CS229 hoeffding inequality notes

James Chuang

April 28, 2017

Contents

My notes on John Duchi's CS229 supplemental notes on Hoeffding's inequality.

basic probability bounds

- a basic question in probability, statistics, and machine learning:
  - given a random variable Z with expectation $\mathbb{E}[Z]$, how likely is Z to be close to its expectation?
  - more precisely, how close is it likely to be?
  - therefore, we would like to compute bounds of the following form for $t \geq 0$

$$P(Z \geq \mathbb{E}[Z] + t) \text{ and } P(Z \leq \mathbb{E}[Z] - t)$$

- **Markov's inequality**
  - Let $Z \geq 0$ be a non-negative random variable. Then for all $t \geq 0$,

$$P(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}$$

  - i.e., Markov's inequality puts a bound on the probability that a random variable is greater than a non-negative value t
  - Proof:
    - note: $P(Z \geq t) = \mathbb{E}\left[\mathbf{1}\left\{Z \geq t\right\}\right]$
      - consider the two possible cases for Z:
        - if $Z \geq t$, then $\mathbf{1}\{Z \geq t\} = 1$:

$$\begin{aligned} Z &\geq t \\ \frac{Z}{t} &\geq 1 \\ \frac{Z}{t} &\geq \mathbf{1}\{Z \geq t\} \end{aligned}$$

        - if $Z < t$, then $\mathbf{1}\{Z \geq t\} = 0$:

$$\begin{aligned} \frac{Z}{t} &\geq 0 && \text{Z and t both } > 0 \\ \frac{Z}{t} &\geq \mathbf{1}\{Z \geq t\} \end{aligned}$$

        - so in general, $\frac{Z}{t} \geq \mathbf{1}\{Z \geq t\}$
    - thus:

$$\begin{aligned} P(Z \geq t) &= \mathbb{E}\left[\mathbf{1}\left\{Z \geq t\right\}\right] \\ P(Z \geq t) &\leq \mathbb{E}\left[\frac{Z}{t}\right] \\ P(Z \geq t) &\leq \frac{\mathbb{E}[Z]}{t} \end{aligned}$$

- **note:** this is the proof given in the notes, but this proof from Wolfram Alpha makes more sense to me
- essentially all other bounds on probabilities are variations on Markov's inequality
  - the first variation uses second moments – the variance – of a random variable rather than simply its mean, and is known as Chebyshev's inequality
- **Chebyshev's inequality**
  - Let Z be any random variable with Var$(Z) < \infty$. Then, for $t \geq 0$,

$$P\left((Z - \mathbb{E}[Z])^2 \geq t^2\right) \leq \frac{\text{Var}(Z)}{t^2}$$

or equivalently,

$$P\left(|Z - \mathbb{E}[Z]| \geq t\right) \leq \frac{\text{Var}(Z)}{t^2}$$

  - i.e., Chebyshev's inequality puts a bound on the probability that a random variable is greater than t away from its expected value $\mathbb{E}[Z]$
  - Proof:

$$\begin{aligned}
&P\left((Z - \mathbb{E}[Z])^2 \geq t^2\right) \\
&\leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{t^2} && \text{by Markov's inequality} \\
&\leq \frac{\text{Var}(Z)}{t^2}
\end{aligned}$$

  - a nice consequence of Chebyshev's inequality:
    - averages of random variables with finite variance converge to their mean (this is the **weak law of large numbers**)
    - an example:
      - suppose Z_i are i.i.d. with finite variance and $\mathbb{E}[Z_i] = 0$
      - define $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$
      - then:

$$\begin{aligned}
&\text{Var}\left(\bar{Z}\right) \\
&= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) && \text{def. } \bar{Z} \\
&= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Z_i\right) && \text{property of variance} \\
&= \frac{n\text{Var}\left(Z_1\right)}{n^2} && \text{Var}\left(Z_i\right) \text{ are all equal, since } Z_i \text{ are i.i.d.} \\
&= \frac{\text{Var}\left(Z_1\right)}{n}
\end{aligned}$$

      - in particular, for any $t \geq 0$ (remember $\mathbb{E}[Z_i] = 0$):

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq t\right) \leq \frac{\text{Var}\left(Z_1\right)}{nt^2} \qquad \text{Chebyshev's inequality}$$

      - so, $P\left(\left|\bar{Z}\right| \geq t\right) \to 0$ for any $t > 0$

moment generating functions

- often, we want sharper – even exponential – bounds on the probability that a random variable exceeds its expectation by much
  - to accomplish this, we need a stronger condition than finite variance
  - **moment generating functions** are natural candidates for this condition:
    - for a random variable Z, the moment generating function of Z is the function:

$$M_Z(\lambda) := \mathbb{E}\left[\exp(\lambda Z)\right]$$

- the moment generating function may be infinite for some λ

Chernoff bounds

- **Chernoff bounds** use moment generating functions to give exponential deviation bounds
  - Let Z be any random variable
  - then, for any $t \geq 0$

$$\begin{aligned} P\left(Z \geq \mathbb{E}[Z] + t\right) &\leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] e^{-\lambda t} \\ &\leq \min_{\lambda \geq 0} M_{Z-\mathbb{E}[Z]}\left(\lambda\right) e^{-\lambda t} \end{aligned}$$

and

$$\begin{aligned} P\left(Z \leq \mathbb{E}[Z] - t\right) &\leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(\mathbb{E}[Z] - Z)}\right] e^{-\lambda t} \\ &\leq \min_{\lambda \geq 0} M_{\mathbb{E}[Z]-Z}\left(\lambda\right) e^{-\lambda t} \end{aligned}$$

  - proof of the first inequality (the second inequality is identical):
    - for any $\lambda > 0$:
      - $Z \geq \mathbb{E}[Z] + t$ iff $e^{\lambda Z} \geq e^{\lambda \mathbb{E}[Z] + \lambda t}$, i.e. $e^{\lambda(Z - \mathbb{E}[Z])} \geq e^{\lambda t}$
      - thus,

$$\begin{aligned} P\left(Z - \mathbb{E}[Z] \geq t\right) &= P\left(e^{\lambda(Z-\mathbb{E}[Z])} \geq e^{\lambda t}\right) \\ &\leq \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right] e^{-\lambda t} && \text{by Markov's inequality} \end{aligned}$$

      - since the choice of $\lambda > 0$ did not matter, we can take the best one by minimizing the right side of the bound w.r.t. λ
        - note that the bound still holds at $\lambda = 0$
  - the important result: **Chernoff bounds "play nicely" with summations**
    - this is a consequence of the moment generating function
    - assume that Z_i are independent, then:

$$\begin{aligned} &M_{Z_1 + \cdots + Z_n}(\lambda) \\ &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} Z_i\right)\right] && \text{def. MGF} \\ &= \mathbb{E}\left[\prod_{i=1}^{n} \exp\left(\lambda Z_i\right)\right] && \text{just exponent properties} \\ &= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] && Z_i \text{ are independent} \\ &= \prod_{i=1}^{n} M_{Z_i}(\lambda) \end{aligned}$$

    - this means that **when we calculate a Chernoff bound of a sum of i.i.d. variables, we only need to calculate the moment generating function for one of them**:
      - suppose Z_i are i.i.d. and (for simplicity) mean zero. Then:

$$
\begin{aligned}
P\left(\sum_{i=1}^{n} Z_i \geq t\right) &\leq \mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} Z_i\right)}\right] e^{-\lambda t} && \text{Chernoff bound with } \mathbb{E}\left[Z_i\right]=0 \\
&\leq M_{\sum_{i=1}^{n} Z_i}(\lambda) e^{-\lambda t} && \text{rewrite in terms of MGF} \\
&\leq \prod_{i=1}^{n} M_{Z_i}(\lambda) e^{-\lambda t} && \text{the rule derived above} \\
&\leq \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda Z_i\right)\right] e^{-\lambda t} && \text{def. MGF} \\
&\leq \left(\mathbb{E}\left[e^{\lambda Z_1}\right]\right)^n e^{-\lambda t} && Z_i \text{ are i.i.d.}
\end{aligned}
$$

moment generating function examples

- now we give several examples of moment generating functions, which enable us to give a few nice deviation inequalities as a result
- for all of our examples, we will have very convenient bounds of the form

$$
M_Z(\lambda) = \mathbb{E}\left[e^{\lambda Z}\right] \leq \left(\frac{C^2\lambda^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}
$$

- ,for some $C \in \mathbb{R}$ (which depends on the distribution of Z)
  - this form is 'nice' for applying Chernoff bounds
- begin with the classical normal distribution, $Z \sim \mathcal{N}\left(0, \sigma^2\right)$. Then,

$$
\mathbb{E}\left[\exp\left(\lambda Z\right)\right] = \exp\left(\frac{\lambda^2\sigma^2}{2}\right)
$$

- **DERIVATION GOES HERE**
- a second example is the **Rademacher random variable**, aka the random sign variable:
  - let $S = 1$ with probability $\frac{1}{2}$ and $S = -1$ with probability $\frac{1}{2}$:

$$
\mathbb{E}\left[e^{\lambda S}\right] \leq \left(\frac{\lambda^2}{2}\right)
$$

  - derivation:

$$
\begin{aligned}
\mathbb{E}\left[e^{\lambda S}\right] &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(\lambda S)^k}{k!}\right] && \text{Taylor expansion: } e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}\left[S^k\right]}{k!} \\
&= \sum_{k=0,2,4,\ldots} \frac{\lambda^k}{k!} && \text{for } k \text{ odd, } \mathbb{E}\left[S^k\right] = 0; \text{ for } k \text{ even, } \mathbb{E}\left[S^k\right] = 1 \\
&= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\
&\leq \sum_{k=0}^{\infty} \frac{\left(\lambda^2\right)^k}{2^k \cdot k!} && (2k)! \geq 2k \cdot k! \text{ for all } k = 0, 1, 2, \ldots \\
&\leq \sum_{k=0}^{\infty} \left(\frac{\lambda^2}{2}\right)^k \frac{1}{k!} \\
&\leq \exp\left(\frac{\lambda^2}{2}\right) && \text{Taylor expansion}
\end{aligned}
$$

- we can apply this inequality in a Chernoff bound to see how large a sum of i.i.d. random signs is likely to be:
- $Z = \sum_{i=1}^{n} S_i$, where $S_i \in \{\pm 1\}$, so $\mathbb{E}[Z] = 0$

$$
\begin{aligned}
P(Z > t) &\leq \mathbb{E}\left[e^{\lambda Z}\right] e^{\lambda t} \\
&\leq \mathbb{E}\left[e^{\lambda S_1}\right]^n e^{-\lambda t} \\
&\leq \exp\left(\frac{n\lambda^2}{2}\right) e^{-\lambda t}
\end{aligned}
$$

minimize w.r.t. λ :

$$
\begin{aligned}
&\frac{\partial}{\partial \lambda}\left(\frac{n\lambda^2}{2} - \lambda t\right) \\
&= n\lambda - t = 0 \\
\lambda &= \frac{t}{n}
\end{aligned}
$$

$$
P(Z \geq t) \leq \exp\left(-\frac{t^2}{2n}\right)
$$

$$
P\left(\sum_{i=1}^{n} S_i \geq \sqrt{2n \log \frac{1}{\delta}}\right) \leq \delta \qquad \text{let } t = \sqrt{2n \log \frac{1}{\delta}}
$$

- so, $Z = \sum_{i=1}^{n} S_i = O\left(\sqrt{n}\right)$ with extremely high probability- the sum of n independent random signs is essentially never larger than $O(\sqrt{n})$

Hoeffding's lemma and Hoeffding's inequality

- **Hoeffding's inequality**: a powerful technique for bounding the probability that sums of bounded random variables are too large or too small
  - perhaps the most important inequality in learning theory
  - Let $Z_1, \ldots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, where $-\infty < a < b < \infty$. Then:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\geq t\right)\leq\exp\left(-\frac{2nt^2}{\left(b-a\right)^2}\right)$$

and

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\leq -t\right)\leq\exp\left(-\frac{2nt^2}{\left(b-a\right)^2}\right)$$

for all $t\geq 0$

- proof of Hoeffding's inequality using Chernoff bounds and Hoeffding's lemma:
  - **Hoeffding's lemma**:
    - Let Z be a bounded random variable with $Z\in[a,b]$. Then,

      $$\mathbb{E}\left[\exp\left(\lambda\left(Z-\mathbb{E}[Z]\right)\right)\right]\leq\left(\frac{\lambda^2\left(b-a\right)^2}{8}\right)\quad\text{for all }\lambda\in\mathbb{R}$$

    - A proof of a slightly weaker version of this lemma with a factor of 2 instead of 8 using the random sign moment generating bound and **Jensen's inequality**
      - Jensen's inequality states: if $f:\mathbb{R}\to\mathbb{R}$ is a *convex* function, then:

        $$f\left(\mathbb{E}[Z]\right)\leq\mathbb{E}\left[f(Z)\right]$$

      - to remember this inequality:
        - think of $f(t)=t^2$
        - note that if $\mathbb{E}[Z]=0$, then $f\left(\mathbb{E}[Z]\right)=0$, while we generally have $\mathbb{E}[Z^2]>0$
      - we will use a technique in probability theory known as **symmetrization** (this is a common technique in probability theory, machine learning, and statistics):
        - Let Z' be an independent copy of Z with the same distribution, so that $Z'\in[a,b]$ and $\mathbb{E}[Z']=\mathbb{E}[Z]$, but Z and Z' are independent. Then:

          $$\begin{aligned}&\mathbb{E}_Z\left[\exp\left(\lambda\left(Z-\mathbb{E}_Z[Z]\right)\right)\right]\\&=\mathbb{E}_Z\left[\exp\left(\lambda\left(Z-\mathbb{E}_{Z'}[Z']\right)\right)\right]&&\text{$\mathbb{E}_Z$ and $\mathbb{E}_{Z'}$ indicate expectations w.r.t. Z and Z'}\\&\leq\mathbb{E}_Z\left[\mathbb{E}_{Z'}\exp\left(\lambda\left(Z-Z'\right)\right)\right]&&\text{Jensen's inequality applied to } f(x)=e^{-x}\end{aligned}$$

        - so, we have:

          $$\mathbb{E}[\exp\left(\lambda\left(Z-\mathbb{E}[Z]\right)\right)]\leq\mathbb{E}\left[\exp\left(\lambda\left(Z-Z'\right)\right)\right]$$

        - now, we note the following: the difference $Z-Z'$ is symmetric about zero, so that if $S\in\{-1,1\}$ is a random sign variable, then $S(Z-Z')$ has exactly the same distribution as $Z-Z'$

          $$\begin{aligned}\mathbb{E}_{Z,Z'}\left[\exp\left(\lambda\left(Z-Z'\right)\right)\right]&=\mathbb{E}_{Z,Z',S}\left[\exp\left(\lambda S\left(Z-Z'\right)\right)\right]\\&=\mathbb{E}_{Z,Z'}\left[\mathbb{E}_S\left[\exp\left(\lambda S\left(Z-Z'\right)\right)\right]\mid Z,Z'\right]\end{aligned}$$

        - now, use inequality (3) from the notes (i.e., $\mathbb{E}\left[e^{\lambda S}\right]\leq\exp\left(\frac{\lambda^2}{2}\right)\ \forall\ \in\mathbb{R}$):

          $$\mathbb{E}_S\left[\exp\left(\lambda S\left(Z-Z'\right)\right)\mid Z,Z'\right]\leq\exp\left(\frac{\lambda^2\left(Z-Z'\right)^2}{2}\right)$$

        - by assumption, we have $|Z-Z'|\leq(b-a)$, so $(Z-Z')\leq(b-a)^2$, giving:

          $$\mathbb{E}_{Z,Z'}\left[\exp\left(\lambda\left(Z-Z'\right)\right)\right]\leq\exp\left(\frac{\lambda^2\left(b-a\right)^2}{2}\right)$$

- this is the result, with a factor of 2 instead of 8

- now, use Hoeffding's lemma with the Chernoff bound to prove Hoeffding's inequality:

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\geq t\right) &= P\left(\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\geq nt\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\right)\right]e^{-\lambda nt} \quad \text{Chernoff bound} \\
&\leq \left(\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(Z_i-\mathbb{E}[Z_i])}\right]\right)e^{-\lambda nt} \\
&\leq \left(\prod_{i=1}^{n}e^{\frac{\lambda^2(b-a)^2}{8}}\right)e^{-\lambda nt} \quad \text{Hoeffding's lemma}
\end{aligned}
$$

- rewriting and minimizing over $\lambda \geq 0$:

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\mathbb{E}\left[Z_i\right]\right)\geq t\right) &\leq \min_{\lambda\geq 0}\exp\left(\frac{n\lambda^2\left(b-a\right)^2}{8}-\lambda nt\right) \\
&\leq \exp\left(-\frac{2nt^2}{\left(b-a\right)^2}\right)
\end{aligned}
$$