CS5275 Lecture 2: Concentration Inequalities

Jonathan Scarlett

January 20, 2025

Useful references:

- Blog post by Jeremy Kun[1]
- First section of Boucheron *et al.*'s "Concentration Inequalities" notes[2]
- Chapter 2 of Vershynin's book "High Dimensional Probability"
- CS-style course notes: Chapter 2 of USyd course `https://ccanonne.github.io/teaching/COMPx270`
- CS-style textbooks: "Concentration of Measure for the Analysis of Randomized Algorithms" (Panconesi/Dubhashi) and "Randomized Algorithms" (Motwani/Raghavan)

Categorization of material:

- Core material: Sections 1–4 and Examples 1–4 of Section 6
- Extra material: Section 5, rest of Section 6, Section 7

(Exam will strongly focus on "Core". Take-home assessments may occasionally require consulting "Extra".)

1 Introduction

- Given a random variable Y, how "concentrated" is Y (e.g., around its mean)? We will particularly be interested in the case where $Y = f(X_1, \dots, X_n)$ is a function of n independent random variables $X_1, \dots, X_n$, such as the empirical average $Y = \frac{1}{n}\sum_{i=1}^n X_i$.
- Let $Y = Y_n$ to make explicit that Y depends on n. Roughly, a concentration inequality is an inequality stating that there exists a deterministic value m such that

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

  where $\text{TailBound}(n, t)$ ideally decreases to 0 rapidly as n increases.

  - Typically $m = \mathbb{E}[Y_n]$ (other choices may include $m = \text{median}(Y_n)$ or $m = 0$), and often $\text{TailBound}(n, t)$ decreases exponentially, such as $\text{TailBound}(n, t) \sim e^{-cnt^2}$ for some $c > 0$.
  - Such results are useful because they tell us that *the behavior of Y_n becomes more and more predictable as n increases*; namely, we know that Y_n will be very close to m with high probability.

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

  - In statistics, Y may be a quantity being estimated from data. In computer science, Y can represent the outcome of a randomized algorithm. There are many other applications in information theory, statistical learning theory, statistical physics, random graph theory, random matrix 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, $\mathbb{P}[|Y_n - \mu| > \epsilon] \le e^{-n\cdot\psi(\epsilon)}$ 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 $\epsilon_n\sqrt{n} \to \infty$.

- 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 *non-negative* random variable. Then $\mathbb{P}[Z \ge t] \le \frac{\mathbb{E}[Z]}{t}$.
  - Proof: Suppose for simplicity that Z is continuous with density f_Z (if Z is discrete, just replace integrals by summations below). Then:

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

  - Note that 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

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

  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

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

- 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

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

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

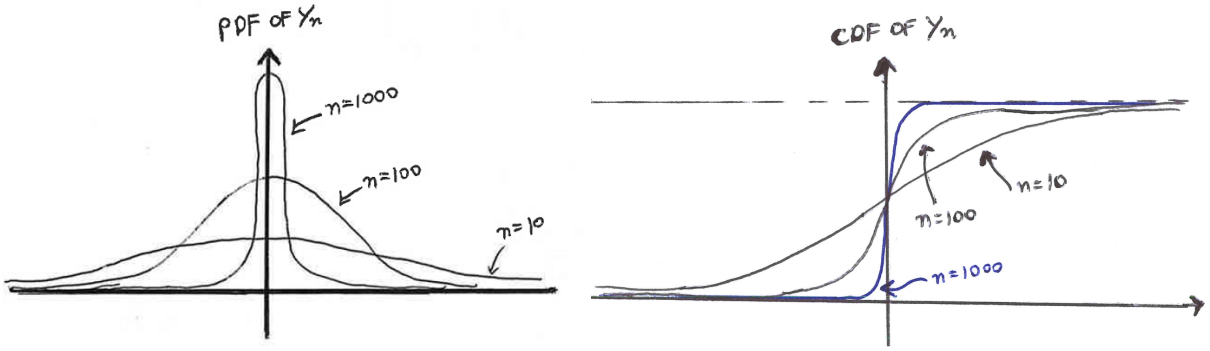
- By simply optimizing over all λ in the Chernoff bound, we have for any random variable Z that

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

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 sharper concentration of $Y_n = \frac{Z}{n}$ 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

$$\mathbb{P}\left[\frac{1}{n}\big|Z - \mathbb{E}[Z]\big| \geq \epsilon\right] \leq \frac{\mathrm{Var}[X]}{n\epsilon^2}.$$

  - 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

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

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

$$\mathbb{P}[Z \geq n\epsilon] \leq \exp\big(-n\psi_X^*(\epsilon)\big). \tag{2}$$

  - 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,

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

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

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

    When we sum n independent copies $Z = X_1 + \cdots + X_n$, analogous reasoning applied to (2) gives

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

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

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 $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, then $\psi_X^*(t) \geq \frac{t^2}{2\sigma^2}$. This motivates the following definition.
- **Definition.** A *zero-mean* random variable X is said to be *sub-Gaussian* with parameter σ^2 if $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, $\forall \lambda > 0$. Denote the set of all such random variables by $\mathcal{G}(\sigma^2)$.
  - Note: Sub-Gaussian variables are very "light-tailed" (tails decaying like e^{-ct^2}). Similar concepts also exists for distributions whose tails are less light, notably including *sub-exponential* (tails decaying like e^{-ct} e.g., see Vershynin's book).
- Properties of sub-Gaussian random variables:
  1. $\mathbb{P}[|X| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ (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.
- Equivalent definitions: Sometimes checking whether $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$ can be difficult, e.g., because the MGF is complicated or has no closed-form expression. To verify the sub-Gaussian property, it is useful to note that the following statements are all equivalent for zero-mean X:
  1. (MGF) There exists $K_0 > 0$ such that $\psi_X(\lambda) \leq K_0^2\lambda^2$ for all $\lambda > 0$ (i.e., the above definition of sub-Gaussianity with $K_0^2 = \frac{\sigma^2}{2}$).
  2. (Tail Behavior) There exists $K_1 > 0$ such that $\mathbb{P}[|X| \geq t] \leq 2\exp\left(-\frac{t^2}{K_1^2}\right)$ for all $t \geq 0$.
  3. (Moments) There exists $K_2 > 0$ such that $\mathbb{E}[|X|^p]^{1/p} \leq K_2\sqrt{p}$ for all $p \geq 1$.

  The proofs are omitted here (e.g., see Proposition 2.5.2 of Vershynin's book). The quantities K_0, K_1, K_2 may differ in general, but they all match to within a constant factor (and thus all play a similar role as σ above).

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 appendix.

- Using this result and the first sub-Gaussian property above, we find that for $X \in [a, b]$,

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

  - Although the theorem assumed $\mathbb{E}[X] = 0$, we can always replace X by $X - \mu$ and $[a, b]$ by $[a - \mu, b - \mu]$, so the difference between the upper and lower limit is still $b - a$.

  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

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

  When we have all $a_i = a$ and all $b_i = b$, this simplifies to

$$\mathbb{P}\bigg[\frac{1}{n}\big|Z - \mathbb{E}[Z]\big| > \epsilon\bigg] \leq 2\exp\bigg(-\frac{2n\epsilon^2}{(b-a)^2}\bigg),$$

  or even more simply, in the commonly-encountered scenario where $a = 0$ and $b = 1$, we have

$$\mathbb{P}\bigg[\frac{1}{n}\big|Z - \mathbb{E}[Z]\big| > \epsilon\bigg] \leq 2e^{-2n\epsilon^2}.$$

- To keep the expressions simple, we discuss the latter case in further detail. This bound can be viewed as fixing (ϵ, n) and asking how small the deviation probability is. It is also to keep in mind two equivalent statements:

  - If we want the deviation probability to be upper bounded by some $\delta > 0$, and ϵ is given, then we get the following condition by setting $2e^{-2n\epsilon^2} = \delta$ and re-arranging:

$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}.$$

    This amounts to fixing (ϵ, δ) and asking how large n needs to be be.

  - Similarly, we can fix (δ, n) and ask what the smallest possible ϵ could be, giving $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$. This is consistent with how the Central Limit Theorem shows that most of the probability is within $O\big(\frac{1}{\sqrt{n}}\big)$ of the true mean.

5 (**Optional**) 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 given in the appendix of this document.

6 Example Applications

Example 1: Estimating Population Statistics

- Suppose that we have a huge population of voters for an upcoming two-party election, and we want to accurately predict the proportion that will vote for Party A. (Note: We can adapt this example to estimating other things, like prevalence of a disease.)
- Strategy: Choose a voter uniformly at random[3] and ask how they will vote (assuming they will respond honestly). Repeat this n times, and let $\hat{p}$ be the fraction of those n that responded Party A.
- Question: How large should n be to ensure (via Hoeffding's inequality) that $|p^* - \hat{p}| \leq 0.02$ with probability at least 0.95, where p^* is the true proportion and $\hat{p}$ is the estimate?
- Analysis: The n binary observations $X_1, \ldots, X_n$ (1 if Party A, 0 if Party B) are Bernoulli(p^*), and letting $Z = X_1, \ldots, X_n$, it follows that $\hat{p} = \frac{Z}{n}$ and $p^* = \frac{\mathbb{E}[Z]}{n}$. Hence, we can apply Hoeffding's inequality directly to get

$$\mathbb{P}\big[|\hat{p} - p^*| > 0.02\big] \leq 2e^{-2n(0.02)^2}.$$

  Equating this with 0.05 and re-arranging, we get that $n = \lceil \frac{1}{2(0.02)^2} \log \frac{2}{0.05} \rceil = 4612$ suffices.

- Exercise: Adapt this example to a scenario where every voter answers honestly with probability exactly 0.9 (independently of all other questions/answers).

Example 2: Typical Sequences.

- Let $(U_1, \ldots, U_n)$ be i.i.d. random variables drawn from a PMF P_U. Assume that U is integer-valued and finite, only taking values $\{1, \ldots, m\}$ for some integer m.
- Question. How many occurrences of each value $u \in \{1, \ldots, m\}$ 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,

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

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

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

  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 $n \geq \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$, then the above probability is at most δ.

[3] Probably the main reason that actual polls can be quite inaccurate is that the people they poll are *not* uniformly random.

- 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 3: 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 4: Randomized Algorithms

- As an exercise, you can try the following: Suppose that a randomized algorithm produces the correct output with probability at least 2/3. Show that by independently running the algorithm n times and letting the final output be the one output the highest number of times, we can boost the success probability to any target $1-\delta$ with a number of trials satisfying $n = O\big(\log \frac{1}{\delta}\big)$.

Example 5: Estimation Under Heavy-Tailed Noise.

- A fundamental primitive in statistics and related areas is estimating the mean of a random variable from independent samples (i.e., given $X_1, \ldots, X_n$ each drawn from P_X, estimate $\mu = \mathbb{E}[X]$).
- If $X - \mu$ is sub-Gaussian for $X \sim P_X$, then we know that the empirical mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ works well, with $e^{-cn\epsilon^2}$ decay of the probability of being ϵ-far from the correct value (for some constant c).
- What if we only know that $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}[X]$ are finite, but the higher moments may be infinite? This occurs for *heavy tailed* distributions, which are often used to model outliers in the data.
- At first glance it looks hopeless to consider Hoeffding's inequality (as X is not only unbounded but heavy-tailed!), but in fact with a more clever choice of estimator, all we need is Hoeffding's and Chebyshev's inequality, and we can get the sub-Gaussian level of accuracy mentioned above!
- The more clever estimator is called *median of means*:
  - Split the n samples into K blocks of size B, so that $n = KB$ (we will ignore rounding/divisibility issues here and below, as they are insignificant for the interesting regimes of parameter choices)
  - For $k = 1, \ldots, K$, let $\hat{\mu}_k$ be the empirical mean computed using only the samples in block k.
  - The final estimate is $\hat{\mu} = \text{median}(\hat{\mu}_1, \ldots, \hat{\mu}_K)$.
- By the definition of median, if $|\hat{\mu} - \mu| > \epsilon$, then at least half of the values in $\hat{\mu}_1, \ldots, \hat{\mu}_K$ must be ϵ-far from μ. Hence,

$$\mathbb{P}\big[|\hat{\mu} - \mu| > \epsilon\big] \le \mathbb{P}\bigg[\sum_{k=1}^{K} Z_k \ge \frac{K}{2}\bigg],$$

  where Z_k equals 1 if $|\hat{\mu}_k - \mu| > \epsilon$, and $Z_k = 0$ otherwise.

- Defining $p_\epsilon = \mathbb{E}[Z_k] = \mathbb{P}[|\hat{\mu}_k - \mu| > \epsilon]$ and $t = \frac{1}{2} - p_\epsilon$, the above right-hand side is equivalent to

$$\mathbb{P}\Big[\sum_{k=1}^{K}(Z_k - p_\epsilon) \geq Kt\Big],$$

  and as long as $t > 0$ (which we will verify shortly), this is upper bounded by e^{-2Kt^2} by Hoeffding's inequality. Substituting back $t = \frac{1}{2} - p_\epsilon$, we have proved that

$$\mathbb{P}\big[|\hat{\mu} - \mu| > \epsilon\big] \leq e^{-2K(1/2-p_\epsilon)^2}.$$

- Next, by the definition $p_\epsilon = \mathbb{P}[|\hat{\mu}_k - \mu| > \epsilon]$ and the fact that $\hat{\mu}_k$ is the empirical average of B samples, we can simply apply Chebyshev's inequality to obtain $p_\epsilon \leq \frac{\sigma^2}{B\epsilon^2}$, and substituting $B = \frac{n}{K}$ gives $p_\epsilon \leq \frac{K\sigma^2}{n\epsilon^2}$. In particular, if we choose $K = \frac{n\epsilon^2}{4\sigma^2}$, we get $p_\epsilon \leq \frac{1}{4}$ (which gives the desired property $t > 0$ mentioned above), and the previous display equation becomes

$$\mathbb{P}\big[|\hat{\mu} - \mu| > \epsilon\big] \leq e^{-K/8} = \exp\Big(-\frac{n\epsilon^2}{32\sigma^2}\Big).$$

  This is the desired sub-Gaussian style concentration! (Note: The factor of 32 can be improved via a more careful analysis)

- **Caveats/discussion:**
  - Setting $K = \frac{n\epsilon^2}{4\sigma^2}$ requires knowing ϵ and σ in advance, which may be questionable. On the other hand, setting the above upper bound $e^{-K/8}$ to a target value δ gives $K = 8\log\frac{1}{\delta}$, so we can actually set K given only knowledge of such a target δ, which may be more natural.
  - Also note that since K should be an integer, more care is needed with rounding if (e.g.) $\frac{n\epsilon^2}{4\sigma^2} < 1$; the above result may not hold as stated in such scenarios.

Other uses:

- See "Randomized Algorithms" (Motwani/Raghavan) for an entire book on randomized algorithms where these techniques are prominent.
- See `https://www.comp.nus.edu.sg/~scarlett/CS5339_notes/09-Theory_Notes.pdf` for the use of Hoeffding's inequality in statistical learning theory, a theoretical branch of machine learning.
- (This list could be made much longer!)

Other useful concentration bounds:

- As a simple example of where Hoeffding's inequality can be weaker than ideal, suppose that we are interested in the probability that Binomial(n, p) takes value 0. This event has probability $(1-p)^n \leq e^{-pn}$, Hoeffding's inequality would only give a bound of e^{-2p^2n}, which is much weaker when p is small. The concentration bounds below serve as alternatives that circumvent this weakness.
- Bernstein's inequality: If $Z = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$ with the X_i being i.i.d. and satisfying $|X_i| \leq M$ (with probability one), then

$$\mathbb{P}[Z \geq t] \leq \exp\Big(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{1}{3}Mt}\Big).$$

For example, when $X_i \sim \text{Bernoulli}(p)$ we have $M = 1$ and $\mathbb{E}[X_i^2] = p$, and setting $t = \epsilon n$ gives a bound of $\exp\big(-\frac{n\epsilon^2}{2(p+\epsilon/3)}\big)$. This is "Hoeffding-like" when p is "large", but has better $e^{-\Theta(n\epsilon)}$ behavior when p is "small" (e.g., $p \le \epsilon \ll 0$).

  - Note: More general forms of Bernstein's inequality consider *all* moments of the random variable, and drop the requirement of the random variable being bounded. See for example Chapter 2 of Vershynin's textbook.

- Binomial tail bounds: Since the binomial distribution arises especially frequently, it's useful to highlight some of its most widely-used tail bounds. Letting $Z \sim \text{Binomial}(n, p)$ (i.e., Z is the sum of n independent Bernoulli(p) variables), we have the following:

  - The Chernoff bound can be simplified to give

$$\begin{aligned} \mathbb{P}[Z \le \gamma n] &\le e^{-nD(\gamma\|p)} \quad \text{if } \gamma \le p \\ \mathbb{P}[Z \ge \gamma n] &\le e^{-nD(\gamma\|p)} \quad \text{if } \gamma \ge p, \end{aligned}$$

    where $D(a\|b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ (see *KL divergence* or *relative entropy* in the upcoming lecture on information theory). These bounds are usually tight to within a $\Theta\big(\frac{1}{\sqrt{n}}\big)$ factor (which is usually insignificant compared to the exponential terms).

  - The following weakened bounds are often more "user-friendly":

$$\begin{aligned} \mathbb{P}[Z \ge (1+\delta)np] &\le \exp\big(-np\big((1+\delta)\log(1+\delta) - \delta\big)\big) \quad \text{for } \delta > 0 \\ \mathbb{P}[Z \le (1-\delta)np] &\le \exp\big(-np\big((1-\delta)\log(1-\delta) + \delta\big)\big) \quad \text{for } \delta \in (0,1) \end{aligned}$$

    These can also be further weakened to the following particularly simple bounds:

$$\begin{aligned} \mathbb{P}[Z \ge (1+\delta)np] &\le \exp\left(-np\cdot\frac{1}{3}\delta^2\right) \quad \text{for } \delta \in (0,1) \\ \mathbb{P}[Z \le (1-\delta)np] &\le \exp\left(-np\cdot\frac{1}{3}\delta^2\right) \quad \text{for } \delta \in (0,1). \end{aligned}$$

    All four of these are particularly useful when p is small (e.g., decreasing as n increases), in which case Hoeffding's inequality may not be powerful enough.

7 Beyond Sums of Independent Random Variables

In many scenarios throughout machine learning and statistics, we would like to establish concentration random variables that are not sums of independent random variables. This is often much more difficult, but there exist tools for this purpose – below are just two examples (without proofs).

Bounded differences and McDiarmid's inequality.

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

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

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

- Example 1: 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

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

we find that f satisfies the bounded difference property with $c_{ij} = 1$. This is because adding (resp., removing) an edge at most amounts to needing to add (resp. being able to remove) one color.

- Example 2: Suppose that we throw m balls into n bins uniformly at random. Let $X_1, \ldots, X_m$ be random variables giving the bin indices of the balls. Then if we are interested in a function $f(x_1, \ldots, x_m)$ such as *the number of empty bins* or *the number of bins with at least 2 balls*, we clearly have that $f(\cdot)$ changes by at most one whenever a single index X_i changes. Thus, we again have the bounded differences property with $c_i = 1$.

- **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

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

- This is 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 $X_i \in [a_i, b_i]$ with $c_i = b_i - a_i$).

- **Example (kernel density estimation):** Suppose that we have i.i.d. samples $X_1, \ldots, X_n$ with each X_i being drawn from some probability density function $\psi(x)$, and we would like to know the function ψ (as best we can). The Kernel Density Estimation (KDE) method does this by estimating

$$\widehat{\psi}(x) = \frac{1}{n}\sum_{i=1}^n K(x - X_i)$$

for some "kernel" K such that $\int_{-\infty}^{\infty} K(z)dz = 1$ (e.g., a Gaussian-like curve, so that we try to put more density near each X_i but "smooth it out" away from that point). Suppose that we are interested in the overall error $Z = \int_{-\infty}^{\infty} |\psi(x) - \widehat{\psi}(x)|dx$.

Observe that $Z = f(X_1, \ldots, X_n)$, where the function f is non-linear and quite complicated. Despite this, we can easily deduce its concentration behavior – whenever a single data point is changed, only a single term in $\frac{1}{n}\sum_{i=1}^n K(x - X_i)$ gets affected, and since K integrates to one we can easily check that

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq \frac{2}{n}.$$

Hence, McDiarmid's inequality with $c_i = \frac{2}{n}$ gives the concentration bound

$$\mathbb{P}\big(|Z - \mathbb{E}[Z]| \geq t\big) \leq 2e^{-nt^2},$$

which signifies very sharp concentration around the average when n is large.

- Note: Understanding $\mathbb{E}[Z]$ itself may still be difficult (and strongly dependent on K and ψ), but whatever value it happens to be, we now know that there is sharp concentration around it.

(Optional**) Martingales and Azuma's inequality.**

- A sequence $Y_1, \ldots, Y_n$ is said to be a *martingale* if it holds for each time step i that $\mathbb{E}[Y_{i+1} \,|\, Y_1, \ldots, Y_i] = Y_i$ (no matter which $Y_1, \ldots, Y_i$ values we condition on).[4]
  - In other words, conditioned on the sequence so far, the next element neither increases nor decreases *on average*.
  - Example: If $Y_i = \sum_{i=1}^{n} X_i$ where $X_1, \ldots, X_n$ are independent zero-mean random variables, then it's easy to check that $Y_1, \ldots, Y_n$ is a martingale. (But there are many other examples of martingales for which the increments $Y_{i+1} - Y_i$ are not independent.)
  - Intuition: If a gambler is playing in a *fair* casino, then what happened so far may impact which games they play or how large their bets are, creating some dependence structure. But no matter which (fair) games or what bet sizes, on average there is no gain or loss. (In reality, there would always be an average loss, leading to a related notion called a *super-martingale*.)
    - Naturally, this concept arises for similar reasons in *sequential decision-making* problems in theoretical computer science, machine learning, etc.
  - Doob martingale: Although seemingly different, the martingale idea is closely related to McDiarmid's inequality (in fact, the theorem below can be used to prove McDiarmid's inequality). Briefly, this connection stems from the *Doob martingale*, where for a random variable of the form $A = f(Z_1, \ldots, Z_n)$ (with independent Z_i's) we define $X_i = \mathbb{E}[A|Z_1, \ldots, Z_i]$. This can easily be shown to produce a martingale. As a concrete example, if $Z_1, \ldots, Z_n$ were edges in a random graph, then the sequence $X_1, \ldots, X_n$ would have an interpretation of *revealing the edges one-by-one* and seeing how some property $f(\cdot)$ develops (e.g., number of triangles formed in the graph).
- **Theorem (Azuma's Inequality).** If $Y_0, Y_1, \ldots, Y_n$ is a martingale and it holds with probability one that $|Y_{i+1} - Y_i| \leq c_i$ for all i (i.e., *bounded increments*), then it holds that

$$\mathbb{P}\big[|Y_n - Y_0| \geq n\epsilon\big] \leq 2\exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n} c_i^2}\right).$$

  - The extra term Y_0 is included for convenience, and whether or not it's included is just a matter of renaming (e.g., shifting n to $n+1$). We could also just specialize to the case that $Y_0 = 0$.
  - In the case of independent sums $Y_i = Y_0 + \sum_{i=1}^{n} X_i$, if we assume that $X_i \in [a_i, b_i]$, then we get $c_i = b_i - a_i$, and the result precisely reduces to Hoeffding's inequality.

[4] Strictly speaking it is also required that $\mathbb{E}[|Y_i|]$ is finite for each i, but this is a minor technical condition.

(**Optional**) Appendix: Proving Bounded RVs are Sub-Gaussian

Claim: Any bounded random variable $Z \in [a, b]$ has variance at most $\text{Var}[Z] \leq \frac{(b-a)^2}{4}$.
Proof:

- It suffices to show $\text{Var}[Z] \leq \frac{1}{4}$ when $Z \in [0, 1]$, since then the general case follows by shifting and re-scaling.
- We have

$$\mathbb{E}[(Z - c)^2] = \mathbb{E}[Z^2] - 2c\mathbb{E}[Z] + c^2,$$

  which is minimized at $c = \mathbb{E}[Z]$ (check by setting the derivative to zero). Therefore, $\text{Var}[Z] \leq \mathbb{E}[(Z-c)^2]$ for any c.
- Setting $c = \frac{1}{2}$ and using the fact that $Z \in [0, 1]$, we conclude that $\text{Var}[Z] \leq \frac{1}{4}$, as required.

Claim: (Recall that the log moment generating function is defined as $\psi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}],\ \lambda \geq 0$.) Assuming that $\mathbb{E}[X] = 0$, we have $\psi_X(0) = 0$, $\psi'_X(0) = 0$, and $\psi''_X(\lambda) = \text{Var}[Z]$ for any $\lambda > 0$, where Z is a random variable with PDF $f_Z(z) = e^{-\psi_X(\lambda)} e^{\lambda z} f_X(z)$. (Note: Z implicitly depends on λ)
Proof:

- (i) Since $\psi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, we have $\psi_X(0) = \log 1 = 0$.
- (ii) By direct differentiation, we have $\psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$, which implies $\psi'_X(0) = \mathbb{E}[X] = 0$ (recall that we assumed $\mathbb{E}[X] = 0$ above).
- (iii) Differentiating a second time, we have $\psi''_X(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}] \cdot \mathbb{E}[e^{\lambda X}] - \mathbb{E}[Xe^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2}$. To simplify this, note that for any function $g(x)$, we have

$$\begin{aligned}
\mathbb{E}[g(X)e^{\lambda X}] &= \int f_X(x) e^{\lambda x} g(x)\, dx \\
&= \int f_Z(x) e^{-\psi_X(\lambda)} g(x)\, dx \\
&= e^{-\psi_X(\lambda)} \int f_Z(x) g(x)\, dx \\
&= \mathbb{E}[e^{\lambda X}]\mathbb{E}[g(Z)],
\end{aligned}$$

  where we applied the definition of Z, factored out the constant $e^{-\psi_X(\lambda)}$, and substituted the definition of ψ_X. It follows that $\psi''_X(\lambda) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$, which is simply $\text{Var}[Z]$.

Claim: $\psi_X(\lambda) \leq \frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$ for any zero-mean random variable taking values in $[a, b]$. In other words, X is sub-Gaussian with parameter $\frac{(b-a)^2}{4}$.
Proof:

- By a (particular form of the) second-order Taylor expansion, we have

$$\psi_X(\lambda) = \psi_X(0) + \lambda \psi'_X(0) + \frac{\lambda^2}{2} \psi''_X(\theta)$$

  for some $\theta \in [0, \lambda]$.

- The claim is then immediate from the previous two claims upon noticing that by the definition of $f_Z(z) = e^{-\psi_X(\lambda)} e^{\lambda z} f_X(z)$, the random variable Z inherits X's property of taking values on $[a, b]$. (Outside that range, f_X is zero and hence so is f_Z.)