

# MAT377 Lecture Notes

ARKY!! :3C

'24 Fall Semester

## Contents

1	Day 1: Introduction to Probability (Sep. 4, 2024)	2
2	Day 2: Expectations and Distributions (Sep. 9, 2024)	4
3	Day 3: Distributions, Stability Property, Moments (Sep. 11, 2024)	9
4	Day 4: Independence and Dependence (Sep. 16, 2024)	11
6	Day 6: Ulam's Problem; Chebyshev Inequality, Stirling Approximation, and Erdős-Renyi Random Graphs (Sep. 23, 2024)	14
10	Day 10: Erdős-Renyi Random Graphs and Cliques, Chebyshev Inequality, Moment Generating Function (Oct. 7, 2024)	17
12	Day 12: Inequalities (Oct. 16, 2024)	20

## §1 Day 1: Introduction to Probability (Sep. 4, 2024)

Link to [textbook](#).

We start with a sampler problem that on the surface, seems unrelated to probability. Let  $v_1, \dots, v_n \in \mathbb{R}^n$  be unit vectors on the unit sphere, i.e.t  $\|v_i\| = 1$ . If we are to pick  $\varepsilon_i = \{-1, 1\}$  at random, what is our expectation on how large will

$$\sum_{i=1}^n \varepsilon_i v_i$$

be? We could brute force and average out over all probabilities as follows,

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \left( \sum_{i=1}^n \varepsilon_i v_i \right) = \sum_{i=1}^n \left( \frac{1}{2^n} \underbrace{\sum_{\varepsilon \in \{-1, 1\}^n} \varepsilon_i}_{=0} \right) v_i = 0.$$

Now, consider that

$$\begin{aligned} \frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \left| \sum_{i=1}^n \varepsilon_i v_i \right|^2 &= \frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \sum_{i_1, i_2=1}^n \varepsilon_{i_1} \varepsilon_{i_2} \langle v_{i_1}, v_{i_2} \rangle \\ &= \sum_{i_1, i_2=1}^n \left( \frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \varepsilon_{i_1} \varepsilon_{i_2} \right) \langle v_{i_1}, v_{i_2} \rangle \end{aligned}$$

To simplify the bracketed summation, we could consider the following two cases:

- If  $i_1 \neq i_2$ , we would have that

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \varepsilon_{i_1} \varepsilon_{i_2} = \frac{2^{n-2}}{2^n} \sum_{\substack{\varepsilon_{i_1} \in \{-1, 1\} \\ \varepsilon_{i_2} \in \{-1, 1\}}} \varepsilon_{i_1} \varepsilon_{i_2} = 0.$$

- If  $i_1 = i_2$ , we would have

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \varepsilon_{i_1} \varepsilon_{i_2} = \frac{2^{n-1}}{2^n} \sum_{\varepsilon_i \in \{-1, 1\}} \varepsilon_i \varepsilon_i = 1.$$

By linearity of expectation, we obtain

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \left| \sum_{i=1}^n \varepsilon_i v_i \right|^2 = n,$$

and

$$\frac{1}{2^n} \sum_{\varepsilon \in \{-1, 1\}^n} \left| \sum_{i=1}^n \varepsilon_i v_i \right| = c\sqrt{n},$$

where  $c$  is a positive real constant.

We now abstract a few probability terms;

- $\Omega$  is a sample space, i.e. the set of possible outcomes.
- Let  $P$  denote probability, i.e. a mapping of subsets of  $\Omega$  to  $[0, 1]$  (read: probability of getting these subsets of  $\Omega$ ); the probability of an event  $\varepsilon$  out of  $S$  occurring is given by  $P(\varepsilon \in S, S \subset \{-1, 1\}^n) = \frac{1}{|S|}$ , assuming that each event in  $S$  is equally likely. With this, we have three important properties of  $P$  to define:
  1.  $P(\Omega) = 1$ ; the chance of an event in the probability space happening is 1.
  2. Let  $\mathcal{F}$  be a collection of subsets  $A_1, \dots, A_n$ . Then

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i),$$

given that  $A_i \cap A_j = \emptyset$  for all  $1 \leq i, j \leq n$ . This is linearity of expectation.

3.  $P(A^C) = 1 - P(A)$ , which is a property of set complement.
- When our collection  $\mathcal{F}$  of subsets of  $\Omega$  satisfy the following properties, we call it a  $\sigma$ -algebra:
    1.  $\emptyset \in \mathcal{F}$ ,
    2. Closed under countable union:  $A_n \in \mathcal{F} \implies \bigcup_{i=1}^n A_i \in \mathcal{F}$ ,
    3. Closed under complement:  $A_i \in \mathcal{F} \implies A_i^C \in \mathcal{F}$ .

In a finite sample space, the power set  $\mathcal{F} = \mathcal{P}(\Omega)$  is one such example of a  $\sigma$ -algebra. As an example, let  $\Omega = [0, 1)$ ; then  $P([a, b]) = b - a$  (wlog, let  $a < b$ ). Now, let  $\mathcal{F} = \mathcal{P}([0, 1))$ . Define the equivalence  $x \sim y$  if  $x - y \in \mathbb{Q}$ .

- We now introduce the axiom of choice; Let  $A$  be a set containing one element of each equivalence class from the above defined equivalence. Consider  $\tau_q A := \{A\} + q$ ; let us claim that

$$\bigcup_{q \in \mathbb{Q}} \tau_q A = [0, 1),$$

which is a countable union of  $[0, 1)$ , since  $\mathbb{Q}$  is countable. We have that  $P(A) = P(\tau_q A)$  because intervals don't change size under shifting by  $q$ . However, observe that

$$P([0, 1)) = P\left(\bigcup_{q \in \mathbb{Q}} \tau_q A\right) = \sum_{q \in \mathbb{Q}} P(\tau_q A).$$

Then either

$$\begin{aligned} P(A) = 0 &\implies P(\tau_q A) = 0 \implies P([0, 1)) = 0, \text{ or} \\ P(A) \neq 0 &\implies P(\tau_q A) \rightarrow \infty \implies P([0, 1)) \rightarrow \infty, \end{aligned}$$

which doesn't make sense (for now). This shows that we need to pick our  $\sigma$ -algebra properly; observing that the intersection of two  $\sigma$ -algebras is also a  $\sigma$ -algebra, it is appropriate to let  $\mathcal{F}$  be the smallest  $\sigma$ -algebra containing  $[a, b]$ . This is called a *Borel Set*.<sup>1</sup>

---

<sup>1</sup>this is confusing. ill check later

## §2 Day 2: Expectations and Distributions (Sep. 9, 2024)

Course administrative details first; starting next week, office hours will be held on Monday from 11:15am to 12:15pm. Recap of last lecture:

- A probability space  $\Omega$  is the set of all possible outcomes of an “experiment,” i.e. a countable set of individual events  $\{\omega_1, \dots, \omega_n\}$  (we will cover continuous probability later on).
- $\mathcal{F} = \mathcal{P}(\Omega)$  is the set of all subsets of  $\Omega$ .
- $P(A) = \sum_{\omega \in \Omega} P(\omega)$  is the probability of an outcome in  $A \in \mathcal{F}$  occurring.

A random variable  $X$  is a function  $\Omega \rightarrow \mathbb{R}$ , aka the measurement of the event, and the expectation of the random variable,  $EX$ , is given by  $\sum_{\omega \in \Omega} X(\omega)P(\omega)$ . Is expectation well behaved? No. For example, consider the St. Petersburg Paradox; suppose you are playing a game in the casino; every time you flip a coin, your prize money doubles if it lands on heads (read: double or nothing lfg!!!). Then we may consider the set of outcomes to be the number of consecutive heads, i.e.

$$\begin{aligned}\Omega &= \{1, 2, 3, \dots\}, \\ P(n) &= \frac{1}{2^n}, \\ X(n) &= 2^n.\end{aligned}$$

Clearly, the chance of getting  $n$  heads in a row is  $2^{-n}$ , and assuming your prize money started at 1 dollar, you would win  $2^n$  dollars for said  $n$  heads. Taking the expectation of this game, we find

$$EX = \sum_{n=1}^{\infty} 2^n \cdot \frac{1}{2^n} = \sum 1 = \infty.$$

It doesn't make sense to expect to win infinite amounts of money from this game<sup>2</sup> unless you had unlimited wealth to start with. With this in mind, we insist on

$$\sum_{\omega \in \Omega} |X(\omega)| P(\omega) < \infty$$

within the context of this class.

**Theorem 2.1** (Linearity of Expectation).  $X$  is linear; i.e.,  $E[ax + by] = aEx + bEy$ .

We start with a lemma:

**Lemma 2.2.** Let us have a bijective map  $\pi : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ . Then

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} C_{\pi(n,m)}$$

if all  $c_n \geq 0$  or if either side is absolutely convergent.

---

<sup>2</sup>martingale strat lfg,..

We proceed to prove this with casework.

- Suppose  $c_n \geq 0$ ; then using the bijective nature of  $\pi$ , we may choose large enough  $N, M$  such that

$$\sum_{n=1}^K c_n \leq \sum_{n=1}^N \sum_{m=1}^M c_{\pi(n,m)}$$

for any choice of  $K$ . Conversely, we may pick

$$\sum_{n=1}^N \sum_{m=1}^M c_{\pi(n,m)} \leq \sum_{n=1}^K c_n$$

for any  $N, M$  by picking  $k \geq \max_{1 \leq n \leq N} \{\pi(n, m)\}$ . Now, let  $M \rightarrow \infty$ ; we have

$$\sum_{n=1}^N \sum_{m=1}^{\infty} c_{\pi(n,m)} \leq \sum_{n=1}^{\infty} c_n,$$

then let  $N \rightarrow \infty$  to get

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} c_{\pi(n,m)} \leq \sum_{n=1}^{\infty} c_n.$$

As per earlier, we also see that LHS is greater or equal to RHS, which implies equality.  $\square$

- Now, suppose  $\sum_{n=1}^{\infty} |c_n| < \infty$ . Let  $c_n = a_n - b_n$ , where  $a_n = c_n 1(c_n \geq 0)$  and  $b_n = c_n 1(c_n < 0)$ . Then we obtain

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{\pi(n,m)}, \quad \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} b_{\pi(n,m)}$$

as per our proof above. Summing both, we conclude that equality holds for absolute convergence as well.  $\square$

For now, let  $X$  take values  $\{a_1, a_2, \dots\}$  (countably many). Consider

$$P'(a_n) = P(X = a_n) = P(\underbrace{\{\omega \mid X(\omega) = a_n\}}_{X^{-1}(a_n)})$$

as the probability of a pre-image (or,  $P' = P \circ X$ ). We see that  $P'$  is a probability on  $\mathbb{R}$  (concentrated on  $\{a_1, a_2, \dots\}$ ), and  $0 \leq P'(a_n) \leq 1$  for any  $n$ ; from now, we will call  $P'$  a *distribution* of  $X$ . Here are some examples of distributions:

- The Bernoulli distribution: let  $0 \leq p \leq 1$ . Then consider a coin with  $p$  chance to land on heads, and  $1 - p$  on tails; then  $Ber_p$  is given by  $\Omega = \{H, T\}$ ,

$$\begin{aligned} X(H) &= 1, P(X = 1) = p, \\ X(T) &= 0, P(X = 0) = 1 - p. \end{aligned}$$

- Flip  $N$  coins, with  $X = \{0, 1, \dots, N\}$  being the number of heads we obtain. Then

$$P(X = \ell) = \binom{N}{\ell} p^{\ell} (1 - p)^{N - \ell},$$

and the expected value is given by

$$EX = \sum_{\ell=0}^N \ell \binom{N}{\ell} p^\ell (1-p)^{N-\ell}.$$

Using linearity of expectation, we see  $EX = EX_1 + \dots + EX_n = Np$  by separating each coinflip.

Expectation enjoys the change of variables property;<sup>3</sup>

$$EX = \sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{n=1}^{\infty} a_n P'(a_n).$$

To see this, consider partitioning the probability space  $\Omega$  into  $X^{-1} = \{\omega_{nm} \mid 1 \leq m \leq M_n\}$  in terms of their measurement from  $X$  (where  $X(\omega_{ni}) = X(\omega_{nj}) = a_n$  for any  $1 \leq i, j \leq M_n$ )<sup>4</sup>, and write

$$\sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{n=1}^{\infty} \sum_{m=1}^{M_n} X(\omega_{nm})P(\omega_{nm}), \quad (\text{by Lemma})$$

where we may note that mapping each individual  $\omega \in \Omega$  to some index  $nm$  is bijective since it is a partition. We continue by writing

$$\begin{aligned} &= \sum_{n=1}^{\infty} \sum_{m=1}^{M_n} a_n P(\omega_{nm}) = \sum_{n=1}^{\infty} a_n \left( \sum_{m=1}^{M_n} P(\omega_{nm}) \right) \\ &= \sum_{n=1}^{\infty} a_n P(X = a_n), \end{aligned}$$

where we may note  $P(X = a_n) = P'(a_n)$ . □

The probability distribution of any given random variable  $X$  also approaches 0 at its tail. Specifically, we have that  $\lim_{t \rightarrow \infty} P(x \geq t) = 0$ . To prove this, we start by observing that  $P(x \geq t)$  is monotone decreasing; consider

$$P(X \geq n) = \sum_{m=n}^{\infty} P(m \leq X < m+1).$$

Clearly, the sum is convergent, as the sum of probabilities is equal to 1. Using the fact that the tail of a convergent series approaches 0, we conclude that  $P(m \leq X < m+1) \rightarrow 0$  as  $m \rightarrow \infty$ , and so  $P(X \geq n) \rightarrow 0$  as  $n \rightarrow \infty$ .

<sup>3</sup>read: sum of value of outcome multiplied by the chance it occurs over all  $\omega$  is the same as going over each value individually and multiplying the chance you roll into it

<sup>4</sup>read  $M_n$  as a counter of how many outcomes in  $\Omega$  have the same measurement of  $a_n$

**Lemma 2.3** (Expectation of Random Variable in terms of Integral). The expectation of a random variable  $X$  may be expressed as  $EX = \int_0^\infty P(X \geq t) dt$  for  $X \geq 0$ .<sup>5</sup>

Let us start by considering the case where  $X$  takes integer values only;

$$\begin{aligned} EX &= \sum_{n=1}^{\infty} nP(X = n) = \sum_{n=1}^{\infty} \sum_{m=1}^n P(X = n) \\ &= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} P(X = n) \\ &= \sum_{m=1}^{\infty} P(X \geq m). \end{aligned}$$

For the general case, let us start by writing  $a_n = \int_0^\infty 1(t \leq a_n) dt$  by the layer cake decomposition. Then

$$\begin{aligned} EX &= \sum_{n=1}^{\infty} a_n P(X = a_n) = \sum_{n=1}^{\infty} \left( \int_0^\infty 1(t \leq a_n) dt \right) P(X = a_n) \\ &\stackrel{(*)}{=} \int_0^\infty \left( \sum_{n=1}^{\infty} 1(t \leq a_n) P(X = a_n) \right) dt \quad (\text{Fubini}) \\ &= \int_0^\infty \sum_{a_n \geq t} P(X = a_n) dt \\ &= \int_0^\infty P(X \geq t) dt \end{aligned}$$

To resolve  $(*)$  without the use of Fubini's theorem, we may write

$$\begin{aligned} \sum_{n=1}^{\infty} \int_0^\infty 1(t \leq a_n) P(X = a_n) dt &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \int_{m-1}^m 1(t \leq a_n) P(X = a_n) dt \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \int_{m-1}^m 1(t \leq a_n) P(X = a_n) dt \\ &= \sum_{m=1}^{\infty} \lim_{N \rightarrow \infty} \sum_{n=1}^N \int_{m-1}^m 1(t \leq a_n) P(X = a_n) dt \\ &= \sum_{m=1}^{\infty} \lim_{N \rightarrow \infty} \int_{m-1}^m \left( \sum_{n=1}^N 1(t \leq a_n) P(X = a_n) \right) dt \\ &= \sum_{m=1}^{\infty} \int_{m-1}^m \left( \lim_{N \rightarrow \infty} \sum_{n=1}^N 1(t \leq a_n) P(X = a_n) \right) dt \\ &= \int_0^\infty \left( \sum_{n=1}^{\infty} 1(t \leq a_n) P(X = a_n) \right) dt, \end{aligned}$$

which we conclude by removing the auxiliary summations, since it is enough to know that  $\sum_{n=1}^N 1(t \leq a_n) P(X = a_n) \rightarrow \sum_{n=1}^{\infty} 1(t \leq a_n) P(X = a_n)$  uniformly in  $t \in [m-1, m]$ ; i.e.,

$$\left| \sum_{n=N+1}^{\infty} 1(t \leq a_n) P(X = a_n) \right| \leq \sum_{n=N+1}^{\infty} P(X = a_n) \rightarrow 0$$

as  $N \rightarrow \infty$  as per earlier (since the tail goes to 0).

<sup>5</sup>intuition: layer cake formula, but compile them together in level sets.

We also briefly went over examples multinomial distributions at the end of class;

- Suppose  $X_1, \dots, X_n$  are independent, and let  $P(X_i = j) = p_j$  for  $j = 1, \dots, k$ . Let  $\Omega = \{n_1, \dots, n_k\}, n_j \geq 0, n_1 + \dots + n_k = n\}$  (read:  $k$ -sided dice rolled  $n$  times, where  $n_j$  denotes the number of times  $j$  came up). Then

$$P((n_1, \dots, n_k)) = \binom{n}{n_1, \dots, n_k} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}.$$

- The geometric distribution; let  $0 < p < 1$  denote the probability of getting a head, and let us toss a coin until we get a heads. Let the outcome of  $X$  denote the number of tosses it took. Then

$$P(X = n) = (1 - p)^{n-1} p,$$

and we may check  $\sum_{n=1}^{\infty} P(x = n) = 1$  by geometric series.

- The Poisson distribution; let  $\lambda > 0$ . Then

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

for  $n = 0, 1, 2, \dots$



### §3 Day 3: Distributions, Stability Property, Moments (Sep. 11, 2024)

Recall the Poisson distribution

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

where  $n = 0, 1, \dots$  and  $\lambda > 0$ . We introduce the stability property: let us consider the independent random variables  $X_1, X_2$  distributed as follows,

$$\begin{aligned} X_1 &\sim \text{Poiss}_{\lambda_1}, \\ X_2 &\sim \text{Poiss}_{\lambda_2}. \end{aligned}$$

Then we have that  $X_1 + X_2 \sim \text{Poiss}_{\lambda_1 + \lambda_2}$ . To prove this, write

$$\begin{aligned} P(X_1 + X_2 = n) &= \sum_{m=0}^n P(X_1 = m, X_2 = n - m) \\ &= \sum_{m=0}^n P(X_1 = m)P(X_2 = n - m) \\ &= \sum_{m=0}^n \frac{\lambda_1^m}{m!} e^{-\lambda_1} \frac{\lambda_2^{n-m}}{(n-m)!} e^{-\lambda_2} \\ &= \frac{1}{n!} \sum_{m=0}^n \underbrace{\frac{n!}{m!(n-m)!}}_{(\lambda_1 + \lambda_2)^n} \lambda_1^m \lambda_2^{n-m} e^{-(\lambda_1 + \lambda_2)} \\ &= \frac{1}{n!} (\lambda_1 + \lambda_2)^n e^{-(\lambda_1 + \lambda_2)}. \end{aligned}$$

Binomials also have a related property; let

$$\begin{aligned} X_1 &\sim \text{Bin}(n_1, p), \\ X_2 &\sim \text{Bin}(n_2, p). \end{aligned}$$

Given that  $X_1, X_2$  are independent, we know that  $X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$ . To prove this, we may just write

$$\begin{aligned} X_1 &= y_1 + \dots + y_{n_1}, \\ X_2 &= y_{n_1+1} + \dots + y_{n_1+n_2}, \\ X_1 + X_2 &= y_1 + \dots + y_{n_1+n_2} \sim \text{Bin}(n_1 + n_2, p). \end{aligned}$$

Moreover, we also have  $\text{Bin}(n, \frac{\lambda}{n}) \xrightarrow{n \rightarrow \infty} \text{Poiss}_{\lambda}$ . This is called the *law of little numbers*. To prove this, we have

$$\begin{aligned} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} &= \frac{\lambda^k}{k!} \underbrace{\frac{n(n-1)\dots(n-k+1)}{n^k}}_{\rightarrow 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{n-k}}_{\rightarrow e^{-\lambda}} \\ &\xrightarrow{n \rightarrow \infty} \frac{\lambda^k}{k!} e^{-\lambda} = \text{Poiss}_{\lambda}. \end{aligned}$$

There are two examples of Poisson distributions that we will go over: shark attacks and radioactive decay. (but we didn't go over it ig?)

**Theorem 3.1** (Doebelin). Let  $X_i$  be independent random variables distributed by  $\text{Ber}_{p_i}$ , where  $0 < p_i < 1$ . Let us have  $S_n = X_1 + \dots + X_n$  with  $\lambda = p_1 + \dots + p_n$ . Then

$$\left| P(S_n \in A) - \sum_{n \in A} \frac{\lambda^n}{n!} e^{-\lambda} \right| \leq \sum_{i=1}^n p_i^2,$$

where  $A \subset \{0, 1, \dots\}$ .

To prove this, let  $y$  be a random variable where  $y \sim \text{Pois}_p$ , then  $P(y=0) = e^{-p} > 1-p$ . Define  $\Omega^\perp = \{-1, 0, 1, 2, \dots\}$ ; then we have  $P_p(-1) = 1-p$ ,  $P_p(0) = e^{-p} - 1 + p$ ,  $P_p(k) = \frac{p^k}{k!} e^{-p}$  for  $k = 1, 2, 3, \dots$ . Moreover, define

$$X(\omega) = \begin{cases} 0 & \omega = -1 \\ 1 & \omega \geq 0 \end{cases}, \quad y(\omega) = \begin{cases} 0 & \omega = 0, 1 \\ \omega & \omega \geq 1 \end{cases}.$$

Then  $P(x=y) = 1-p + pe^{-p} \geq 1-p + p(1-p) = 1-p^2$ , so  $P(x=y) \leq p^2$ . Take  $\Omega = (\Omega_+)^n$  and  $X_i(\omega) = X(\omega)$ , and let us have

$$P(\omega) = \prod_{i=1}^n P_{p_i}(\omega_i)$$

where  $X_i$  are independently distributed by  $\text{Ber}_p$ , and  $y_i$  are independently distributed by  $\text{Pois}_{p_i}$ . Finally, let us have

$$S_n = X_1 + \dots + X_n, \\ S'_n = y_1 + \dots + y_n,$$

then  $P(S_n \neq S'_n) \leq \sum_{i=1}^n P(X_i \neq y_i) \leq \sum_{i=1}^n p_i^2$ , which means  $S'_n \sim \text{Pois}_\lambda$ .<sup>6</sup> □

We now define *moments*. For a random variable  $X$ ,  $E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega)$ , as long as  $E|X| < \infty$ . Then moments are given by  $EX^n$  where  $n = 1, 2, \dots$  (??) For example, let  $X = \text{Pois}_\lambda$ . Then we have for  $n = 1$ ,

$$\begin{aligned} EX &= \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \sum_{n=0}^{\infty} \frac{\lambda^n}{(n-1)!} e^{-\lambda} \\ &= \sum_{n=0}^{\infty} \frac{\lambda^{n+1}}{n!} e^{-\lambda} \\ &= \lambda. \end{aligned}$$

For  $n = 2$ , we have

$$\begin{aligned} EX^2 &= \sum_{n=0}^{\infty} n^2 \frac{\lambda^n}{n!} e^{-\lambda} \\ &= E(X(X-1)) + EX \\ &= \lambda^2 + \lambda. \end{aligned}$$

<sup>6</sup>reminder: review this proof, i'm stupid and don't really get it.

## §4 Day 4: Independence and Dependence (Sep. 16, 2024)

Let  $(\Omega, P)$  be our probability space. We define

$$P(A | B) := \frac{P(A \cap B)}{P(B)}$$

to be the conditional probability, i.e. probability of  $A$  given  $B$ , as long as  $P(B) > 0$  (this is called Bayes' Rule). If  $P(A | B) = P(A)$ , then  $A$  is said to be independent of  $B$ . In particular, if  $A_1, \dots, A_n$  are independent, then

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i).$$

If the above is true for only pairs of events  $A_i, A_j$ , then we say that they are pairwise independent.

Let  $\Omega_i, P_i$  be probability spaces, and consider  $\Omega = \Omega_1 \times \dots \times \Omega_n = \prod_{i=1}^n \Omega_i$ , where we define a probability event in  $\omega \in \Omega$  to be  $(\omega_1, \dots, \omega_n) = \omega$  with  $\omega_i \in \Omega_i$ . Specifically, we have

$$P(\omega) := \prod_{i=1}^n P_i(\omega_i).$$

For example, let  $A = A_1 \times \dots \times A_n$ , and  $A_i \in \Omega_i$ . then

$$P(A) := \sum_{\omega \in A} P(\omega) = \sum_{\substack{\omega_i \in A_i \\ i=1, \dots, n}} \prod_{i=1}^n P_i(\omega_i) = \prod_{i=1}^n \sum_{\omega_i \in A_i} P_i(\omega_i) = \prod_{i=1}^n P_i(A_i).$$

Let us have random variables  $X_i : \Omega_i \rightarrow \mathbb{R}$  where  $1 \leq i \leq n$ . Then  $X_i^{(\omega)} = f_i(\omega_i)$  are independent if  $P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i)$ . In other words,  $X_i^{-1}(A_i)$  are independent.<sup>7</sup> We may continue simplifying the expression as follows,

$$\prod_{i=1}^n P(X_i \in A_i) = \prod_{i=1}^n P_i(f_i(\omega_i) \in A_i) = \prod_{i=1}^n P(X_i = x_i).$$

Now, suppose  $X, Y$  are independent and  $f, g$  are functions. Then we claim that  $f(X), g(Y)$  are independent. To check this, let us write

$$\begin{aligned} P(f(X) = a, g(Y) = b) &= P(X = f^{-1}(a), Y = g^{-1}(b)) \\ &= P(X \in f^{-1}(a), Y \in g^{-1}(b)) \\ &= P(f(X) = a)P(g(Y) = b). \end{aligned}$$

We can also do this with grouping; let  $\{1, \dots, n\} = \bigcup_{k=1}^m I_k$  with  $I_k$  disjoint; i.e., we're sorting  $[n]$  into disjoint subsets  $I_k$ . Then let  $y_k = f_k(\{x_i\}_{i \in I_k})$  for some function  $f_k : \mathbb{R}^{|I_k|} \rightarrow \mathbb{R}$ , and we have that  $y_k$  are independent. To prove this, observe that

$$\begin{aligned} P(y_1 \in A_1, \dots, y_m \in A_m) &= P(f_1 \in A_1, \dots, f_m \in A_m) \\ &= P(\{X_i\}_{i \in I_1} \in f_1^{-1}(A_1), \dots, \{X_i\}_{i \in I_m} \in f_m^{-1}(A_m)) \\ &= \prod_{j=1}^m P(\{X_i\}_{i \in I_j} \in f_j^{-1}(A_j)) \\ &= \prod_{j=1}^m P(y_j \in A_j). \end{aligned}$$

<sup>7</sup>note on board: for any  $A_1, \dots, A_n$  borel sets, intervals are enough, like  $(-\infty, x_i]$ . confusion?

We need to show that  $P(\{X_i\}_{i \in I_1} = b_1, \{X_i\}_{i \in I_2} = b_2) = P(\{X_i\}_{i \in I_1} = b_1)P(\{X_i\}_{i \in I_2} = b_2)$ ; but as per earlier, this is true.

Now, suppose our random variables  $X_i$ s are independently binomial distributed. Then

$$X_1 + \cdots + X_{m_1} \sim \text{Bin}(m_1, p) \sim \text{Poiss}_{\lambda_1}, \quad (\lambda_1 = pm_1)$$

$$X_{m_1+1} + \cdots + X_{m_1+m_2} \sim \text{Bin}_{m_2,p} \sim \text{Poiss}_{\lambda_2} \quad (\lambda_2 = pm_2)$$

We may combine the groupings above to get  $X_1 + \cdots + X_{m_1+m_2} \sim \text{Bin}(m_1 + m_2, p) \sim \text{Poiss}_{\lambda_1+\lambda_2}$ .

**Lemma 4.1.** If  $X$  and  $Y$  are independent and  $E[|X|] < \infty$ ,  $E[|Y|] < \infty$ , then  $E[XY] = E[X]E[Y]$ .<sup>8</sup>

First, assume  $X, Y > 0$ . Let us directly write

$$\begin{aligned} E[XY] &= \sum_{\omega \in \Omega} X(\omega)Y(\omega)P(\omega) \\ &= \sum_{n,m} a_n b_m P(X = a_n, Y = b_m) \\ &= \sum_{n,m} a_n b_m P(X = a_n)P(Y = b_m) \\ &= \sum_n a_n P(X = a_n) \sum_m b_m P(Y = b_m) \\ &= E[X]E[Y]. \end{aligned}$$

In the case that the random variables are not necessarily non-negative, we may simply consider

$$\begin{aligned} X &= X1(X \geq 0) - |X|1(X < 0) = X_+ - X_-, \\ Y &= Y1(Y \geq 0) - |Y|1(Y < 0) = Y_+ - Y_-. \end{aligned}$$

However, do note that in the OPPOSITE direction that  $E[XY] = E[X]E[Y]$  does NOT imply that  $X, Y$  are independent. It is true that  $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$  for “lots of”  $f, g$  would imply that  $X, Y$  independent (if this is true for *all*  $f, g$ , then it is independent), but this is unreliable.

Using Fubini’s theorem, we may consider  $X, Y$  on non-discrete probability spaces, and write

$$\begin{aligned} E[f(X, Y)] &= \sum_{n,m} f(a_n, b_m)P(X = a_n, Y = b_m) \\ &\stackrel{\text{if indep.}}{=} \sum_{n,m} f(a_n, b_m)P(X = a_n)P(Y = b_m) \\ &\stackrel{\text{if “nice”}}{=} \sum_n \left[ \sum_m f(a_n, b_m)P(Y = b_m) \right] P(X = a_n). \end{aligned}$$

We say that the above is “nice” if  $f \geq 0$ , or  $E[f(x, y)] < \infty$ , or

$$\sum_n \left[ \sum_m |f(a_n, b_m)| P(Y = b_m) \right] P(X = a_n) < \infty.$$

---

<sup>8</sup>i’m sick of the no bracket nonsense

Alternatively, if we don't have our "nice" cases, we have

$$\sum_n \left[ \sum_m f(a_n, b_m) P(Y = b_m \mid X = a_n) \right] P(X = a_n).$$

Now, we introduce the conditional distribution  $P(y = b_m \mid X = a_n)$ , where the distribution is  $Y$  given  $X = a_n$ . We can write the expectation

$$E[g(Y) \mid X = a_n] = \sum_m g(b_m) P(y = b_m \mid x = a_n),$$

i.e. the conditional expectation of  $g(Y)$  given  $X = a_n$ . For example, let  $X_1, X_2, \dots$  be i.i.d.  $\text{Ber}_p$  and  $N$  indep.  $\text{Pois}_\lambda$ . Then  $Y = X_1 + \dots + X_N$  has

$$P(Y = k) = \sum_{n=0}^{\infty} P(Y = k, N = n) = \sum_{n=0}^{\infty} P(y = k \mid N = n) P(N = n).$$

If  $N = n$ , then  $Y = X_1 + \dots + X_n$ , and we have

$$P(Y = k \mid N = n) = P(X_1 + \dots + X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Specifically,

$$\begin{aligned} P(y) &= \sum_{n=0}^{\infty} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{\lambda^n}{n!} e^{-\lambda} \\ &= \frac{(\lambda p)^k}{k!} \left( \sum_{n=0}^{\infty} \frac{(1-p)^{n-k}}{(n-k)!} \lambda^{n-k} \right) e^{-\lambda} \\ &= \frac{(\lambda p)^k}{k!} e^{-p\lambda} \sim \text{Pois}_{p\lambda}. \end{aligned}$$

Now for another example; let  $X_1, X_2, \dots$  be i.i.d.  $\text{Ber}_{y_2}$ ; i.e. let  $x_i \in \{0, 1\}^{\mathbb{N}}$ ; let  $x \in [0, 1) = \Omega$ ,  $X = 0, X_1, X_2, \dots$ . Let  $P$  be on  $[0, 1)$ . Then  $P([a, b)) = b - a$  where  $b > a$ ; we claim that they are i.i.d.  $\text{Ber}_{y_2}$ , which is proven by subdividing the intervals (whatever this means).

Let  $P(X_1 = x_1, \dots, X_n = x_n)$ . Then this is equal to

$$\begin{aligned} &= P(X_n = x_n \mid x_1 = x_1, \dots, X_{n-1} = x_{n-1}) P(X_1 = x_1, \dots, X_{n-1} = x_{n-1}) \\ &= \prod_{k=0}^{n-1} P(X_{k+1} = x_{k+1} \mid X_1 = x_1, \dots, X_k = x_k). \end{aligned}$$

In this specific kind of system where the probability of  $X_{k+1}$  only depends on the ones the step right before, we call it a *Markov Chain*, i.e. a probabilistic version of dynamical systems.

## §6 Day 6: Ulam's Problem; Chebyshev Inequality, Stirling Approximation, and Erdős-Renyi Random Graphs (Sep. 23, 2024)

Let us throw  $n$  balls into  $n$  boxes. Then consider  $N$  to be the number of empty boxes, and we have

$$P(N = k) = \frac{1}{n^n} \binom{n}{k} \sum_{\ell=0}^k (-1)^\ell \binom{k}{\ell} (k - \ell)^n,$$

with accompanying expectation

$$E[N] = \sum_{k=0}^n P(N = k).$$

However, we may simplify the expression as follows; let  $N$  instead be written as a sum of indicators, i.e.

$$N = \sum_{i=1}^N 1(\text{the } i\text{th box is empty}),$$

yielding

$$E[N] = \sum_{i=1}^N P(\text{the } i\text{th box is empty}) = nP(\text{the } i\text{th box is empty}) = n \frac{(n-1)^n}{n^n}.$$

Next example; the longest increasing subsequence of a random permutation. Let  $S_n$  be the set of bijections  $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ . Then  $|S_n| = n!$  where each  $\sigma$  has probability  $\frac{1}{n!}$ , and we define an increasing subsequence to be given by  $\sigma(i_1) < \sigma(i_2) < \dots < \sigma(i_k)$  for  $i_1 < i_2 < \dots < i_k$ . Let  $L_n(k)$  be the longest increasing subsequence of  $\sigma$ . Then  $L_n$  is a random variable; it remains to ask how big  $L_n$  is (Ulam's Problem). It is proven (though not in this class) that

$$\frac{L_n}{\sqrt{n}} \rightarrow 2$$

as  $n \rightarrow \infty$ , i.e. the expectation  $E[L_n] \sim 2\sqrt{n}$ <sup>9</sup>.

We prove a looser bound for now; let  $N_k$  be the number of increasing subsequences of length  $k$ . Then

$$N_k = \sum_{i_1 < \dots < i_k} 1(\sigma(i_1) < \dots < \sigma(i_k)),$$

and we may write

$$\begin{aligned} E[N_k] &= \sum_{i_1 < \dots < i_k} P(\sigma(i_1) < \dots < \sigma(i_k)) \\ &= \binom{n}{k} \frac{1}{k!} = \frac{n!}{(n-k)!(k!)^2}. \end{aligned}$$

We now present Chebyshev's inequality,

$$P(X \geq x) \leq \frac{E[X1(X \geq x)]}{x},$$

where  $x > 0$ . Using  $X1(X \geq x) \geq x1(X \geq x)$ , we have  $E[X1(X \geq x)] \geq xP(X \geq x)$ . Returning to earlier, we obtain  $P(N_k > 0) \leq E[N_k] = \frac{n!}{(n-k)!(k!)^2}$ .

<sup>9</sup>for more, see [here](#) :3

**Theorem 6.1** (Stirling's Formula).  $n! \sim \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n}$ .

To prove this, start by considering the Gamma function  $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ , which gives  $\Gamma(n) = (n-1)!$ . Let us consider the following;

$$\begin{aligned}
 \Gamma(n+1) &= \int_0^\infty t^n e^{-t} dt \\
 &= \int_0^\infty e^{n \log t - t} dt && \text{(Substitute } t = nx) \\
 &= n e^{n \log n} \underbrace{\int_0^\infty e^{n(\log x - x)} dx}_{\text{of the form } \int e^{nf(x)} dx} \\
 &\approx \int e^{nf(x^*) + \frac{f''(x^*)}{2}(x-x^*)^2} dx \\
 &= e^{-n} \int e^{-\frac{n}{2}(x-1)^2} dx && \text{(Substitute } x = 1 + \frac{y}{\sqrt{n}}) \\
 &= \frac{e^{-n}}{\sqrt{n}} \int_{-\infty}^\infty e^{-\frac{y^2}{2}} dy \\
 &= \frac{e^{-n}}{\sqrt{n}} \sqrt{\int_{-\infty}^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2+y^2)} dx dy} \\
 &= \frac{e^{-n}}{\sqrt{n}} \sqrt{\int_0^\infty \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr} \\
 &= \frac{e^{-n}}{\sqrt{n}} \sqrt{2\pi \int_0^\infty r e^{-\frac{r^2}{2}} dr} \\
 &= \frac{e^{-n}}{\sqrt{n}} \sqrt{2\pi}.
 \end{aligned}$$

*(I don't know where the hell this went. Oh well.)*

We may now insert Stirling's formula into Ulam's problem to obtain

$$P_n(N_k > 0) \leq \frac{n!}{(n-k)!(k!)^2} \approx c \frac{n^{n+\frac{1}{2}} e^{-n}}{k^{2k+1} e^{-2k} (n-k)^{n-k+\frac{1}{2}} e^{-(n-\frac{1}{2})}},$$

which cancels nicely. Using  $k! \geq k^k e^{-k}$ , we get  $P(N_k > 0) \leq (\frac{e\sqrt{n}}{k})^{2k}$ . We may make the bound nicer by writing

$$\left(\frac{e\sqrt{n}}{k}\right)^{2k} \leq \left(\frac{e}{3}\right)^{6\sqrt{n}}, \quad (k = 3\sqrt{n})$$

and using  $(\frac{e}{3})^6 \leq e^{-\frac{1}{2}}$  yields that it is less than  $e^{-\frac{n}{2}}$ . Thus, we have that  $P(L > \sqrt[3]{n}) \leq \sum_{m=3\sqrt{n}}^\infty e^{-\frac{\sqrt{m}}{2}}$ .

We now cover Erdős-Renyi random graphs. Let us have a graph on vertices  $V = \{v_1, \dots, v_n\}$ , and edges  $E = \{e_{ij}\} \subset V \times V$ . We have that  $e_{ij}$  is in  $E$  with probability  $p$ , and not there with probability  $1 - p$ , considered independently over all undirected pairs  $(i, j)$ .

Define a clique to be a complete subgraph of any graph  $G(n, p)$ , and let us have  $\omega(G)$  to be the clique number, i.e. the size of the largest clique of  $G$ . This is approximately  $C_p \log n$ , where  $C_p$  is some constant. Let us have  $N_k$  as the number of cliques of size  $k$ . To calculate the expectation, let us have

$$N_k = \sum_{V' \subset V, |V'|=k} 1_{\text{all } e_{ij} \text{ for all } i, j \in V' \text{ exists}}.$$

Then we have

$$E[N_k] = \binom{n}{k} p^{\binom{k}{2}} =: f(k).$$

Observe that we have

$$\frac{f(k+1)}{f(k)} = \frac{\binom{n}{k+1} p^{\binom{k+1}{2}}}{\binom{n}{k} p^{\binom{k}{2}}} = \frac{n-k}{k+1} p.$$

Observing that  $f(1) = n$  and  $f(n) = p^{\frac{n(n-1)}{2}} \ll 1$ , we see that  $f$  is unimodal. In particular, there is a unique point  $k_0$  such that  $f(k_0) \geq 1 > f(k_0 + 1)$ . Thus,

$$\left(\frac{n}{k} - 1\right)^k p^{\frac{k(k-1)}{2}} \leq f(k) \leq n^k p^{\frac{k(k-1)}{2}}.$$

In particular, the right hand side is less than 1 if  $np^{\frac{k-1}{2}} < 1$ , and this evaluates out to  $k > C_p \log n$ . The left hand side is greater than 1 when  $k \leq \frac{\log(\frac{n}{k}-1)}{|\log p|} + 1$ .



## §10 Day 10: Erdős-Renyi Random Graphs and Cliques, Chebyshev Inequality, Moment Generating Function (Oct. 7, 2024)

We start at *Example 1.5.5* in Panchenko. Consider the Erdős-Renyi random graph,  $G(n, p)$ , where  $n$  is the number of vertices of the graph, and  $p$  is the probability that an edge is in the graph. Then a clique subset of  $V = \{v_1, \dots, v_n\}$  is a complete graph, and we denote  $N_k$  to be the number of cliques of size  $k$ . We also define

$$f(k) = \mathbb{E}N_k = \binom{n}{k} p^{\binom{k}{2}},$$

i.e.  $f(1) = n$ ,  $f(n) = p^{\binom{n}{2}}$ . Then  $k_0$ , defined as  $f(k_0) \geq 1 > f(k_0 + 1)$ , has  $k_0 \sim c_p \log n$ , where  $c_p = \frac{2}{|\log p|}$ . We also have

$$f(k+1) = \frac{n-k}{k+1} p^k f(k) \leq n p^k f(k), \quad f(k_0 + m) \leq \frac{1}{n^m} (1 - \varepsilon).$$

Then we also get bounds on  $k_0$ , which is greater than  $\frac{(2-\varepsilon) \log n}{|\log p|}$ , and so  $p_0^k \leq \frac{1}{n^{2-\varepsilon}}$ .

Using Chebyshev, we get that

$$\mathbb{P}(N_{k_0+m+1} > 0) \leq \mathbb{E}[N_{k_0+m+1}] \leq \frac{1}{n^{m(1-\varepsilon)}}.$$

Observe that we have  $\frac{(2-\varepsilon) \log n}{|\log p|} < k < k_0$ , and  $p^k \leq \frac{1}{n^{2-\varepsilon}}$ , and

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1} p^k \leq n p^k \leq \frac{1}{n^{1-\varepsilon}}.$$

We want to prove that there are lots of cliques there. To do this, we use the second moment method,  $\mathbb{P}(|N_k - \mathbb{E}N_k| \geq x) \leq \frac{\text{Var}(N_k)}{x^2}$ . Using the fact that  $N_k$  is defined as

$$N_k = \sum_{\substack{W \subset V \\ |W|=k}} 1(W \text{ is a clique}),$$

we get that

$$\text{Var } N_k = \sum_{W, W'} \text{Cov}(1_W, 1_{W'}).$$

If  $|W \cap W'| \leq 1$ , then  $1_W, 1_{W'}$  are clearly independent, so  $\text{Cov}(1_W, 1_{W'}) = 0$ . If  $W$  and  $W'$  share more than 1 vertex, though, call this number  $i$ . Then

$$\text{Cov}(1_W, 1_{W'}) = \mathbb{E}[1'_W 1_W] - \mathbb{E}[1_W] \mathbb{E}[1_{W'}] \leq \mathbb{E}[1'_W 1_W] = \mathbb{P}(W, W' \text{ cliques}).$$

Note that there are  $\binom{n}{k} \binom{k}{i} \binom{n-k}{k-i}$  pairs of  $W, W'$  with  $i$  vertices in common. Thus, we have that the probability of  $W, W'$  being cliques is given by

$$\mathbb{P}(W, W' \text{ cliques}) = p^{\binom{k}{2}} p^{\binom{k}{2} - \binom{i}{2}}.$$

Then we may write

$$\begin{aligned} \text{Var}(N_k) &\leq \sum_{i=2}^k \binom{n}{k} \binom{k}{i} \binom{n-k}{k-i} p^{\binom{k}{2}} p^{\binom{k}{2} - \binom{i}{2}} \\ &= f^2(k) \sum_{i=2}^k \frac{\binom{k}{i} \binom{n-k}{k-i}}{\binom{n}{k}} p^{-\binom{i}{2}}. \end{aligned}$$

Using Chebyshev from earlier, we get

$$\mathbb{P}(|N_k - \mathbb{E}N_k| \geq \delta \mathbb{E}N_k) \leq \frac{1}{\delta^2} \sum_{i=2}^k a(i).$$

It remains to show that the right hand side is small for  $k$  in the range. This just leads to tedious calculation; to start, let us check  $a(2)$ :

$$\begin{aligned} a(2) &= \frac{\binom{k}{2} \binom{n-k}{k-2}}{\binom{k}{2}} p^{-\binom{2}{2}} \\ &= \frac{k^2(k-1)^2}{2} \frac{(n-k) \dots (n-2k+3)}{n(n-1) \dots (n-k+1)} \frac{1}{p} \\ &\leq \frac{k^4}{(n-k)^2} \frac{1}{2p} \leq c_p \frac{(\log n)^4}{n^2}. \end{aligned}$$

We also have that

$$a(k) = \frac{1}{\binom{n}{k}} p^{-\binom{k}{2}} = \frac{1}{f(k)}.$$

Let us define  $b(i)$  as the ratio of two consecutive  $a(i)$ , and write

$$b(i) = \frac{a(i+1)}{a(i)} = \frac{\binom{k}{i+1} \binom{n-k}{k-i-1}}{\binom{k}{i} \binom{n-k}{k-i}} p^{-(i+1)+\binom{i}{2}} = \frac{(k-i)^2}{(i+1)(n-2k+i+1)} p^{-i}.$$

We may derive the properties of  $a$  from  $b$ ; in particular, we like  $b$  more because it is “nicer” than  $a$ . In particular, if  $i \leq \frac{1}{3} \frac{\log n}{|\log p|}$ , then  $p^{-i} \leq n^{\frac{1}{3}}$ , and we have  $b(i) \sim n^{-\frac{2}{3}} \lesssim 1$ .

Likewise, if  $i > \frac{3}{2} \frac{\log n}{|\log p|}$ , then  $p^{-i} > n^{\frac{3}{2}}$ . Then  $b(i) \gtrsim \frac{n^{\frac{1}{2}}}{k} > 1$ . We have that  $\frac{b(i+1)}{b(i)} \gtrsim \frac{1}{p} > 1$ , so  $b$  is increasing.

We now return to sums of independent random variables. Let  $X_1, \dots, X_n$  be i.i.d., and  $\mathbb{E}X_i^2 < \infty$ . Let us have  $\overline{X_n} = \frac{1}{n} \sum_{i=1}^n x_i$ , and

$$\mathbb{P}(|\overline{X_n} - \mathbb{E}X_1| > \varepsilon) = \frac{\text{Var } \overline{X_n}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

where we let  $\varepsilon^2$  be the variance of  $X_i$ . We could do better, i.e.

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[e^{\lambda x}]}{e^{\lambda t}}.$$

This gives

$$\mathbb{P}(X_1 + \dots + X_n \geq t) \leq \frac{\mathbb{E}[e^{\lambda(X_1 + \dots + X_n)}]}{e^{\lambda t}} = \frac{EE[e^{\lambda X_1}]^n}{e^{\lambda t}}.$$

This is called exponential Chebyshev. For example, let  $X = \sum \varepsilon_i a_i$ , where  $\varepsilon_i$  are independent Rademacher with  $\mathbb{P}(\varepsilon_i = 1) = \frac{1}{2} = \mathbb{P}(\varepsilon_i = -1)$  (read: coinflip distribution). Then we may apply exponential Chebyshev and optimize to get (Theorem 3.1 in Panchenko)

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right) \leq e^{-\frac{t^2}{2 \sum_{i=1}^n a_i^2}}.$$

We also have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \varepsilon_i a_i\right| \geq t\right) \leq 2e^{-\frac{t^2}{2 \sum_{i=1}^n a_i^2}}.$$

By the law of large numbers for a fair coin, we may take  $a_i = \frac{1}{n}$  as per  $\overline{X_n}$ , and we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right| \geq t\right) \leq 2e^{-\frac{nt^2}{2}}. \quad \square$$

Small digression; let  $X_1, \dots$  be i.i.d., and consider  $\mathbb{E}[e^{\lambda x}] < \infty$ . We call the LHS  $M(\lambda)$ , i.e. the moment generating function. In particular,  $M'(0) = \mathbb{E}[X]$ . We also have  $M''(0) = \mathbb{E}[X^2] = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \mathbb{E}[X^n]$ , meaning that we have  $X^n \leq C_\lambda[e^{\lambda X} + e^{-\lambda X}]$ . Notice that we may write

$$\mathbb{P}\left(\frac{X_1 + \dots + X_n}{n} > t\right) \leq \frac{M(\lambda)^n}{e^{\lambda tn}} = e^{-n\{\lambda t - \log M(\lambda)\}}.$$

Optimizing this over  $\lambda$ , we get that the LHS is less than or equal to  $e^{-nI(t)}$ , where  $I(t) = \sup_{\lambda} \{\lambda t - \log M(\lambda)\}$ .

Let  $a_1^2 + \dots + a_n^2 = 1$ . Then  $\mathbb{E}X^{2k+1} = 0$ , and we may write

$$\begin{aligned} \mathbb{E}X^{2k} &= \int_0^\infty 2kt^{2k-1} \mathbb{P}(|X| > t) dt \\ &\leq \int_0^\infty 2kt^{2k-1} 2e^{-\frac{t^2}{2}} dt \\ &= k2^{k+1} \int_0^\infty u^{k-1} e^{-u} du \\ &= k2^{k+1} \Gamma(k) \\ &= k2^{k+1} (k-1)! = 2^{k+1} k!. \end{aligned}$$

Let  $Z \sim N(0, 1)$ . Then we have

$$\mathbb{E}[e^{\lambda z^2}] = \int_{-\infty}^{\infty} e^{\lambda z^2 - \frac{z^2}{2}} \frac{dz}{\sqrt{2\pi}} < \infty$$

if and only if  $\lambda < \frac{1}{2}$ . Let  $Y = X^2 - 1$ . Then  $e^x \leq 1 + x + \frac{x^2}{2} + \sum_{k=3}^{\infty} \frac{x_+^k}{k!}$ . We denote  $x_+ = x$  if  $x \geq 0$ , and 0 otherwise.  $\square$

## §12 Day 12: Inequalities (Oct. 16, 2024)

Let  $X_i$  be independently distributed  $B(p)$ , and consider  $X_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then (where  $q > p$ )

$$\mathbb{P}(\overline{X_n} \geq q) \leq e^{-n\lambda q} \mathbb{E}[e^{\lambda x_i}]^n = e^{-n\{\lambda q - \log(1-p+pe^\lambda)\}}.$$

We want to take the supremum of  $\{\lambda q - \log(1-p+pe^\lambda)\}$ ; then we have

$$f'(\lambda) = q - \frac{pe^\lambda}{1-p+pe^\lambda},$$

where we note if  $f'(\lambda) = 0$ , then  $q(1-p) + qpe^\lambda = pe^\lambda$  implies  $e^\lambda = \frac{q(1-p)}{p(1-q)} > 1$ . We write  $D(q||p)$  to be the Kullback-Leibler divergence, aka relative entropy, i.e.  $H(q||p)$ . Let  $M$  be a differentiable manifold of dimension  $n$ . Then let  $D : M \times M \rightarrow [0, \infty)$ , and consider  $M$  as a parameterized family of probability measures. Then

$$\begin{aligned} D(q, p) &\geq 0; \\ D(q, p) &= 0; && \text{(if and only if } q = p) \\ D(p, p + dp) & && \text{(should be positive def. quadratic in } dp) \end{aligned}$$

We check that this is indeed true.

(a) We check that  $D$  is non-negative.

$$\begin{aligned} & q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \\ &= q \left( -\log \frac{p}{q} \right) + (1-q) \left( -\log \frac{1-p}{1-q} \right) \\ &\geq q \left( 1 - \frac{p}{q} \right) + (1-q) \left( 1 - \frac{1-p}{1-q} \right) \\ &= q - p + 1 - q - (1-p) = 0. \end{aligned}$$

(b) We now check that it is identically zero iff  $q = p$ . Intuitively, there is no entropy needed to move  $p \rightarrow q$  if they are equal.

$$\begin{aligned} & \inf_q \underbrace{\left\{ q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \right\}}_{g'(a)} \\ & g'(a) = \log \frac{q}{p} - q \frac{p}{q} \frac{1}{p} - \log \frac{1-q}{1-p} = 1 \\ \implies & \log \frac{q}{p} = \log \frac{1-q}{1-p}, \end{aligned}$$

which occurs only when  $p = q$ .

(c) We leave the third alone for now.

An application of this is to classification algorithms. Consider  $t$  a classifier, and  $E_n$  an empirical error, i.e.

$$E_n(t) = \frac{1}{n} \sum_{i=1}^m L(y_i, f(x_i)),$$

where  $L$  is some loss function, and we consider  $(X_i, Y_i)$  i.i.d.  $\mathcal{F} = \{f_1, \dots, f_N\}$ . The generalization error  $E(f) = E[L(X, f(y))]$ , and suppose we have

$$\mathbb{P}(E_n(f) \geq E(f) + \varepsilon) < e^{-cn\varepsilon^2}.$$

Then  $\mathbb{P}(\forall f \in \mathcal{F}, E(f) \leq E_n(f) + \varepsilon) \geq 1 - Ne^{cn\varepsilon^2}$ . Let  $\delta = Ne^{-Cn\varepsilon^2}$ ; to get confidence  $1 - \delta$  that the generalization error is within  $\varepsilon$ ; then we need  $n = \frac{1}{c\varepsilon^2} \log \frac{N}{\delta}$ .