6.7720/18.619/15.070 Lecture 1
Introduction to Discrete Probability & Stochastic Processes

Kuikui Liu

February 3, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials created by Guy Bresler and David Gamarnik, who taught previous iterations of this course. We are grateful for the discussions we had with them. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Introduction

This course is about *stochastic processes*, which really is just a fancy name for a collection of random variables $\mathcal{X} = \{X_\iota\}_{\iota \in \mathcal{I}}$ drawn from some high-dimensional probability distribution, where $\mathcal{I}$ is some "index set". Typically this collection is used to model some real-world phenomenon, e.g. the evolution of stock prices, or properties of materials. The stochastic process could be handed to you by a colleague, or you might have to engineer and analyze a process as part of some greater, overarching endeavor (which a priori might not have any randomness in it!). Our goal will be to lay the mathematical foundations for how to rigorously study such processes. The ideas we will develop in this course primarily rest upon three pillars.

Models We will see many examples of stochastic processes. Perhaps the simplest example is just a sequence of independent coin flips, i.e. each X_i takes value 0 or 1 with equal probability. But we will go far beyond this, and consider stochastic processes where

- there are possibly complex dependencies between the random variables, and
- the index set $\mathcal{I}$ has tangible meaning (e.g. $\mathcal{I} = \mathbb{N}$ representing the arrow of time, $\mathcal{I}$ is the set of edges/hyperedges of a graph/hypergraph, $\mathcal{I}$ is the set of indices of a matrix/tensor, $\mathcal{I}$ is a subclass of Boolean functions, etc.).

As the title of the course suggests, our focus will primarily be on *discrete* random variables. Nonetheless, ideas from the continuous world will play an essential role in our journey.

Questions Once we have our models, we can ask all sorts of questions regarding the properties of our stochastic processes. These questions originate from numerous disparate fields of research.

- **Statistics:** Given samples of a stochastic process, what properties of the underlying distribution of the process can we infer (e.g. parameter estimation)? Can we distinguish between two different distributions (i.e. *hypotheses*) from which the samples could have been generated? These types of problems broadly fall under the umbrella of high-dimensional *statistical inference.*
- **Computation:** One can ask whether or not various natural computational tasks associated to the stochastic process are tractable or intractable. For instance, if $\mathcal{X}$ is a collection of indicator random variables describing a graph, we can consider classical problems like coloring, clique, and Hamiltonian cycle, all of which are well-studied in complexity theory and theoretical computer science more broadly. As another example, if we are given a description of the distribution of $\mathcal{X}$, we can try to design efficient algorithms for generating samples from that distribution to get a sense of what a "typical realization" of $\mathcal{X}$ "looks like" on a computer.

- **Physics:** Suppose we have a *family* of stochastic processes $\mathcal{X}_p$ *parametrized* by some $p \in \mathbb{R}$. How do the properties of $\mathcal{X}_p$ change as we vary p? As with many large-scale physical systems, an intriguing phenomenon we will encounter is the presence of a *phase transition* at some p_c, where the properties of $\mathcal{X}_p$ suddenly and dramatically change when we perturb p around p_c by even a tiny amount.

Techniques Finally, we will need to develop tools to answer the questions we ask. We roughly group them into a few unifying themes.

- The Probabilistic & Moment Methods
- Concentration of Measure
- Comparison Methods (e.g. Coupling)
- Algorithmic Methods

Disclaimer: Both in this course and in real life, almost none of the problems we will encounter will be nice enough to admit closed-form solutions. Hence, many of our results and techniques will come in the form of inequalities and asymptotic estimates.

1.1 Basic Tools

In this subsection, we briefly review some of the basic inequalities that we will use repeatedly throughout this course.

- **Markov's Inequality:** For a *nonnegative* random variable X, we can always upper bound the upper tail probabilities by

  $$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}, \qquad \forall t > 0.$$

  To use such an inequality, the only information we need about X is its expectation, which makes it widely applicable. For instance, if X decomposes as a weighted sum $\sum_{i=1}^n w_i X_i$ of (possibly dependent) smaller random variables, then $\mathbb{E}[X]$ decomposes as $\sum_{i=1}^n w_i \cdot \mathbb{E}[X_i]$, a fact known as *linearity of expectation*. Note that this holds regardless of how correlated $X_1, \ldots, X_n$ are!

- **Chebyshev's Inequality:** If we additionally have control over the variance of our random variable, then we get stronger bounds on the probability that X deviates from its expectation: For any real-valued random variable X (not necessarily nonnegative),

  $$\Pr\left[|X - \mathbb{E}[X]| \geq t\right] \leq \frac{\mathrm{Var}(X)}{t^2}, \qquad \forall t > 0.$$

- **Union Bound:** If $E_1, \ldots, E_n$ is a collection of events, then

  $$\Pr\left[\bigcup_{i=1}^n E_i\right] \leq \sum_{i=1}^n \Pr[E_i],$$

  with equality if and only if $E_1, \ldots, E_n$ are *mutually exclusive*, i.e. pairwise disjoint.[1]

2 Introduction to Percolation

We begin by discussing one of the most fundamental models of a random network, and some of its basic properties.

Definition 1 (Bond Percolation)**.** *For a graph $G = (V, E)$ and $p \in [0, 1]$,* (bond) percolation *on G yields a random subgraph $H = (V, F)$ with $F \subseteq E$, where each edge $e \in E$ is included in F independently with probability p.*

[1] Technically, you can allow nonempty intersections, but they must have zero measure.

We isolate a particular special case, namely the case where G is the *complete graph* and $E = \binom{V}{2}$. This is an extremely famous and well-studied model of random graphs.

Definition 2 (Erdös–Rényi Random Graph). *For $n \in \mathbb{N}$ and $p \in [0,1]$, the* Erdös–Rényi random graph *$\boldsymbol{G}(n,p)$ is a random undirected simple graph on vertex set V with $|V| = n$ where each pair of distinct vertices $\{u,v\} \in \binom{V}{2}$ is included in the edge set E independently with probability p. We sometimes abuse notation and write $G \sim \boldsymbol{G}(n,p)$ for G drawn from the Erdös–Rényi distribution.*

Fact 2.1. *For any $n \in \mathbb{N}$ and $p \in [0,1]$, the expected number of edges in $\boldsymbol{G}(n,p)$ is $p \cdot \binom{n}{2}$.*

Hence, if p is a constant independent of n, we expect to see a very dense graph when sampling $G \sim \boldsymbol{G}(n,p)$. We will also be interested in sparser situations where $p = p_n$ is allowed to depend on n. For example, we may take $p_n = \frac{c}{n^\alpha}$, $p_n = \frac{c \ln n}{n}$, or $p_n = \frac{c}{n}$ for fixed constants $c > 0$ and $\alpha \in [0,1]$. What will be important is how p_n scales as n grows. In the remainder of this lecture, we will drop the subscript, as everything will be with respect to $G \sim \boldsymbol{G}(n,p)$. All asymptotics will be in the large n limit.

2.1 Local Structures: Triangles in Erdös–Rényi

Definition 3 (Triangle). *For a graph G, a triangle is a triple of distinct vertices $\{u,v,w\} \in \binom{V}{3}$ such that every pair is connected by an edge in G. We let T_G denote the number of triangles in G.*

Lemma 2.2. *For any $n \in \mathbb{N}$ and $p \in [0,1]$,*

$$\mathbb{E}\left[T_G\right] = p^3 \cdot \binom{n}{3} = (1 + o(1)) \cdot \frac{p^3 n^3}{6}.$$

Proof. For each triple $\{u,v,w\} \in \binom{V}{3}$, the probability that it forms a triangle in $G \sim \boldsymbol{G}(n,p)$ is precisely p^3. The claim then follows by linearity of expectation. □

Proposition 2.3. *Let $0 < p \leq 1$ be an arbitrarily fixed constant independent of n. Then for every $\epsilon > 0$ (again independent of n),*

$$\Pr\left[\left|\frac{T_G}{p^3 n^3/6} - 1\right| > \epsilon\right] \to 0 \qquad \textit{as} \qquad n \to \infty.$$

In other words, $\frac{T_{\boldsymbol{G}(n,p)}}{p^3 n^3/6}$ converges to 1 in probability, *sometimes written as $\frac{T_{\boldsymbol{G}(n,p)}}{p^3 n^3/6} \xrightarrow{\mathsf{P}} 1$, as $n \to \infty$.*

Proof. We use Chebyshev's Inequality to bound the probability that T_G deviates from its expectation. Let $\mathcal{I}_{uvw}$ denote the indicator random variable for whether or not the triple of vertices $\{u,v,w\}$ forms a triangle in the graph. Our goal is to estimate the variance of T_G and then compare it to its squared expectation. Expanding using the linearity of expectation, we have

$$\mathbb{E}\left[T_G^2\right] = \left[\sum_{\{u,v,w\},\{x,y,z\} \in \binom{V}{3}} \mathbb{E}\left[\mathcal{I}_{uvw}\mathcal{I}_{xyz}\right]\right].$$

Each term is equal to the probability that some subset of edges are present. These probabilities depend on the number of overlapping vertices between the triples $\{u,v,w\}$ and $\{x,y,z\}$. We have four cases.

1. If $\{u,v,w\} = \{x,y,z\}$, then $\mathbb{E}[\mathcal{I}_{uvw}\mathcal{I}_{xyz}] = p^3$. There are $\binom{n}{3} = O(n^3)$ such terms.
2. If $\{u,v,w\}, \{x,y,z\}$ share exactly two vertices, then $\mathbb{E}[\mathcal{I}_{uvw}\mathcal{I}_{xyz}] = p^5$ corresponding to the fact that the structure $\mathcal{I}_{uvw}\mathcal{I}_{xyz}$ captures is a complete graph on 4 vertices with a single edge removed. There are $\binom{n}{4} \cdot \binom{4}{2} = O(n^4)$ such terms.
3. If $\{u,v,w\}, \{x,y,z\}$ share exactly one vertex, then $\mathbb{E}[\mathcal{I}_{uvw}\mathcal{I}_{xyz}] = p^6$ corresponding to the fact that the structure $\mathcal{I}_{uvw}\mathcal{I}_{xyz}$ captures is the graph consisting of two triangles joined at a single vertex. There are $\binom{n}{3} \cdot 3 \cdot \binom{n-3}{2} = O(n^5)$ such terms.
4. If $\{u,v,w\}, \{x,y,z\}$ do not share any vertices, then $\mathbb{E}[\mathcal{I}_{uvw}\mathcal{I}_{xyz}] = \mathbb{E}[\mathcal{I}_{uvw}] \cdot \mathbb{E}[\mathcal{I}_{xyz}] = p^6$. There are $\binom{n}{3} \cdot \binom{n-3}{3}$ such terms.

Putting together these calculations, we get

$$\mathrm{Var}\left(T_G\right) = \mathbb{E}[T_G^2] - \mathbb{E}[T_G]^2 \leq O_p(n^5),$$

where the constant depends on p. It follows that

$$\begin{aligned} \Pr\left[\left|\frac{T_G}{p^3n^3/6} - 1\right| > \epsilon\right] &= \Pr\left[\left|T_G - \frac{p^3n^3}{6}\right| > \epsilon \cdot \frac{p^3n^3}{6}\right] \\ &\leq \frac{6}{\epsilon^2} \cdot \frac{\mathrm{Var}\left(T_G\right)}{p^6n^6} && \text{(Chebyshev's Inequality)} \\ &\leq O_{\epsilon,p}(1/n), \end{aligned}$$

which decays to 0 as $n \to \infty$, for every fixed $\epsilon > 0$ and $p \in (0, 1]$. □

Remark 1. By being more careful with the calculations, one can show that the correct order for the variance of T_G is actually $O_p(n^4)$. This can actually be shown using some of the general-purpose tools we'll discuss in a future lecture.

2.2 Global Structure: The Erdös–Rényi Connectivity Phase Transition

Theorem 2.4. *For $p_n = \frac{c \ln n}{n}$ where $c \in \mathbb{R}_{\geq 0}$ is a constant, we have*

$$\lim_{n\to\infty} \Pr\left[\boldsymbol{G}(n, p_n) \text{ is connected}\right] = \begin{cases} 1, & \text{if } c > 1 \\ 0, & \text{if } c < 1. \end{cases}$$

Remark 2. At $c = 1$, one can show that $\lim_{n\to\infty} \Pr\left[\boldsymbol{G}(n, p_n) \text{ is connected}\right]$ equals some explicit constant lying strictly between 0 and 1.

Connectivity is a global property of a graph. The usual definition states that for every pair of vertices $u, v \in V$, there is a path from u to v using the edges of the graph. The "dual" viewpoint is that for every subset of vertices $\emptyset \subsetneq S \subsetneq V$, there exists an edge $\{u, v\}$ in the graph such that $u \in S$ and $v \in \overline{S} \overset{\mathsf{def}}{=} V \setminus S$; in other words, there is an edge crossing every *cut* in the graph. This is an equivalent definition of connectivity which is more amenable to calculations. In particular,

$$\Pr\left[\boldsymbol{G}(n, p_n) \text{ is disconnected}\right] = \Pr\left[\exists \emptyset \subsetneq S \subsetneq V \text{ s.t. } E\left(S, \overline{S}\right) = \emptyset\right]$$

and

$$\Pr\left[E\left(S, \overline{S}\right) = \emptyset\right] = (1 - p_n)^{|S| \cdot |V \setminus S|}, \qquad \forall \emptyset \subsetneq S \subsetneq V,$$

where we write $E\left(S, \overline{S}\right)$ for the subset of edges of G crossing the cut $\left(S, \overline{S}\right)$. This already suggests a way to upper bound the probability that G is disconnected. In particular, by combining the preceding two displays with the Union Bound (and the fact that S and $\overline{S}$ induce the same cut), we have

$$\Pr\left[\boldsymbol{G}(n, p_n) \text{ is disconnected}\right] \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1 - p_n)^{k \cdot (n-k)}. \tag{1}$$

We will show that if $c > 1$, then the right-hand side converges to 0 as $n \to \infty$. If $c < 1$, we will see that the right-hand side actually grows with n due to the contribution from the $k = 1$ term in the summation, which signals the presence of isolated vertices. These certainly certify disconnectedness of the graph, and so the $c < 1$ case will proceed by lower bounding the probability of those events.

2.3 Heuristic Reasoning

Before we dive into calculations, let's try to get a heuristic understanding of the right-hand side of Eq. (1). Perhaps the most unwieldy component is the binomial coefficient $\binom{n}{k}$. However, an extremely useful approximation is

$$\binom{n}{k} \text{``} \approx \text{''} \exp\left(H_e\left(\frac{k}{n}\right) \cdot n\right) \qquad \text{where} \qquad H_e(p) \overset{\mathsf{def}}{=} -p \ln(p) - (1 - p) \ln(1 - p).$$

Here $H_e(p)$ is the *entropy* of $\mathsf{Ber}(p)$ measured in base e. One can formalize this approximation using *Stirling's approximation* for the factorial (see below), but one must be careful about the size of the errors in this approximation, even when $k \ll n$. Nonetheless, if our goal is to merely get a sense of what we're dealing with, we can treat it as an equality.

Using this approximation and $1 - p_n \approx e^{-p_n}$ (which is valid for small p_n; note that $1 - x \leq e^{-x}$ for all x), we can approximate the right-hand side of Eq. (1) by

$$\text{“}\lesssim\text{”} \sum_{k=1}^{\lfloor n/2 \rfloor} \exp\left(n \cdot \left(H_e(k/n) - c\ln(n) \cdot \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)\right)\right).$$

In order for this to decay to 0 as $n \to \infty$, we certainly need each term to decay to 0, and in particular, we need

$$H_e(k/n) < M \cdot \frac{k}{n} \cdot \left(1 - \frac{k}{n}\right), \qquad \forall k = 1, \ldots, \lfloor n/2 \rfloor,$$

where $M = c\ln n$. Now if we compare $H_e(p)$ with the function $f(p) = Mp(1-p)$ (e.g. by plotting them), we find that $f(p)$ upper bounds $H_e(p)$ precisely for p lying in a symmetric interval around $1/2$ as soon as $M \geq e$. The crossover point for when $H_e(p) \geq Mp(1-p)$ occurs at p^* and $1 - p^*$, where $p^* \approx \exp(1 - M)$ for M large. If $M = c\ln n$, this yields $p^* \asymp \frac{1}{n^c}$. In order for this to be smaller than $\frac{k}{n}$ for every $k = 1, \ldots, \lfloor n/2 \rfloor$ in the large n limit, we certainly need $c > 1$. On the other hand, as soon as $c < 1$, the $k = 1$ term becomes macroscopic and dominates the sum. This roughly explains the transition at $c = 1$ and why the scaling is $p_n \asymp \frac{\ln n}{n}$.

2.4 Proof of Theorem 2.4: The Case $c > 1$

Our goal is to show that the right-hand side of Eq. (1) decays to 0 as $n \to \infty$ when $c > 1$. Towards this, recall that one way to state Stirling's approximation is,

$$\lim_{k\to\infty} \frac{k!}{k^{k+1/2}e^{-k}} = 1,$$

which in particular implies that $\ln(k!) \geq \left(k + \frac{1}{2}\right)\ln k - k - C$ for some universal constant $C > 0$. Combining this with the standard inequalities $\binom{n}{k} \leq \frac{n^k}{k!}$ and $1 - x \leq e^{-x}$, we obtain the upper bound

$$e^C \cdot \sum_{k=1}^{\lfloor n/2 \rfloor} \exp\left(k\ln n - k(n-k)p_n - \left(k + \frac{1}{2}\right)\ln k + k\right). \tag{2}$$

To show that this is $o(1)$, we separate the terms in the sum into two regimes. Fix a small constant $\epsilon > 0$ such that $(1 - \epsilon)c > 1$ (which exists since $c > 1$).

- Suppose $1 \leq k \leq \epsilon n$. In this case, note that $\exp\left(-\left(k + \frac{1}{2}\right)\ln k + k\right) \leq O(1)$ (in fact, the exponent is decreasing and becomes negative already for $k \geq 3$) so we can ignore this term. Since $(1 - \epsilon)c > 1$, we have that

  $$k\ln n - k(n-k)p_n < k\ln n - k(1-\epsilon)n \cdot \frac{c\ln n}{n} = -((1-\epsilon)c - 1)k\ln n.$$

  Hence, if we set $\widehat{c} = (1-\epsilon)c - 1 > 0$, then the first ϵn terms of Eq. (2) are upper bounded by

  $$O(1) \cdot \sum_{k=1}^{\epsilon n} n^{-\widehat{c}\cdot k} \leq O\left(n^{-\widehat{c}}\right),$$

  which clearly decays to 0 as $n \to \infty$.

- Suppose $\epsilon n \leq k \leq \lfloor n/2 \rfloor$. Then observe that

  $$k\ln n - \left(k + \frac{1}{2}\right)\ln k + k \leq k\ln n - k\ln(\epsilon n) + k = k\left(1 + \ln(1/\epsilon)\right)$$

while $k(n-k)p_n \geq \frac{1}{2}ck\ln n$. It follows that the remaining terms of Eq. (2) are upper bounded by

$$\sum_{k=\epsilon n}^{\lfloor n/2 \rfloor} \left(\frac{e}{\epsilon \cdot n^{c/2}}\right)^k \leq O\left(n^{-c/2}\right),$$

which also decays to 0 as $n \to \infty$.

2.5 Proof of Theorem 2.4: The Case $c < 1$

Observe that

$$\Pr\left[\boldsymbol{G}(n, p_n) \text{ is disconnected}\right] \geq \Pr\left[\exists \text{ isolated vertex}\right].$$

Hence, it suffices to show that the right-hand side converges to 1 when $c < 1$. Towards this, let $\mathcal{I}_v$ be the indicator random variable for whether or not the vertex v is isolated, and let $\mathcal{I} = \sum_{v\in V}\mathcal{I}_v$ be the total number of isolated vertices. Observe that the expectation of $\mathcal{I}$ (which is the $k = 1$ term in the right-hand side of Eq. (1)) is given by

$$\begin{aligned}
\mathbb{E}\left[\mathcal{I}\right] &= n(1-p_n)^{n-1} \\
&= n \cdot \exp\left((n-1)\ln\left(1 - \frac{c\ln n}{n}\right)\right) \\
&= n \cdot \exp\left(-c\ln n - o(\ln n)\right) && \text{(Using } \ln(1-x) = -x - o(x) \text{ for small } x\text{)} \\
&\geq n^{1-c-o(1)}.
\end{aligned}$$

This lower bound on the expectation by itself is not enough to obtain a $1 - o(1)$ lower bound on the probability that there exists an isolated vertex. To complete the proof, we compute the variance of $\mathcal{I}$ and compare it with $\mathbb{E}\left[\mathcal{I}\right]^2$. We have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{I}^2\right] &= \sum_{v\in V}\mathbb{E}[\mathcal{I}_v] + \sum_{u\neq v}\mathbb{E}\left[\mathcal{I}_u\mathcal{I}_v\right] && \text{(Expanding and using } \mathcal{I}_v^2 = \mathcal{I}_v\text{)} \\
&= n(1-p_n)^{n-1} + n(n-1)(1-p_n)^{2(n-2)+1}. && \text{(Direct Calculation)}
\end{aligned}$$

It follows that

$$\begin{aligned}
\Pr\left[\mathcal{I} = 0\right] &\leq \Pr\left[\mathcal{I} \leq \frac{1}{2}\mathbb{E}[\mathcal{I}]\right] \\
&\leq \frac{\mathrm{Var}\left(\mathcal{I}\right)}{\left(\mathbb{E}[\mathcal{I}]/2\right)^2} && \text{(Chebyshev's Inequality)} \\
&= 4 \cdot \left(\frac{1}{n(1-p_n)^{n-1}} + \frac{n-1}{n} \cdot \frac{1}{1-p_n} - 1\right) \\
&\leq \frac{4}{n^{1-c-o(1)}} + \frac{4p_n}{1-p_n} && \text{(Using } \mathbb{E}[\mathcal{I}] \geq n^{1-c-o(1)}\text{)} \\
&\leq o(1).
\end{aligned}$$

The claim is immediate.