Bounded, Subgaussian and Subexponential r.v.s

High Dim Probability & Linear Algebra for ML and Sig Proc

Namrata Vaswani

Iowa State University

Chapter 2 of book (Vershynin's book)

For a non-negative r.v. Z,

$$\Pr(Z > s) \leq \frac{\mathbb{E}[Z]}{s}$$

Proof: easy application of integral identity

$$\mathbb{E}[Z] \geq \int_0^s \Pr(Z > \tau) d\tau \geq \Pr(Z > s)(\int_0^s d\tau) = \Pr(Z > s)s$$

Applications: basic ideas

1. Apply this to $Z = |X - \mu|$ with $\mu = \mathbb{E}[X]$, to get Chebyshev's inequality.
2. Apply this to $Z = e^{tX}$ for any $t \geq 0$. notice e^{tX} is always non-negative.

$$\Pr(X > s) = \Pr(e^{tX} > e^{ts}) \leq e^{-ts}\mathbb{E}[e^{tX}] = e^{-ts} M_X(t)$$

   Since this bound holds for all $t \geq 0$, we can take a $\min_{t \geq 0}$ of the RHS or we can substitute in any convenient value of t.
3. To get a bound for $\Pr(X < -s)$, use $Z = e^{-tX}$ for $t \geq 0$.

4. Useful for sums of independent r.v.s: if $S = \sum_{i=1}^m X_i$ with X_i's independent, then $M_X(\lambda) = \prod_i M_{X_i}(\lambda)$. So then we get

$$\Pr(\sum_i X_i > s) \leq \min_{\lambda \geq 0} e^{-\lambda s} M_{\sum_i X_i}(\lambda) = \min_{\lambda \geq 0} e^{-\lambda s} \prod_i \mathbb{E}[e^{\lambda X_i}]$$

5. Use exact expression for MGF or a bound on MGFs (e.g. Hoeffding's lemma bounds the MGF of any bounded r.v.)
6. Followed by often using $1 + x \leq e^x$ or using $cosh(x) \leq e^{x^2/2}$ (or other bounds) to simplify things. Basic point is to try to get a summation over i in the exponent.
7. Final step: either minimizer over $\lambda \geq 0$ by differentiating the expression or a pick a convenient value of $\lambda \geq 0$ to substitute.
8. Similar approach to bound $\Pr(\sum_i X_i < -s)$. Combine both to bound $\Pr(|\sum_i X_i| > s)$.
9. *disregard this in first read:* Final final step that is used sometimes: suppose get a bound $g(s)$ but want to show $g(s) \leq f(s)$ for some simpler expression $f(s)$: try to show that $g(s) - f(s)$ is a decreasing function for the desired range of s values with $g(0) - f(0) = 0$ or something similar: this is used in Chernoff inequality for $Bern(p_i)$ r.v.s. for small s setting.

Chebyshev's inequality I

Given n independent r.v.s X_i with variance $\sigma^2 < \infty$. Then,

$$\Pr(|\sum_i (X_i - \mathbb{E}[X_i])| > t) \leq n\sigma^2/t^2$$

Proof:

- apply Markov's inequality to $|\sum_i (X_i - \mathbb{E}[X_i])|^2$, and then use independence to argue that $\mathbb{E}[|\sum_i (X_i - \mathbb{E}[X_i])|^2] = \sum_i \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = n\sigma^2$.

Notice that this does not make any assumption on the distribution of the r.v.s, does not require bounded-ness or sub-Gaussianity or sub-expo. Tradeoff: the probability bound is much looser

Hoeffding's inequality

1. Symmetric Bernoulli: Hoeffding inequality
   Let X_i, $i = 1, 2, \ldots, n$ are independent symmetric Bernoulli r.v.s. Then

$$\Pr(|\sum_i a_i X_i| \geq t) \leq 2\exp\left(-\frac{t^2}{2\|a\|^2}\right)$$

   Proof idea
   - $\mathbb{E}[\exp(\lambda a_i X_i)] = (e^{\lambda a_i} + e^{-\lambda a_i})/2 = cosh(\lambda a_i)$
   - Show $cosh(x) \leq e^{x^2}/2$ (Ex 2.2.3)
   - conclude $\Pr(|\sum_i a_i X_i| \geq t) \leq exp(-\lambda t + \lambda^2 \sum_i a_i^2/2)$; minimize over λ.
2. General bounded r.v.s (including $Bern(p_i)$): Hoeffding inequality
   Let X_i, $i = 1, 2, \ldots, n$ are independent bounded r.v.s with $\Pr(m_i \leq X_i \leq M_i) = 1$. Then

$$\Pr(|\sum_i (X_i - E[X_i])| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_i (M_i - m_i)^2}\right)$$

   Proof: use Hoeffding's lemma: this bounds the MGF of a *zero mean* and *bounded* r.v..:

- Hoeffding's Lemma: Suppose $\mathbb{E}[X] = 0$ and $\Pr(X \in [a, b]) = 1$, then

$$M_X(s) := \mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}} \text{ if } s > 0$$

  - ★ Proof: use Jensen's inequality followed by mean value theorem, see `http://www.cs.berkeley.edu/~jduchi/projects/probability_bounds.pdf`

Chernoff's inequality

1. *Bern*(p_i) r.v.s: Chernoff inequality

Let X_i, $i = 1, 2, \ldots, n$ are independent Bernoulli r.v.s. with $X_i \sim Bern(p_i)$ and let $\mu = \sum_i p_i$.

- For a $t > \mu$,

$$\Pr(\sum_i X_i \geq t) \leq \exp(-\mu)(\frac{e\mu}{t})^t$$

- For a $0 \leq \delta < 1$,

$$\Pr(|\sum_i X_i - \mu| \geq \delta\mu) \leq 2\exp(-c\delta^2\mu)$$

Proof idea:

- For $t > \mu$: exact MGF expression, $1 + x \le e^x$, use $\lambda = \log(t/\mu)$ where $\mu := \sum_i p_i$. For $t < \mu$: exact MGF expression, $1 + x \le e^x$, set $\lambda = \log(1 + \delta)$ (obtained as the minimizer), then use this: for $0 < x < 1$, $\log(1 + x) \ge x/(1 + x/2)$. Finally use $1/(2 + \delta) < 1/3$ for $\delta < 1$ to get a bound of $\exp(-\mu\delta^2/3)$.
  - ★ for showing the last inequality, use this: show $g(\delta) \le f(\delta)$ by showing $g(\delta) - f(\delta)$ is decreasing in δ for $\delta \in [0, 1]$ and $g(0) - f(0) = 0$.

Bernstein for general bounded r.v.s

1. General bounded r.v.s: Bernstein inequality
Let X_i, $i = 1, 2, \ldots, n$ are independent bounded r.v.s with $\Pr(-M_i \le X_i \le M_i) = 1$. Then

$$\Pr(|\sum_{i=1}^{n}(X_i - E[X_i])| \ge t) \le 2\exp\left(-\frac{0.5t^2}{\sum_i \sigma_i^2 + 0.33(max_i M_i)t}\right)$$

where $\sigma_i^2 := \mathbb{E}[(X_i - E[X_i])^2]$. Assume $\sigma_i^2 \le \sigma_{mx}^2$ and $M_i \le M_{mx}$. Also simplify above further to get

$$\Pr(|\sum_{i=1}^{n}(X_i - E[X_i])| \ge t) \le 2\exp\left(-c\min(\frac{t^2}{n\sigma_{mx}^2}\frac{t}{M_{mx}}\right)$$

- Proof: use MGF bound of Ex 2.8.5

When $t > n\sigma^2/M_{mx}$ the prob bnd grows as $\exp(-t/M_{mx})$. When t is smaller, it grows as $\exp(-t^2/n\sigma^2)$. In this small t regime, we have $\exp(-t^2/n\sigma^2)$ decay.
In this small t regime, if $\sigma^2 \ll M_{mx}^2$, then the Bernstein bound is better than the Hoeffding bound (which always grows as $\exp(-t^2/nM_{mx}^2)$)

Hoeffding inequality only uses the bounds, but not the variance of X_is. It is not very tight if the variance is much smaller than the square of the range. This issue is addressed by use of Chernoff inequality for $Bern(p_i)$ r.v.s., and use of Bernstein inequality for general bounded r.v.s.

"variance much smaller than the square of the range" : $\sigma_{mx}^2/M_{mx}^2 \ll 1$ or more generally $\sum_i \sigma_i^2 \ll \sum_i M_i^2$

- $a \ll b$ means a/b is less than $O(1)$

equivalent for Bernoulli: $\sum_i p_i \ll n$, e.g., $\sum_i p_i \in O(\log n)$: this happens for sparse random graphs

Application: Boosting randomized algorithms

- Ex 2.2.8 of book (Boosting) : Suppose algo works correctly w.p. $0.5 + \delta$ (a little better than random guess). Run the algo n independent times and take majority vote. Show that answer correct w.p. $1 - \epsilon$ if $n \geq \frac{1}{2\delta^2} \log(1/\epsilon)$
- Ex 2.2.9 (Robust estimation / Median of Means):

Application: bounding degrees of dense or sparse random graphs, use Chernoff for sparse graphs

- Proposition 2.4.1 : Dense graphs are almost regular
  proof: use Chernoff for small deviations (Ex 2.3.5) for degree of one node i; then union bound to "unfix" i
- Problem 2.4.2, 2.4.3, 2.4.4
- Chernoff for $Bern(p_i)$ r.v.s gives a better bound than Hoeffding for bounded r.v.s when $p_i \ll 1/2$.
  The reason is Hoeffding does not use knowledge of p_i, only the fact that a Bernoulli r.v. is lower and upper bounded by $m_i = 0, M_i = 1$.

Sub-Gaussian and Sub-Exponential r.v.'s I

1. Definition and Properties of a sub-Gaussian r.v. X: for constants $K_i = CK$, the following are equivalent:
   1. $\Pr(|X| > t) \leq 2\exp(-t^2/K_1^2)$
   2. $\|X\|_{L_p} := \mathbb{E}[|X|^p]^{1/p} \leq K_2\sqrt{p}$
   3. $\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2\lambda^2)$ for $|\lambda| \leq 1/K_3$
   4. $\mathbb{E}[\exp(X^2/K_4^2)] \leq 2$
   5. If $\mathbb{E}[X] = 0$, then $\mathbb{E}[exp(\lambda X)] \leq \exp(K_5^2\lambda^2)$ for all λ.
2. Sub-Gaussian norm: can be defined as the smallest value of K for which any of the above properties hold.
   We use the second one here since that is easiest to interpret

   $$\|X\|_{\psi_2} := \sup_{p \geq 1} \frac{1}{\sqrt{p}} \mathbb{E}[|X|^p]^{1/p}$$

   (used in Vershynin's tutorial article)
   We can also define subG norm as the smallest value of K for which $\exp(X^2/K^2) \leq 2$:

   $$\|X\|_{\psi_2} := \inf_{K>0:\exp(X^2/K^2)\leq 2} K$$

   (this is used in the book)

3. Examples: Gaussian, Bernoulli, bounded
4. Sub-Gaussian Hoeffding inequality: Let $X_1, X_2, \dots X_n$ be independent zero-mean subG with subG norm K_i.
   Then $\sum_i X_i$ is also subG with subG norm $K = \sqrt{C \sum_i K_i^2}$.
   - Proof: Chernoff bounding followed by use of sub-G property.

Theorem (Sub-Gaussian Hoeffding inequality)

Let $X_1, X_2, \dots X_n$ be independent zero-mean subG r.v.s with subG norm K_i. Then, for every $t \geq 0$,

$$\Pr(|\sum_i X_i| \geq t) \leq 2 \exp\left(-c \frac{t^2}{\sum_i K_i^2}\right)$$

   - Proof: follows from above.

5. Definition/Properties of a sub-exponential r.v. X: for constants $K_i = CK$, the following are equivalent
   1. $\Pr(|X| > t) \leq 2 \exp(-t/K_1)$
   2. $\|X\|_{L_p} := \mathbb{E}[|X|^p]^{1/p} \leq K_2 p$
   3. $\mathbb{E}[\exp(\lambda |X|)] \leq \exp(K_3 \lambda)$ for $|\lambda| \leq 1/K_3$
   4. $\mathbb{E}[\exp(|X|/K_4)] \leq 2$

5. If $\mathbb{E}[X] = 0$, then $\mathbb{E}[exp(\lambda X)] \leq \exp(K_5^2 \lambda^2)$ for $|\lambda| \leq 1/K_5$

6. Proof main ideas
   - a ==> b: Integral identity, Gamma function property, $p^{1/p} \leq C$.
   - b ==> c: Taylor expansion, Sterling $p! > (p/e)^p$, $1/(1-x) < e^{2x}$
   - c ==> d: use $\lambda = c/K_3$, pick c so that $e^c = 2$.
   - d==> a : use Chernoff bounding for $|X|$
   - b ==> e: Taylor expansion, Sterling $p! > (p/e)^p$, $1 + x < e^x$
   - e ==> b: option 1: see book. option 2: Chernoff bounding should work to go from e to a

7. Sub-expo norm,

$$\|X\|_{\psi_1} := \sup_{p \geq 1} \frac{1}{p} \mathbb{E}[|X|^p]^{1/p}$$

8. Square of a sub-Gaussian is sub-expo with $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$

   proof:
   - immediate consequence of property d

9. If X, Y are sub-Gaussian with subG norms K_X, K_Y, then XY is sub-exponential with sub-expo norm $K_X K_Y$. In other words,

$$\|\mathbf{X}Y\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

   Proof:
   - consider normalized rvs X/K_X, Y/K_Y (here K_X, K_Y are their subG norms)
   - try to bound $\mathbb{E}[\exp(|XY|)]$ (property d) using $\mathbb{E}[X^2] \leq 2$ property for subG rvs
   - use Young's inequality twice: $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$
10. Examples: square of a sub-Gaussian,
11. Sub-exponential Bernstein inequality

Theorem (Sub-exponential Bernstein inequality)

Let $X_1, X_2, \ldots X_n$ be independent zero-mean sub-expo r.v.s with sub-expo norm K_i. Then, for every $t \geq 0$,

$$\Pr(|\sum_i X_i| \geq t) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_i K_i^2}, \frac{t}{\max_i K_i}\right)\right)$$

- Proof: Chernoff bounding; followed by use of sub-expo property v to bound the MGF of each term; pick λ as the minimum of the constraint on it and the value obtained by unconstrained minimiz over it.

12. Centering: if X is sub-G with sub-G norm K, then $X - \mathbb{E}[X]$ is subG with sub-G norm at most CK. Same for sub-expo r.v.s as well.
13. Comparing the different inequalities: Chebyshev, Bernstein, and Hoeffding
    - Hoeffding applies to the lightest tailed r.v.s (subGaussians). The probability exponent depends only $\sum_i K_{G,i}^2$ where $K_{G,i}$ is subG norm of X_i.
    - Bernstein applies to sub-expo r.v.s which are heavier tailed than subG but still somewhat "well-behaved". it depends on both $\sum_i K_{e,i}^2$ and $\max_i K_{e,i}$. The latter can be problematic sometimes for sums of sub-expo r.v.s that are such that $\max_i K_{e,i}$ is not small enough.
    - Chebyshev needs the least assumptions, applies to any r.v. with finite mean and variance. Used for r.v.s that are heavier tailed than sub-expo. It gives the loosest bounds

Truncation idea used in data science / ML: explained with 3 examples

- Truncation used in analyzing the algorithm: see https://arxiv.org/abs/1306.0160, Appendix A (Proof of the Initialization Step)
  - Bound $\sum_i X_i$ where X_i are r. matrices with some entries that are fourth powers of a Gaussian r.v.s. These entries are worse than sub-exponential. Can truncate these entries so each scalar G is truncated. Do this carefully so that it is possible to bound the residual term w.h.p. too.

- Truncation used to modify the algorithm, applied to the observed r.v. (convert it from worse-than-sub-expo to sub-expo) : https://yuxinchen2020.github.io/publications/TruncatedWF_CPAM.pdf (see Sec 2.2), Truncated Wirtinger Flow algorithm paper of Chen and Candes, but as cited there, the idea goes back to older work.
  - Idea: suppose we need to bound a term of the form $\sum_i z_i(y_i, \mathbf{a}_i)^2$ with z_i being indep, zero mean, $sub - expo(K_i)$ r.v.s. Since z_i are sub-expo, z_i^2 are even worse and (to my best knowledge), Chebyshev ineq is the only result to bound such a summation w.h.p. As we already discussed Cheby results in loose bounds. Here y_i and $\mathbf{a}_i$ are the available data/measurements and the known design/measurement vectors used in the algorithm design. And z_i is some function of both of these that is used in the defining error terms that need to be bounded.

- In the TWF context, $z_i = \mathbf{w}'\mathbf{Y}_{mat}\tilde{\mathbf{w}}$ with $\mathbf{w}, \tilde{\mathbf{w}}$ being arbitrary fixed unit vectors and $\mathbf{Y}_{mat} = \sum_i y_i \mathbf{a}_i \mathbf{a}_i'$ with $y_i := (\mathbf{a}_i' \mathbf{x}^*)^2$. See Sec 2.2. of https://yuxinchen2020.github.io/publications/TruncatedWF_CPAM.pdf
- A possible solution: truncate y_i using a carefully chosen large enough threshold to make the y_i's bounded. Here "truncate" is used in the sense of truncated Gaussian: u The threshold itself can depend on the mean of y_is.
- Then, can show that $\sum_i z_i(y_{trunc,i}, \mathbf{a}_i)^2$ is a sum of sub-expo r.v.s that can be bounded.

- Truncation used to modify the algorithm, applied to the observed r.v. (convert from sub-expo to sub-G): used in my work with Sara Nayer:
  - In other settings z_i are indep, zero mean, $subE(K_i)$ r.v.s., which means one can use the sub-expo Bern. But this requires a good enough bound on $\max_i K_i$. In some settings, this is not possible to get
  - Solution: truncate y_is to make them bounded and hence sub-G. Then can argue that z_is are also subG. In this particular setting the sum of subG norms was easy to get a good enough bound on.
  - details: see Sec II-A of https://arxiv.org/pdf/2102.10217.pdf