CS5339 Notes #8:
A Detour Into Concentration of Measure

Jonathan Scarlett

April 3, 2021

Useful references:

- Blog post by Jeremy Kun[1]
- First section of Boucheron *et al.*'s "Concentration Inequalities" notes[2]
- Appendix B of "Understanding Machine Learning" book

1 Introduction

- Given a random variable Y, how "concentrated" is Y (e.g., around its mean)?
- Rough statement: Suppose that we can find a deterministic value m, such that

$$\mathbb{P}[|Y - m| > t] \leq \text{TailBound}(t)$$

  where TailBound(t) decreases drastically to 0 in t.

  - Typically $m = \mathbb{E}[Y]$, and often TailBound(t) decreases exponentially, such as TailBound$(t) \sim e^{-ct}$ or TailBound$(t) \sim e^{-ct^2}$ for some $c > 0$.
  - In statistics, Y can be the estimation/prediction error. In computer science, Y can be the outcomes of randomized algorithms. There are many other applications in information theory, statistical physics, random matrices, statistical learning theory, etc.

- Simple example: Suppose $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$, where the X_i are i.i.d. with mean μ and variance σ^2.
  - **Law of Large Numbers:** $\mathbb{P}[|Y_n - \mu| > \epsilon] \to 0$ as $n \to \infty$.
  - **Central Limit Theorem:** $\mathbb{P}\big[|Y_n - \mu| > \frac{\alpha}{\sqrt{n}}\big] \to 2\Phi\big(-\frac{\alpha}{\sigma}\big)$ as $n \to \infty$, where Φ is the standard normal CDF.
  - **Large Deviations:** Under some technical assumptions, $\mathbb{P}[|Y_n - \mu| > \epsilon] \leq e^{-n \cdot \psi(\epsilon)}$ for some $\psi(\epsilon) > 0$. This type of result is the focus of this lecture.
  - **Moderate Deviations:** Decay rate of $\mathbb{P}[|Y_n - \mu| > \epsilon_n]$ when $\epsilon_n \to 0$ sufficiently slowly so that $\epsilon_n \sqrt{n} \to \infty$.

[1] http://jeremykun.com/2013/04/15/probabilistic-bounds-a-primer/
[2] http://www.econ.upf.edu/~lugosi/mlss_conc.pdf

- In many applications, we want the bounds to be *non-asymptotic* (i.e., holding for any n, as opposed to only in the limit $n \to \infty$).

2 Basic Inequalities

- Markov's inequality. Let Z be a *nonnegative* random variable. Then $\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}$.
  - Proof:

$$\begin{aligned}\mathbb{P}[Z \geq t] &= \int_0^\infty f_Z(z)\mathbf{1}\{z \geq t\}dz \\ &\leq \int_0^\infty \frac{z}{t} f_Z(z)\mathbf{1}\{z \geq t\}dz \\ &\leq \int_0^\infty \frac{z}{t} f_Z(z)dz \\ &= \frac{\mathbb{E}[Z]}{t}.\end{aligned}$$

  - This result definitely doesn't hold in general for RVs that can take negative values (e.g., take $Z \sim N(0,1)$ as a counter-example).

- Markov's inequality applied to functions: Let ϕ denote any *non-decreasing* and *non-negative* function. Let Z be any random variable. Then Markov's inequality gives

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

  where the first inequality uses the non-decreasing property, and the second uses Markov's inequality and the non-negative property.

- Chebyshev's inequality: Choose $\phi(t) = t^2$, and replace Z by $|Z - \mathbb{E}[Z]|$. Then

$$\mathbb{P}\big[|Z - \mathbb{E}[Z]| \geq t\big] \leq \frac{\mathrm{Var}[Z]}{t^2}.$$

- Chernoff bound: Choose $\phi(t) = e^{\lambda t}$ where $\lambda \geq 0$. Then we have

$$\mathbb{P}[Z \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}].$$

  Despite being a simple application of Markov's inequality, this bound is extremely useful.

3 Simplifying the Chernoff Bound

Rewriting the bound.

- The log-moment-generating function $\psi_Z(\lambda)$ of a random variable Z is defined as

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}], \quad \lambda \geq 0.$$

  Observe that the Chernoff bound above can be written as $\mathbb{P}[Z \geq t] \leq e^{-(\lambda t - \psi_Z(\lambda))}$.

– Note: If $\mathbb{E}[e^{\lambda Z}] = \infty$ for some λ, then this value of λ does not give a meaningful bound (but a smaller λ might be OK). If Z is sufficiently heavy-tailed, it could even be that $\mathbb{E}[e^{\lambda Z}] = \infty$ for *all* $\lambda > 0$, in which case, the Chernoff bound cannot be used.

- The Cramér transform of Z is defined as

$$\psi_Z^*(t) = \sup_{\lambda \ge 0} \big(\lambda t - \psi_Z(\lambda)\big). \tag{1}$$

By a direct substitution, setting $\lambda = 0$ would make the right-hand term zero, so since we are maximizing over all $\lambda \ge 0$, we conclude that $\psi_Z^*(t) \ge 0$ for all t.

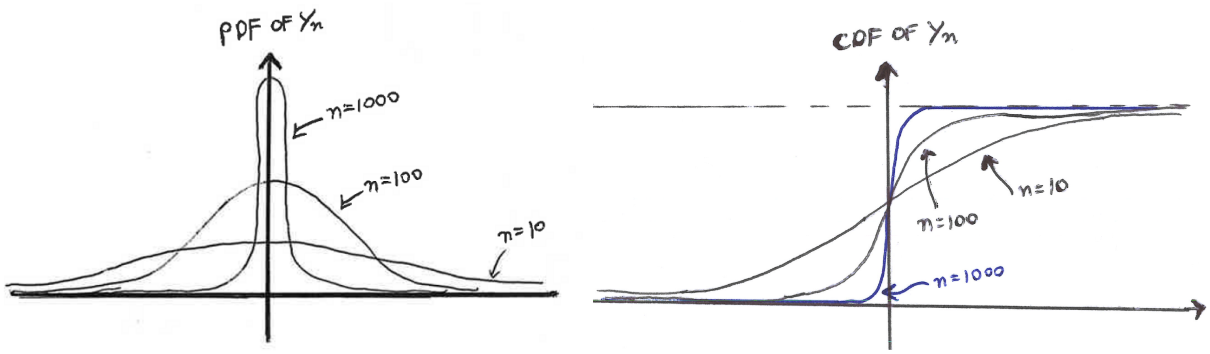
- By simply optimizing over all λ in the Chernoff bound, we have for any random variable Z that

$$\mathbb{P}[Z \ge t] \le \exp(-\psi_Z^*(t)).$$

This is known as the *Cramér-Chernoff Inequality.*

Sums of independent random variables.

- Let $Z = X_1 + \cdots + X_n$ where $\{X_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.). We expect better concentration of $Y_n = \frac{Z}{n}$ to as n increases:



- *Chebyshev's inequality on the sum:* We have $\mathrm{Var}[Z] = n\mathrm{Var}[X]$ (by the i.i.d. assumption), and hence Chebyshev's inequality with $t = n\epsilon$ gives

$$\mathbb{P}\bigg[\frac{1}{n}\big|Z - \mathbb{E}[Z]\big| \ge \epsilon\bigg] \le \frac{\mathrm{Var}[X]}{n\epsilon^2}.$$

  – This is an $O\big(\frac{1}{n}\big)$ probability of a "large" deviation, which can be useful but is typically not the best possible.

- *Cramér-Chernoff inequality on the sum:* We have

$$\begin{aligned}\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}] = \log \mathbb{E}\Big[e^{\lambda \sum_{i=1}^n X_i}\Big] &= \log \mathbb{E}\bigg[\prod_{i=1}^n e^{\lambda X_i}\bigg] \\ &= \log \prod_{i=1}^n \mathbb{E}\big[e^{\lambda X_i}\big] = \log\Big(\mathbb{E}\big[e^{\lambda X}\big]\Big)^n = n\psi_X(\lambda),\end{aligned}$$

where in the second line we used independence and then the identical distribution property. Then the Cramér-Chernoff inequality with $t = n\epsilon$ gives

$$\mathbb{P}[Z \ge n\epsilon] \le \exp\big(-n\psi_X^*(\epsilon)\big).$$

- This is looking better – exponential decay!
- But $\psi_X^*(\epsilon)$ is a bit complicated (it is not a closed-form formula, and it involves an optimization over λ) – can we simplify further?

- *A simple case: Gaussian random variables.*
  - Let $X \sim \mathcal{N}(0, \sigma^2)$.
  - A direct computation yields $\psi_X(\lambda) = \frac{\lambda^2\sigma^2}{2}$ (this requires a bit of integration).
  - Substituting into (1), we get the expression $\lambda t - \frac{\lambda^2\sigma^2}{2}$. Setting the derivative to zero gives the optimal $\lambda^* = \frac{t}{\sigma^2}$, and hence $\psi_X^*(t) = \frac{t^2}{2\sigma^2}$.
  - Therefore,

$$\mathbb{P}[X \ge t] \le \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Since X and $-X$ have the same distribution, the union bound $\Pr[A \cup B] \le \Pr[A] + \Pr[B]$ gives

$$\mathbb{P}[|X| \ge t] \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

or, when we sum n independent copies $Z = X_1 + \cdots + X_n$,

$$\mathbb{P}[|Z| \ge n\epsilon] \le 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

Since this example appears so frequently, it is used as a baseline for a much larger class of distributions with similar concentration behavior.

4 Sub-Gaussian Random Variables and Hoeffding's Inequality

Sub-Gaussian Random Variables.

- From the definition in (1) along with the above Gaussian example, we find that if $\psi_X(\lambda) \le \frac{\lambda^2\sigma^2}{2}$, then $\psi_X^*(t) \ge \frac{t^2}{2\sigma^2}$. This motivates the following definition.

- **Definition.** A *zero-mean* random variable X is said to be *sub-Gaussian* with parameter σ^2 if $\psi_X(\lambda) \le \frac{\lambda^2\sigma^2}{2}$, $\forall \lambda > 0$. Denote the set of all such random variables by $\mathcal{G}(\sigma^2)$.

- Properties of sub-Gaussian random variables:
  1. $\mathbb{P}[|X| \ge t] \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ (as we already proved for Gaussians)
  2. If $X_i \in \mathcal{G}(\sigma_i^2)$ are independent, then $\sum_{i=1}^n a_i X_i \in \mathcal{G}\big(\sum_{i=1}^n a_i^2 \sigma_i^2\big)$ (just like with Gaussians)

The straightforward proofs of these properties are omitted.

- Combining these properties (with $t = n\epsilon$), we find that if $Z = X_1 + \ldots + X_n$ where the X_i are independent and sub-Gaussian with parameter σ^2, then

$$\mathbb{P}[|Z| \geq n\epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right),$$

just like the sum of n independent Gaussians.

Bounded Random Variables.

- An important class of sub-Gaussian random variables is the class of bounded random variables.
- **Theorem.** Let X be a random variable with $\mathbb{E}[X] = 0$, taking values in a bounded interval $[a, b]$. Then we have $X \in \mathcal{G}\big(\frac{(b-a)^2}{4}\big)$.
  - A proof outline is below, with the details left as an optional tutorial exercise.
- Using this result and the first sub-Gaussian property above, we find that for $X \in [a, b]$,

$$\mathbb{P}\big[|X - \mathbb{E}[X]| > t\big] \leq 2\exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

  - Although the theorem assumed $\mathbb{E}[X] = 0$, we can always replace X by $X - \mu$ and $[a, b]$ by $[a - \mu, b - \mu]$, which clearly doesn't change $b - a$.

  Using a similar argument along with the fact that sums of sub-Gaussian variables are sub-Gaussian, we obtain the following.

- **Corollary (Hoeffding's inequality)** Let $Z = X_1 + \cdots + X_n$, where the X_i are independent and supported on $[a_i, b_i]$. Then

$$\mathbb{P}\left[\frac{1}{n}\big|Z - \mathbb{E}[Z]\big| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right).$$

5 Proof Outline: Bounded RVs are Sub-Gaussian

- Main steps of the proof.
  1. Prove that $\mathrm{Var}[Z] \leq \frac{(b-a)^2}{4}$ for any Z bounded on $[a, b]$.
  2. Show $\psi_X(0) = 0$, $\psi'_X(0) = 0$, and $\psi''_X(\lambda) = \mathrm{Var}[Z]$, where Z is a random variable with PDF $f_Z(z) = e^{-\psi_X(\lambda)} e^{\lambda z} f_X(z)$; hence $\psi''_X(\lambda) \leq \frac{(b-a)^2}{4}$ by Step 1.
  3. Taylor expand $\psi_X(\lambda) = \psi_X(0) + \lambda\psi'_X(0) + \frac{\lambda^2}{2}\psi''_X(\theta)$ (for some $\theta \in [0, \lambda]$) and substitute Step 2 to upper bound this by $\frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$.
- The details are left as an optional tutorial exercise.

6 Example Applications

Example 1: Typical Sequences.

- Let $(U_1, \dots, U_n)$ be i.i.d. random variables drawn from a PMF P_U. Assume that U is integer-valued and finite, only taking values $\{1, \dots, m\}$ for some integer m.
- Question. How many occurrences of each value $u \in \{1, \dots, m\}$ occur?
- Let $Z_u = \sum_{i=1}^n \mathbf{1}\{U_i = u\}$. This is a sum of i.i.d. random variables bounded within $[0,1]$, and $\mathbb{E}[Z_u] = nP_U(u)$. So by Hoeffding's inequality,

$$\mathbb{P}\big[\big|Z_u - nP_U(u)\big| \ge n\epsilon\big] \le 2e^{-2n\epsilon^2}.$$

- Since there are m values that U can take, the union bound gives

$$\mathbb{P}\bigg[\bigcup_{u=1,\dots,m}\Big\{\big|Z_u - nP_U(u)\big| \ge n\epsilon\Big\}\bigg] \le 2m \cdot e^{-2n\epsilon^2}.$$

  Re-arranging, we find that probability is upper bounded by $\delta > 0$ under the choice $\epsilon = \sqrt{\frac{\log \frac{2m}{\delta}}{2n}}$. Equivalently, if $n \ge \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$, then the above probability is at most δ.

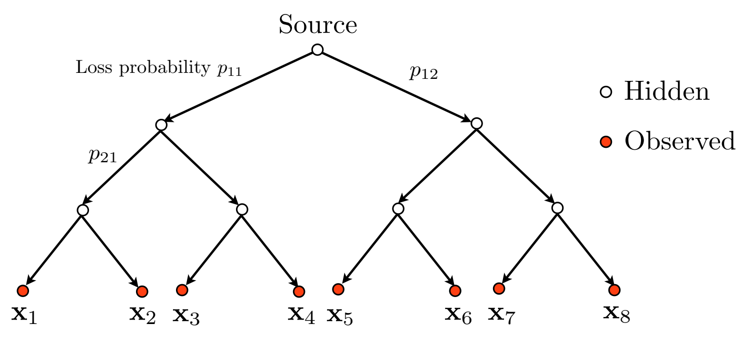
- The above findings can be viewed in at least two ways:
  - With high probability, all of the counts are within $O(\sqrt{n \log m})$ of their mean as n grows large.
  - For the counts to deviate from their mean by at most $n\epsilon$ with high probability, it suffices to have $n = \text{constant} \times \frac{\log m}{\epsilon^2}$ samples.

Example 2: Graph Degree.

- As an exercise, see if you can use the analysis of Example 1 to bound the maximum degree in a random graph with high probability.
  - More precisely, consider a random graph with n nodes, in which each given edge is present with probability p (independent from all other edges). The edges have no direction, so there are $\binom{n}{2}$ potential edges, and the average number of edges is $p\binom{n}{2}$.
  - The degree of a node is defined as the number of edges attached to that node. For a given node, its mean is $(n-1)p$. The maximum degree of the graph is the highest degree among the n nodes.

Example 3: Network Tomography.

- Network tomography problem:



  - Starting at the source, a packet is sent along both branches following the arrows until hitting the leaves (shown in red)
  - Each link has a probability of the packet being lost (independent of all other links)
  - We only get to observe which packets ended up arriving at the leaves.

- In the case of n packets and p leaf nodes, define
  - $X_k^{(i)} = \mathbf{1}\{\text{packet } i \text{ arrives at node } k\}$ for $i = 1, \cdots, n$ and $k = 1, \cdots, p$
  - Goal: Given these n independent samples, reconstruct the tree structure.

- Outline of analysis in the paper [Ni, 2011]:[3]
  - Show that the tree can be recovered from the values $q_{kl} = \mathbb{P}[\text{packet reaches } x_k \text{ and } x_l]$
  - Show robustness, in the sense that any $\hat{q}$ with $|\hat{q}_{kl} - q_{kl}| \leq \epsilon$ suffices
  - Set $\hat{q}_{kl} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_k^{(i)} = 1 \cap X_l^{(i)} = 1\}$, and bound using Hoeffding's inequality:

$$\mathbb{P}[|\hat{q}_{kl} - q_{kl}| > \epsilon] \leq 2\exp(-2n\epsilon^2).$$

  - Apply the union bound over the $\binom{p}{2} \leq \frac{1}{2}p^2$ possible pairs of leaf nodes to conclude that $\mathbb{P}[\text{error}] \leq \delta$ as long as $n \geq \frac{1}{2\epsilon^2}\log\frac{p^2}{2\delta}$.

Example 4: Statistical Learning Theory.

- ...see the next lecture!

7 Bounded Differences

(This section is included for the sake of interest, but we will not make use of it)

- A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property if, for some positive $c_1, .., c_n$,

$$\sup_{x_1,...,x_n,x_i' \in \mathcal{X}} |f(x_1, .., x_i, ..., x_n) - f(x_1, ..., x_i', ..., x_n)| \leq c_i$$

  for all $i = 1, \dots, n$. This means that changing any single input value does not change the output value too much.

- Example. Let $V = \{1, \cdots, n\}$, and let G be a random graph such that each pair $i, j \in V$ is independently connected with probability p. Let

$$X_{ij} = \begin{cases} 1 & (i,j) \text{ are connected} \\ 0 & \text{otherwise.} \end{cases}$$

  The *chromatic number* of G is the minimum number of colors needed to color the vertices such that no two connected vertices have the same color. Writing

$$\text{chromatic number } = f(X_{11}, \cdots, X_{ij}, \cdots, X_{nn}),$$

[3]The first two steps here are not obvious nor particularly easy to prove, so let's take them for granted and focus on the concentration part (third step).

we find that f satisfies the bounded difference property with $c_{ij} = 1$.

  - Countless other examples also exist

- **Theorem (McDiarmid's Inequality).** Let $X_1, ..., X_n$ be independent random variables, and let f satisfy the bounded differences property with c_i's. Then

$$P\Big(|f(X_1, ..., X_n) - \mathbb{E}[f(X_1, ..., X_n)]| > t\Big) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

  - A very useful generalization of Hoeffding's inequality (which is recovered from this result by choosing $f(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ when the random variables satisfy $X_i \in [a_i, b_i]$).
  - Harder to prove (beyond the scope of this course)