CSCI 1951-W Sublinear Algorithms for Big Data **Fall 2020**

Lecture 2: Concentration Inequalities

Lecturer: Jasper Lee *Scribe: Ang Li*

1 Concentration Behavior of Independent Random Processes

Intuition 2.1 Aggregate behavior of lots of random process is typically "well-behaved" (at least to some extent).

Consider n i.i.d. random variables $X_1, ..., X_n$ from a distribution D on $\mathbb{R}$ with $n >> 1$. Assume the distribution has finite mean μ and finite variance σ^2. The following are true about the mean $\bar{X}_n = \frac{1}{n}\sum_i X_i$:

- The Law of Large Numbers: $\bar{X}_n \to \mu$ almost surely as $n \to \infty$.
- The Central Limit Theorem: $\bar{X}_n \to N(\mu, \sigma^2/n)$ in distribution.

To concretely and operationally capture the law of large numbers, what we really want is a *concentration inequality*:

$$\mathbb{P}(\bar{X}_n \geq \epsilon) \leq \text{small}(\epsilon, n)$$

for some function "small".

A weak concentration inequality is Chebyshev's inequality.

Proposition 2.2 (Chebyshev's Inequality) *For a real valued random variable $Y \in \mathbb{R}$ with finite mean and variance, we have*

$$Pr(|Y - \mathbb{E}Y| \geq a) \leq \frac{\sigma^2(Y)}{a^2}.$$

Proof. This can be proved by applying the Markov's inequality to the positive random variable $(Y - \mathbb{E}Y)^2$. □

Corollary 2.3 *Applying the Chebyshev's inequality to the sample mean $\bar{X}_n = \frac{1}{n}\sum_i X_i$, we have*

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}X| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \tag{1}$$

where σ^2 is the variance of the i.i.d random variables X_i.

We can compare the bound derived from Chebyshev's inequality with the bound suggested by the CLT. Consider the Normal pdf:

$$\phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We can approximate the tail probability by

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}X| \geq \epsilon) \approx 2\int_\epsilon^\infty \phi(x; 0, \sigma^2/n) \approx C \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

Compared with (1), we can see that the bound derived from Normal approximation decays like $O(e^{-n})$, which is much faster than the $O(n^{-1})$ bound derived from Chebyshev's inequality.

To strengthen the inequality, one approach is to generalize Chebyshev's inequality by applying the Markov's inequality to the k-th moment $\mathbb{E}(|Y-\mathbb{E}Y|^k)$:

Proposition 2.4 *For a real-valued random variable $Y \in \mathbb{R}$ and $k \in \mathbb{N}$, we have*

$$\mathbb{P}(|Y-\mathbb{E}Y| \geq a) = \mathbb{P}(|Y-\mathbb{E}Y|^k \geq a^k) \leq \frac{\mathbb{E}(|Y-\mathbb{E}Y|^k)}{a^k}.$$

Since the above inequality is true for all $k \in \mathbb{N}$, we can combine them to get a stronger bound:

Corollary 2.5 *With the same setup as the above proposition, we have*

$$\mathbb{P}(|Y-\mathbb{E}Y| \geq a) \leq \inf_{k\in\mathbb{N}} \frac{\mathbb{E}(|Y-\mathbb{E}Y|^k)}{a^k}.$$

This particular corollary can be hard to apply though, because it is generally not clear how to bound the k^{th} moments in nice ways such that we can minimize the right hand side with respect to k both cleanly and tightly.

2 Chernoff Bounds

With the comparison between the bounds from Normal approximation and Chebyshev's inequality, it seems like we can improve the bounds on $\bar{X}_n$ by quite a bit if we believe in the Central Limit Theorem. We start by define the moment generating function of a random variable X

Definition 2.6 The moment generating function $M_X(t)$ of a random variable $X \in \mathbb{R}$ is defined as

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Remark 2.7 Using the standard Taylor expansion, we have

$$M_X(t) = \mathbb{E}(\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}) = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}(X^k).$$

Therefore, the moment generating function combines information from all moments of X. Thus, if instead of using the k^{th} moment to derive concentration, we used the moment generating function instead, this is another way to deriving concentration using all the available moment information about the underlying distribution.

2.1 Generic Chernoff bound

We introduce the generic Chernoff bound which we will use later for more specific cases:

Lemma 2.8 *For a real-valued random variable $Y \in \mathbb{R}$ with finite moment generating function (handwaving the formalization issues here a little), we have*

$$\begin{aligned}\mathbb{P}(Y \geq b) &\leq \inf_{t>0} e^{-tb}\mathbb{E}(e^{tY}) = \inf_{t>0} e^{-tb}M_Y(t)\\ \mathbb{P}(Y \leq b) &\leq \inf_{t<0} e^{-tb}\mathbb{E}(e^{tY}) = \inf_{t<0} e^{-tb}M_Y(t) = \inf_{t>0} e^{tb}M_Y(-t).\end{aligned}$$

Proof. Notice that the random variable e^{tY} is nonnegative and therefore we can apply the Markov's inequality e^{tY} for any $t > 0$:

$$\mathbb{P}(Y \geq b) = \mathbb{P}(e^{tY} \geq e^{tb}) \leq \frac{\mathbb{E}(e^{tY})}{e^{tb}}.$$

Since this is true for all $t > 0$, the inequality also holds for the inf over all possible values of t, which proves the first inequality. For the second one, we can use the fact that

$$\mathbb{P}(Y \leq b) = \mathbb{P}(e^{tY} \geq e^{tb}), \quad t < 0$$

and follow a similar argument.

□

2.2 Chernoff bound for Poisson trials

To apply the generic Chernoff bound to a random variable Y, we typically need the following two steps:

1. Compute or derive an upper bound of the moment generating function $M_Y(t)$
2. Compute the inf over proper ranges of t. In cases when the inf is hard to compute, we can also aim for a particular value of t that's good enough for the bound we need.

Next we derive a bound for a sum of Poisson trials:

Theorem 2.9 *Give n independent Bernoulli random variables $X_i \leftarrow \text{Bernoulli}(p_i)$, we define*

$$X = \sum_{i=1}^{n} X_i, \quad \mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i.$$

Given $\kappa > 0$, we have the following tail bound:

$$\mathbb{P}(X \geq (1+\kappa)\mu) \leq \left(\frac{e^{\kappa}}{(1+\kappa)^{(1+\kappa)}}\right)^{\mu} \tag{2}$$

Furthermore, if $\kappa \in (0, 1]$, we have

$$\mathbb{P}(X \geq (1+\kappa)\mu) \leq \exp\left(-\frac{\mu\kappa^2}{3}\right) \tag{3}$$

Proof. We follow the two-step process to prove the desired bound. We start by deriving an upper bound of the moment generating function $M_X(t)$:

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_i X_i}] = \prod_i \mathbb{E}[e^{tX_i}].$$

For each X_i, we have

$$\mathbb{E}[e^{tX_i}] = p_i \cdot e^t + (1 - p_i) \cdot 1 = 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)}.$$

Therefore

$$M_X(t) \leq \prod_i e^{p_i(e^t-1)} = e^{(\sum_i p_i(e^t-1))} = e^{\mu(e^t-1)}.$$

Using the generic Chernoff bound, we know

$$\mathbb{P}(X \geq (1+\kappa)\mu) \leq \inf_{t>0} e^{-t(1+\kappa)\mu} e^{\mu(e^t-1)} = \inf_{t>0} e^{\mu(e^t-1-(1+\kappa)t)}.$$

Consider the inf of the function $f(t) = e^t - 1 - (1+\kappa)t$, it's a straightforward computation to show that inf is achieved at

$$t_* = \log(1+\kappa).$$

Substitute t_* into the upper bound, we obtain the first inequality. To prove the second inequality, we just need to verify

$$\frac{e^\kappa}{(1+\kappa)^{1+\kappa}} \leq e^{-\kappa^2/3},$$

holds for all $\kappa \in (0,1]$, which can be done numerically. □

Using the generic Chernoff bound for lower tail case, we can obtain very similar result with slightly different bounds:

Theorem 2.10 *For $\kappa \in (0,1)$ and with the same setup as the previous theorem, we have*

$$\mathbb{P}(X \leq (1-\kappa)\mu) \leq \left(\frac{e^{-\kappa}}{(1-\kappa)^{(1-\kappa)}}\right)^\mu \tag{4}$$

$$\mathbb{P}(X \leq (1-\kappa)\mu) \leq \exp^{-\mu\kappa^2/2}. \tag{5}$$

Note the constant of 2 instead of 3 in the denominator of the exponent, in the second inequality.

Combine the bounds for the both lower and upper tails, we have a two-sided tail bound for Poisson trials:

Corollary 2.11 *For $\kappa \in (0,1)$ and with the same setup as the previous theorem, we have*

$$\mathbb{P}(|X-\mu| \geq \kappa\mu) \leq 2\exp\left(-\frac{\mu\kappa^2}{3}\right)$$

Sometimes, we don't know the exact probabilities p_i and as a result the expectation μ is not known either. However, we can extend the Chernoff bounds to cases when we can derive a lower or upper bound of the expectation μ:

Corollary 2.12 *Suppose $\mu \leq \mu_u \leq n$ for some known μ_u, we have*

$$\mathbb{P}(X \geq (1+\kappa)\mu_u) \leq \exp\left(-\frac{\mu_u\kappa^2}{3}\right)$$

This result generalizes to the other 3 inequalities in the above two theorems on Chernoff bounds for Poisson trials.

Proof. Given $X_i \sim \text{Bernoulli}(p_i)$, we can construct a new set of Poisson trials $Y_i \sim \text{Bernoulli}(q_i)$ such that

$$q_i \geq p_i, \quad \sum_i q_i = \mu_u.$$

Let $Y = \sum_i Y_i$. Since it's more likely to get 1 for each Y_i compared to X_i, we have

$$\mathbb{P}(X \geq (1+\kappa)\mu_u) \leq Pr(Y \geq (1+\kappa)\mu_u).$$

Notice that $\mathbb{E}Y = \mu_u$ and apply the Chernoff bound to Y will give the desired inequality. □

Remark 2.13 By using an upper bound instead of exact value, we have increased the tail from $(1+\kappa)\mu$ to $(1+\kappa)\mu_u$ but the resulting upper bound on the probability is also improved:

$$\exp\left(-\frac{\mu_u \kappa^2}{3}\right) \leq \exp\left(-\frac{\mu\kappa^2}{3}\right)$$

On the other hand, this observation does not hold for the lower tail, in that the probability *increases* (and the inequality loosens) if we only have a lower bound of μ.

2.3 Applications

One way to increase the performance of an existing algorithm is to run multiple times and try combine the results in some way. In particular, we are concerned with increasing the success probability of an algorithm. We can apply the Chernoff bounds to bound the number of repetitions.

Example 2.14 *Consider a decision algorithm $\mathcal{A}$ that makes the correct decision with probability $\geq 2/3$. To get an algorithm $\mathcal{A}'$ that makes the decision with probability $\geq 1-\delta$, we can repeat $\mathcal{A}$ for n times and return the majority decision. To decide a lower bound for n, we let E_i be the indicator variable of the event that "i-th run of $\mathcal{A}$ gets the wrong answer". We know*

$$E_i \leftarrow \text{Bernoulli}(p_i), \quad p_i \leq 1/3.$$

The algorithm $\mathcal{A}'$ ends up with the wrong decision if and only if the majority of the runs gives the wrong answer. Therefore

$$\mathbb{P}(\mathcal{A}' \textit{ fails}) = \mathbb{P}\left(\sum_{i=1}^{n} E_i \geq n/2\right) = \mathbb{P}\left(\sum_i E_i \geq \left(1+\frac{1}{2}\right)\frac{n}{3}\right).$$

Since $n/3$ is an upper bound for $\mathbb{E}[\sum_i E_i]$, we can use Corollary 2.12 and conclude

$$\mathbb{P}(\mathcal{A}' \textit{ fails}) \leq \exp\left(-\frac{n}{3}\left(\frac{1}{2}\right)^2\frac{1}{3}\right) = \exp(-\Theta(n))$$

We need $\exp(-\Theta(n)) \leq \delta$, which is true for $n = \Theta(\log\frac{1}{\delta})$.

Example 2.15 *Consider an estimation algorithm $\mathcal{B}$ which outputs an estimate $\hat{b}$ of the unknown parameter b. Suppose $\mathcal{B}$ outputs $\hat{b} \in [b-\epsilon_l, b+\epsilon_u]$ with probability $\geq 2/3$. We want to construct an algorithm $\mathcal{B}'$ that returns an estimate $b' \in [b-\epsilon_l, b+\epsilon_u]$ with probability $\geq 1-\delta$.*

We can run the algorithm $\mathcal{B}$ for n times and obtain n estimates $\hat{b}_i$ and try combine them into a single estimate b'. One obvious option is to take the mean, but this may not be the best idea because we don't know the behavior of $\mathcal{B}$ outside the unknown interval $[b-\epsilon_l, b+\epsilon_u]$. For instance, it's possible that $\mathcal{B}$ outputs $\pm\infty$ with a nontrivial probability which is $\leq 1/3$ and in that case, the mean will very likely be $\pm\infty$.
A better choice is to take the median of the n estimates:

$$b' = \text{median}\{\hat{b}_1, ..., \hat{b}_n\},$$

which is more robust to outliers. To bound the number of iterations n, let E_i^l be the indicator random variable that $\hat{b}_i < b - \epsilon_l$. We know $E_i^l \leftarrow \text{Bernoulli}(p_i)$ with $p_i \leq 1/3$. We have

$$\mathbb{P}(b' < b - \epsilon_l) = \mathbb{P}\left(\sum_{i=1}^{n} E_i^l \geq \frac{n}{2}\right)$$

Using the same argument as in the previous example, we can see that $n = \Theta(\log \frac{1}{\delta})$ is enough to achieve

$$\mathbb{P}(b' < b - \epsilon_l) < \delta/2.$$

Using a similar argument for the upper bound $\mathbb{P}(b' > b + \epsilon_u)$, we can show it's enough to have $n = \Theta(\log \frac{1}{\delta})$ overall.

Example 3 A similar construction can also work for optimization algorithms. Suppose we have an approximation algorithm $\mathcal{C}$ that returns a valid answer that is at least $C - \epsilon$ with probability $\geq 2/3$, but never more than the optimum C of the optimization problem at hand. By repeating $\mathcal{C}$ for $n = \Theta(\log \frac{1}{\delta})$ times and returning the maximum answer, we can construct an algorithm $\mathcal{C}'$ that outputs answer $\geq C - \epsilon$ with probability $\geq 1 - \delta$.

Remark 2.16 One general heuristic is that we should always design algorithms that has query/time complexity $q(\delta)$ such that

$$q(\delta) = O\left(q(1/3) \log \frac{1}{\delta}\right)$$

where δ in $q(\delta)$ is the success probability. If our algorithm has worse than log dependence on δ, then we can typically use one of the above tricks to boost the constant probability version of our algorithm into a high probability one with only a log dependence.