Lecture 3: February 18

Scribe: Athanasios Filippidis, Konstantinos Sotiropoulos *Tail Bounds*

Central Limit Theorem

Theorem 3.1 (Lindeberg-Lévy Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ sequence of i.i.d. random variables each one having expected value μ and variance σ^2. We consider the sum*

$$S_n = X_1 + X_2 + \ldots, X_n.$$

Notice that $\mathbb{E}[S_n] = n\mu, \mathbb{V}ar[S_n] = n\sigma^2$. We normalize S_n so as to obtain a random variable with zero mean and unit variance as follows

$$Z_n := \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathbb{V}ar[S_n]}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Then, as $n \to +\infty$,

$$Z_n \to N(0,1) \text{ in distribution.}$$

Reminder. What does the latter statement, i.e., *convergence in distribution* mean? It means that the CDF of Z_n in the limit of $n \to +\infty$ is equal to the CDF of a normal random variable. Specifically,

$$\mathbf{Pr}[Z_n \geq t] \to \mathbf{Pr}[g \geq t] = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{\frac{-x^2}{2}} dx$$

where g is a random variable drawn from the normal distribution $g \sim N(0,1)$.

Moivre-Laplace theorem, (aka CLT for Binomials). The special case of the CLT when X_is are Bernoulli random variables is particularly relevant to our class, as we will frequently focus on sums of Bernoulli random variables, i.e., binomial RVs (e.g., X_i could be an indicator random variable of the existence of the i-th edge, the sum corresponds to the total number of edges). Due to its importance, despite being a corollary of the CLT, we state it as a theorem.

Theorem 3.2 (Moivre-Laplace theorem)**.** *Let $X_i \sim Ber(p)$, $i = 1, 2, \ldots$ i.i.d. Bernoulli variables and $p \in (0,1)$ is fixed. Again, let*

$$S_n = X_1 + X_2 + \ldots, X_n \sim Binomial(n,p).$$

and S_n the sum of them.

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to N(0,1) \text{ in distribution.}$$

Remark. While the exact computation of the above probabilities is not easy, the exponential decay is what determines the probability. The following double inequality due to Mill's captures this exponential decay

$$(\frac{1}{t}-\frac{1}{t^3})\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \leq \mathbf{Pr}\left[g \geq t\right] \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

Coin Tossing

Let's say we toss a fair coin n times. We want to get a bound on the probability that we see heads at least $\frac{3n}{4}$ times. Let X_i an indicator random variable that takes the value 1, if i^{th} coin toss ends up heads, and 0 otherwise. As before, we denote let $S_n = X_1 + \ldots + X_n$ be the number of heads. By the linearity of expectation, and the independence of the tosses, we have

$$\mathbb{E}\left[S_n\right] = \frac{n}{2}, \mathbb{V}ar\left[S_n\right] = \frac{n}{4}.$$

Let's see how we can bound the $\mathbf{Pr}\left[S_n \geq \frac{3n}{4}\right]$. If we apply Markov's inequality we obtain

$$\mathbf{Pr}\left[S_n \geq \frac{3n}{4}\right] \leq \frac{\mathbb{E}\left[S_n\right]}{\frac{3n}{4}} = \frac{\frac{n}{2}}{\frac{3n}{4}} = \frac{2}{3}.$$

Using Chebyshev's inequality, we can get a better but still much weaker bound than what CLT **hints**.

$$\mathbf{Pr}\left[S_n \geq \frac{3n}{4}\right] = \mathbf{Pr}\left[S_n - \frac{n}{2} \geq \frac{n}{4}\right] \leq \mathbf{Pr}\left[|S_n - E[S_n]| \geq \frac{n}{4}\right] \leq \frac{\mathbb{V}ar\left[S_n\right]}{(\frac{n}{4})^2} = \frac{4}{n}.$$

By hinting, we mean the following. Let's pretend we could use the Moivre-Laplace approximation, the variable $Z_n = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$. Therefore,

$$\mathbf{Pr}\left[S_n \geq \frac{3n}{4}\right] = \mathbf{Pr}\left[S_n - \frac{n}{2} \geq \frac{n}{4}\right] = \mathbf{Pr}\left[\frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \geq \sqrt{\frac{n}{4}}\right] \approx \mathbf{Pr}\left[N(0,1) \geq \sqrt{\frac{n}{4}}\right] \leq \frac{1}{\sqrt{\frac{n}{4}}\sqrt{2\pi}}e^{-\frac{n}{8}} \leq e^{-\frac{n}{8}}.$$

This is great news! The decay appears to be exponential assuming the CLT approximation. Can we conclude that by using this informal approximation we get exponential decay? The answer is no! Why? This approximation ignores the error of approximation. And unfortunately the Berry-Esseen theorem states that the error can be $\frac{1}{\text{poly}(n)}$ in CLT. This much larger than the exponential term, so we cannot use this argument to conclude the exponential decay.

Theorem 3.3 (Berry - Esseen Theorem)**.** *Consider the same setting as in Theorem 3.1. For all n and for all t:*

$$|\mathbf{Pr}\left[Z_n \geq t\right] - \mathbf{Pr}\left[N(0,1) \geq t\right]| \leq \frac{\rho}{\sqrt{n}}$$

where $\rho = \frac{\mathbb{E}\left[|X_1-\mu|^3\right]}{\sigma^3}$.

Perhaps we can improve the RHS (right-hand side) of Theorem 3.3? The answer is negative as the following example shows. If S_n is the number of heads in n tosses of a fair coin, then $\mathbf{Pr}\left[S_n = \frac{n}{2}\right] = \binom{n}{\frac{n}{2}} 2^{-n} \approx \frac{1}{\sqrt{n}}$[1]. At the same time $\mathbf{Pr}\left[S_n = \frac{n}{2}\right] = \mathbf{Pr}\left[Z_n = 0\right]$. Since $N(0,1)$ is continuous, in the continuous regime this probability is 0, which shows that indeed we can have an error approximation of order $O(\frac{1}{\sqrt{n}})$. For this reason we need to develop a different approach to derive the exponential decay. This is described in the next Section.

The Exponential Method

Hoeffding's inequalities

Definition 3.4 (Radamacher random variable)**.** *We call a random variable X symmetric Bernoulli or Radamacher, if*

$$\mathbf{Pr}\left[X = 1\right] = \mathbf{Pr}\left[X = -1\right] = \frac{1}{2}.$$

Notice that if: $X \sim Ber(\frac{1}{2})$, then $2X - 1$ is symmetric Bernoulli and vice versa.

Theorem 3.5 (Hoeffding's Inequality for Radamacher random variables)**.** *Let $a \in R^n$ and $X_1, X_2, \ldots, X_n$ symmetric Bernoulli variables. Then,*

$$\mathbf{Pr}\left[\sum a_i X_i \geq t\right] \leq e^{-\frac{t^2}{2||\alpha||_2^2}}.$$

Proof. Assume w.l.o.g. that $||\alpha||_2^2 = 1$[2], and let $\lambda > 0$ be a positive parameter that we will decide later.

$$\mathbf{Pr}\left[\sum a_i x_i \geq t\right] = \mathbf{Pr}\left[\lambda \sum a_i x_i \geq \lambda t\right] = \mathbf{Pr}\left[e^{\lambda \sum a_i x_i} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda \sum a_i x_i}\right]}{e^{\lambda t}}$$

where the last inequality comes from Markov's inequality. We bound the nominator as follows by using the independence and the fact that the X_i.

$$\mathbb{E}\left[e^{\lambda \sum a_i X_i}\right] = \prod_i \mathbb{E}\left[e^{\lambda a_i X_i}\right] = \prod_i \frac{e^{\lambda \alpha_i} + e^{-\lambda \alpha_i}}{2} = \prod_i \cosh(\lambda \alpha_i).$$

We upper bound the RHS using the fact that $\cosh(x) \leq e^{x^2/2}$. To see why this is the case, notice that for $x \geq 0$, $\ln(\cosh(x)) = \int_0^x tanh(t)dt$ and $\tanh(t) \leq t$, so $\ln(\cosh(x)) \leq \int_0^x t dt = \frac{x^2}{2}$. Exponentiating we get the desired inequality for $x \geq 0$. The case for $x \leq 0$ is the same due to symmetry. Using this fact, we get the following upper bound for the nominator,

$$\prod_i \cosh(\lambda \alpha_i) \leq \prod_i e^{\frac{\lambda^2 \alpha_i^2}{2}} = e^{\frac{\lambda^2}{2}}.$$

[1] Apply Stirling's formula for the factorial $n! \approx n^{n+\frac{1}{2}} \cdot e^{-n}$

[2] Notice that we can just rescale $\vec{a}$.

Putting everything together, and by setting $\lambda = t$ we obtain the desired upper bound of $e^{-t^2/2}$. □

Our proof technique applies for the symmetric case. Specifically, we obtain the following two-sided Hoeffding bound stated as the next corollary:

Corollary 3.6 (Two-sided Hoeffding's Inequality for Radamacher random variables)**.** *Assume the same setting as in Theorem 3.5. Then,*

$$\mathbf{Pr}\left[|\sum a_i X_i| \geq t\right] \leq 2\exp\big(-\frac{t^2}{2||a||_2^2}\big).$$

Hoeffding's inequality also applies to bounded random variables.

Theorem 3.7 (Hoeffding's Inequality for bounded random variables)**.** *Let X_i i.i.d. random variables, where $X_i \in [m_i, M_i]$ for $i = 1, \ldots, n$. Then,*

$$\mathbf{Pr}\left[|\sum_{i=1}^{n}[X_i - \mathbb{E}[X_i]]| \geq t\right] \leq 2e^{\frac{-t^2}{\sum_{i=1}^{n}(M_i - m_i)^2}}$$

The proof is based on Hoeffding's lemma and the exponential method as we used in the proof of Theorem 3.5. We state this lemma next without proof. For the proof see Wikipedia article.

Lemma 3.8 (Hoeffding's lemma)**.** *If $\mathbb{E}[X] = n$ and $a \leq X \leq b$, then*

$$\mathbb{E}\left[e^{\lambda x}\right] \leq \exp\big(\frac{(b-a)^2}{8}\lambda^2 + \lambda n\big)$$

Chernoff bounds

It is reasonable to ask if we can use more information that we have available for X_is except only for the fact that are bounded. Indeed Chernoff bounds are better than Hoeffding's bound when p_is is small.

Theorem 3.9. *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with parameters p_i $i = 1, \ldots, n$. Again, let $S_n = X_1 + X_2 + \ldots + X_n$, and the expected value of this sum: $\mathbb{E}[S_n] = \mu = p_1 + \ldots p_n$. Then for any $t > m$ we have*

$$\mathbf{Pr}[S_n \geq t] \leq e^{-\mu}(\frac{e\mu}{t})^t.$$

Proof. As before, we introduce $\lambda > 0$, and apply Markov's inequality once we exponentiate:

$$\mathbf{Pr}[S_n \geq t] \overset{\lambda \geq 0}{=} \mathbf{Pr}\left[e^{\lambda S_n} \geq e^{\lambda t}\right] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda S_n}\right]$$

where the last inequality comes from the Markov Inequality. Now,

$$\mathbb{E}\left[e^{\lambda S_n}\right] = \mathbb{E}\left[e^{\lambda}(X_1 + \ldots + X_n)\right] = \prod_i \mathbb{E}\left[e^{\lambda x_i}\right] = \prod_i (p_i e^{\lambda} + (1 - p_i)) \leq \prod_i e^{p_i(e^{\lambda}-1)} = e^{\sum_i p_i(e^{\lambda}-1)} = e^{(e^{\lambda}-1)\mu}$$

where in the last inequality we used the well-known inequality: $1 + x \leq e^x$ for all real x.

So, $\mathbf{Pr}[S_n \leq t] \leq e^{-\lambda t + \mu(e^\lambda - 1)}$. Choosing $\lambda = \ln(\frac{t}{\mu})$ gives the desired upper bound. □

Chernoff bounds refer to a family of bounds. Depending on what one's to prove, some forms may be more convenient than others. The following forms of bounds are useful frequently in practice. endly.

Theorem 3.10 (Multiplicative Chernoff bounds)**.** *Let $X_1, \ldots, X_n$ be independent Poisson trials such that $\mathbf{Pr}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then, for any $\delta > 0$:*

- $\mathbf{Pr}[X \geq (1+\delta)\mu] \leq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^\mu$
- $\mathbf{Pr}[X \leq (1-\delta)\mu] \leq (\frac{e^{-\delta}}{(1-\delta)^{1-\delta}})^\mu$

For $0 < \delta < 1$,

- $\mathbf{Pr}[X \geq (1+\delta)\mu] \leq e^{\frac{-\delta^2\mu}{3}}$
- $\mathbf{Pr}[X \leq (1-\delta)\mu] \leq e^{\frac{-\delta^2\mu}{2}}$,

Combining the above two, we can get the following convenient two-sided Chernoff bound.

Corollary 3.11. *Let $X_1, \ldots, X_n$ independent Poisson random trials, such that $\mathbf{Pr}[X_i] = p_i$. Let $X = \sum_i X_i$ and $\mu = \mathbb{E}[X]$. For $0 < \delta < 1$,*

$$\mathbf{Pr}[|X - \mu| \geq \delta\mu] \leq 2 \cdot e^{-\mu\delta^2/3}$$

Application: Analyzing the degree sequence of $G(n,p)$

We want to apply Chernoff bounds to understand the degree sequence in a $G(n,p)$ graph, where p is well-above the connectivity threshold, i.e., let $p = \frac{\omega(n)logn}{n}$ where $\omega(n) \to +\infty$ is any slowly growing function (e.g., $\omega(n) = o(\log n)$). Of course the degrees are random variables, but as we will see they are extremely concentrated around the mean. Specifically, with high probability the lowest $\delta(G)$ and the highest degree $\Delta(G)$ are asymptotically equal. This means that they are both equal to $\delta(G) = \Delta(G) = np + o(np), G \sim G(n,p)$. To do so, we use a union bound combined with a Chernoff bound, to prove that the probability of the existence of a node deviating from np by $\frac{np}{\omega(n)^{\frac{1}{3}})}$ is negligible.

$$\begin{aligned}\mathbf{Pr}[\exists v \in [n] : |\deg(v) - np| \geq \epsilon \cdot np] &\leq n\mathbf{Pr}[|\deg(v) - np| \geq \epsilon np] \leq 2ne^{\frac{-\epsilon^2\omega(n)log(n)}{3}} \\ &= 2\exp\left(\log n - \frac{1}{3}\omega(n)^{\frac{1}{3}}\log n\right) = o(1).\end{aligned}$$

Degree sequence in Preferential Attachment (PA) graphs

We will show a heuristic argument, based on mean-field approximation to show that the preferential attachment graphs we defined previously in lecture 1. Specifically, we will analyze the following preferential attachment model. The mechanism assumes that nodes arrive sequentially at discrete time-stamps:

- Initially we start with a single node that is connected to itself (self-loop)
- At timestamp $t+1$, node $t+1$ arrives and forms a new connection with an existing node, as follows:
  a) With probability p chooses a node uniformly at random (u.a.r) among the t previous nodes
  b) With probability $1-p$ chooses an edge u.a.r. and then u.a.r. one of the two endpoints of the edge.
  Intuitively, this is the step where the rich-get-richer phenomenon emerges. A high degree node has more edges incident to it, hence it is more likely to be chosen

Proving that the resulting network of this probabilistic procedure follows a power-law degree distribution is a challenging task, see [1] for a formal proof. Instead, we will follow a heuristic used frequently by physicists to see why we expect to see a power law degree distribution. We approximate "discrete" using "continuous", and discrete differences using derivatives. E.g., for a discrete quantity $Y(t)$ we approximate the difference $Y(t+1)-Y(t)$ as

$$\frac{Y(t+1)-Y(t)}{(t+1)-t} \approx \frac{dY(t)}{dt}.$$

Let $X_j(t)$ be the expected degree of node j at timestamp t. The expected degree of this node at $t+1$ satisfies the following:

$$X_j(t+1) = X_j(t) + \frac{p}{t} + (1-p)\frac{X_j(t)}{t}$$
$$\frac{dX_j(t)}{dt} = \frac{p}{t} + (1-p)\frac{X_j(t)}{t}$$

We divide both sides by $p+(1-p)X_j(t)$, integrate both sides. This results in

$$\frac{1}{p+(1-p)X_j(t)}\frac{dX_j(t)}{dt} = \frac{1}{t}dt$$
$$\int \frac{1}{p+(1-p)X_j(t)}\frac{dX_j(t)}{dt}dt = \int \frac{1}{t}dt$$
$$\frac{1}{1-p}\ln(p+(1-p)X_j(t)) = \ln t + c$$

By exponentiating and using the initial condition $X_j(j)=0$ we obtain

$$X_j(t) = \frac{1}{1-p}[(\frac{t}{j})^{1-p} - 1] \tag{3.1}$$

Equation 3.1 gives us the degree of node j at a timestamp t. To see why the degree distribution follows a power-law, we need first to answer the following question:

For a given timestamp t, what is the fraction of nodes that have degree at least d?

Let's see for which j $X_j(t) \geq d$. By simple algebraic manipulation,

$$j \leq t \left[\frac{1-p}{p} d + 1 \right]^{\frac{1}{1-p}}.$$

Hence, the fraction of nodes that has degree at least d is $\left[\frac{1-p}{p} d + 1 \right]^{\frac{1}{1-p}}$. In order to find the probability that a node has degree exactly d, we need to take the derivative, w.r.t. d of the CDF which yields that this number is proportional to $d^{-1-\frac{1}{1-p}}$. Thefefore, we see a power law with slope $1 + \frac{1}{1-p}$.

Emergence of a Giant Component in Random Graphs

We went over the proof of Krivelevic and Sudakov [3] for the phase transition around $p = \frac{1}{n}$ for the emergence of a giant component. This was originally proved (with many more details, see [2]) in the *seminal* work of Erdős and Rényi [2]. In class we proved the following theorem.

For a random graph $G(n,p)$ and any constant $\epsilon > 0$:

1. if $p = \frac{1-\epsilon}{n}$, all connected components of $G(n,p)$ have size at most $\frac{7}{\epsilon^2} \log n$.
2. if $p = \frac{1+\epsilon}{n}$, a (unique) giant component of size linear in n.

The Professor suggested omitting the proof from the scribe, and instead read the paper [3] The proof is simple and elegant. It uses which uses DFS as its main routine. DFS maintains three sets of vertices: S the set of **explored** vertices, U the vertices in the **stack**, and $T = V \setminus (S \cup T)$ the set of vertices that are neither explored, nor in the stack. Also, DFS prioritizes vertices according to a permutation σ. For the sake of this proof, we assume that $\sigma = (1, 2, \ldots, n)$ is the identity permutation. The DFS algorithm is given as input a sequence of i.i.d. *Bernoulli*(p) random variables $\overline{X} = (X_i)_{i=1}^{\binom{n}{2}}$. It gets its $i-th$ query answered positively only if $X_i = 1$. Also, the following observations about DFS will be useful:

- At each iteration, only one vertex moves, either from T to U, or from U to S.
- At each iteration, there are no edges between S and T
- Vertices in the stack U form a path
- As long as $T \neq 0$, we have $|S \cup U| \geq \sum_{i=1}^{t} X_i$
- The addition of every new node to a connected component is caused by a positive answer to a query, thus at time t: $|U| \leq 1 + \sum_{i=1}^{t} X_i$

The following lemma will be also proved to be useful:

Lemma 3.12. *Suppose I have a sequence of $\binom{n}{2}$ Bernoulli Variables $X_1, X_2, \ldots, X_{\binom{n}{2}}$ Bernoulli variables with parameter $\frac{1-\epsilon}{n}$. Let $k = \frac{7}{\epsilon^2} \ln n$. Then, w.h.p. there is no interval I of $k \cdot n$ variables with at least k ones.*

Proof. Using a union bound to upper-limit the probability that at least one interval of length $k \cdot n$ has at least k ones and a standard Chernoff-type bound, we get that:

$$\Pr[\exists \text{I: I has at least k ones}] \leq \binom{n}{2} \cdot \Pr[\sum_{i=1}^{kn} X_i \geq k] \leq n^2 \cdot e^{-\frac{\epsilon^2}{3}(1-\epsilon)k} = o(1)$$

□

This lemma immediately gives the first claim of the theorem. By virtue of contradiction, assume there is a connected component C of size more than k. Then, the time the algorithm discovered the $k+1$ node of the connected component and is about to move it to stack U, we have that $|\Delta S \cup U| = k$, where $\Delta S = S \cap C$. This means that we have exactly k positive answers. As long as we query only pairs of nodes with one side of the pair being one of the k nodes in $\Delta S \cup U$, we have not performed more than kn queries. Therefore, the sequence $\overline{X}$ contains an interval of length at most kn with at least k ones – a contradiction to Lemma 1.

References

[1] Bollobás, Béla and Riordan, Oliver and Spencer, Joel and Tusnády, Gábor *The degree sequence of a scale-free random graph process* Random Structures & Algorithms, vol. 18(3), pages 279-290, 2001

[2] Erdős, Paul and Rényi, Alfréd *On the evolution of random graphs.* Publ. Math. Inst. Hung. Acad. Sci, vol. 5(1), pages 17-60, 1960

[3] Krivelevich, Michael and Sudakov, Benny *The phase transition in random graphs: A simple proof.* Random Structures & Algorithms, vol. 43(2), pages 131–138, 2013