod_{i=1}^n \mathbb{E} e^{sX_i} \le \exp\left\{ \frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st \right\}$$

Now this bound holds for all $s > 0$, so we optimize. To ease notation define the function:

$$f(s) = \frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st$$

Some high school calculus shows this to be minimized at:

$$s_0 = c^{-1} \log(1 + tc/n\sigma^2)$$

and substituting into the above expression gives the result.

□