title: STATS 217
date: 01/06/2019

Stats 217 - Stochastic Processes

Motivating Examples

- Random Walks
  - Drunkard on the street, modeling their walk on the real number line (integer lattice $\mathbb{Z}$)
  - Starting at 0 (initial condition), steps left and right with equal probability, each step is independent of all other steps
  - Questions to answer:
    - What is the probability drunkard reaches home eventually, say position -10.
    - Expected return time to 0?
    - What is the probability he reaches home before falling down manhole, say at position 5
    - Bring in higher dimensions - d=2, d=3 same questions. There are significant qualitative differences to these questions in different dimensions. In dimensions 1, 2 will return to 0 with probability 1, but not in 3 or higher.
  - Other graphs - T_d - infinite d-regular tree. After each step only a unique edge to return to prior position - tend to float away towards infinity for trees with higher degree nodes.
  - Can approximate large finite graphs like social networks with infinite graphs
- Arrivals Process
  - People arrive at a phone booth at a rate of 10 people / hr. Spend ~ 5 minutes in the booth.
  - Can introduce independence between people to create a model. Can rely on a distribution for minutes in the booth instead of average. Then can answer questions like how long on average is the queue.
- Branching Processes
  - Galton-Watson Process - used to model human Y-chromosome DNA haplogroups - transferred from fathers to sons. Alternatively, how long does a surname last if passed from father to progeny.
  - Randomness introduced by mutations. Haplogroup defined by a set of alleles, after a certain number of mutations no longer in the same group.
  - Each man has $Pois(\mu)$ sons, independent of others. What is the probability of name extinction for varying values of μ?

Probability Review (Chapters 1 and 2)

Probability Spaces

- Sample Space: Ω - set of outcomes
- σ-algebra of events: $F,\ A \subset \Omega$
    - Algebra - some set along with some operations that can be applied
    - $\emptyset = \{\} \in F$
    - If $A \in F, \implies A^C = \{\omega \in \Omega : \omega \notin A\} \in F$ - if event A in F so is its complement
    - $A_1, A_2, \ldots \in F \implies \cup_{n=1}^{\infty} A_n \in F$ - if sequence of events in F then so is their union
    - Note: $(A^C \cup B^C)^C = A \cap B$ - then intersections are also in the sigma-algebra
- Probability Measure: $P : F \to [0,1]$. Input an event and it outputs a probability
    - $P(\emptyset) = 0$
    - Additivity: A_1, A_2, A_3 are pairwise disjoint ($A_i \cap A_j = \emptyset\ \forall i \neq j$), then $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$. In general without disjoint assumption: $P(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$ (union bound). This is intuitively since we are counting the intersection multiple times.
    - Consequences: $P(A^C) = 1 - P(A)$, $P(\Omega) = 1,\ P(A) \leq 1\ \forall A$

Random Variables

- Function from $\Omega \to \mathbb{R}$
- CDF $F_X(u) = P(X \leq u)$. $X \leq u$ is the set $\{\omega \in \Omega : X(\omega) \leq u\} \in F$
- Example: $\Omega = \{(a_i)_{i=1}^{\infty} : a_i \in \{-1, +1\}\}$ - space for random walk. Then $X_i = X_i(\omega) = a_i$. Define sum $S_n = \sum_{j=1}^{n} X_j(\omega)$. Hitting time $T_{10} = inf\{n \geq 1 : S_n = 5\}$

Discrete RVs

- X is discrete: has a countable set of outcomes - can be labeled by integers $\{x_i\}_{i=1}^{\infty}$
- PMF: $P(X = x) = p_X(x)$, and $\sum_x p_X(x) = 1$ for $p_X(x) \geq 0$

Continuous RVs

- By defn, X continuous has a PDF: $f_X(x)$ st $\forall a < b \in \mathbb{R},\ P(a \leq X \leq b) = \int_a^b f_X(x)dx$
- CDF $F_X(x) = P(X \leq x) = \begin{cases} \sum_{y \leq x} p_X(y) & \text{discrete} \\ \int_{-\infty}^{\infty} f_X(y)dy & \text{continuous} \end{cases}$

Joint Distributions

- $X_1, \ldots, X_n$ RVs, Joint CDF $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) := P(X_1 \leq x_1, \ldots, X_n \leq x_n)$
- Discrete: $P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$
- Continuous: $P(X_1 \in [a_1, b_1], \ldots, X_n \in [a_n, b_n]) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)dx_n \ldots dx_1$
- Events: $A_1, \ldots, A_n \in F$ are (mutually) independent if $P(A_1 \cap \ldots \cap A_n) = \prod_{i=1}^{n} P(A_i)$
    - Pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j) \forall i, j$. But almost always refer to mutual independence
- Variable independence if $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i),\ \forall x_1, \ldots, x_n$. Alternatively, $P(X_1 \in S_1, \ldots, X_N \in S_N) = \prod_{i=1}^{n} P(X_i \in S_i),\ \forall Si = [a_i, b_i]$ for sets S

Conditional Probability

- Events A,B with $P(B) \neq 0$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Discrete

- Conditional PMF: $p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$. Law of total probability: $\Pr\{X = x\} = \sum_{y=0}^{\infty} p(X = x|Y = y)P(Y = y) = \sum_{y=0}^{\infty} p_{X|Y}(x|y)p_Y(y)$
- Conditional expectation with specified value (Y=y is an event): $E(X|Y = y) = \sum_x xp_{X|Y}(x|y)$
- Conditioning expectation without an event. Denote $\phi(y) = E(X|Y = y)$, then $\phi(Y)$ is a RV $E(X|Y)$.
  - For example, $X \sim Unif(\{1, \ldots, 6\})$ outcome of die roll. $Y = \begin{cases} 1 & if\ X\ even \\ 0 & if\ X\ odd \end{cases} = 1(X\ even)$. Then $E(X|Y) = \begin{cases} 4 & \text{on Y =1} \\ 3 & \text{on Y = 0} \end{cases} = 4 \times 1(Y = 1) + 3 \times 1(Y = 0) = 4Y + 3(1 - Y) = 3 + Y$. Clearly this is a function of Y
  - $E(X) = E[E(X|Y)] = \sum_y p_Y(y)E(X|Y = y) = \sum_y P(Y = y + \sum_x xP(X = x|y = y) = \sum_x x \sum_y P(Y = y)P(X = x|Y = y)$
  - Summed over all y, $\sum_y P(Y = y)P(X = x|Y = y) = P(X = x)$, so the expression reduces to $E(X)$
- Conditional expectation: $E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y)$. Law of total probability: $E[g(X)] = \sum_y E[g(X)|Y = y]p_Y(y) = E\{E[g(X)|Y]\}$. In its final form, this is a function of the RV Y.

Continuous

- $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

Random Sums

- $X = \xi_1 + \cdots + \xi_N$ where N is discrete RV with pmf $p_N(n) = \Pr(N = n)$, N > 0.
- Mixed continuous X / discrete N
  - $F_{X|N}(x|n) = \frac{\Pr\{X \le x \text{ and } N = n\}}{\Pr\{N = n\}}$
  - Law of total probability: $f_X(x) = \sum_{n=0}^{\infty} f_{X|N}(x|n)p_N(n)$
- Moments assuming $E[\xi_k] = \mu$, $\mathrm{Var}[\xi_k] = \sigma^2,\ E[N] = v$, $\mathrm{Var}[N] = \tau^2$
  - $E[X] = E[\xi_k]E[N] = \mu v$
  - $\mathrm{Var}[X] = E[N]Var[\xi_k] + E[\xi_k]^2 Var[N] = v\sigma^2 + \mu^2\tau^2$
- Distribution - n-fold convolution of the density f(z), denoted $f^{(n)}(z)$
  - For $\xi_1, \xi_2, \ldots$ continuous RVs with PDFs f(z)
  - $f^{(1)}(z) = f(z)$
  - $f^{(n)}(z) = \int f^{(n-1)}(z - u)f(u)\mathrm{d}u \quad \text{for } n > 1$
  - X is continuous and has marginal density $f_X(x) = \sum_{n=1}^{\infty} f^{(n)}(x)p_N(n)$

Martingales

- Definition
  - A stochastic process is a martingale if for n = 0, 1, ...
  1. $E[|X_n|] < \infty$
  2. $E[X_{n+1}|X_0, \ldots, X_n] = X_n$
- By taking expectations of second condition, see martingale has constant mean: $E[X_0] = E[X_k] = E[X_n],\ 0 \le k \le n$
- Markov Inequality: for non negative RV X and positive constant λ, $\lambda \Pr\{X \ge \lambda\} \le E[X]$
- Maximal Inequality Theorem:
  - For a nonnegative martingale, $\Pr\{\max_{0 \le n \le m} X_n \ge \lambda\} \le \frac{E[X_0]}{\lambda}$ for $0 \le n \le m$
  - $\Pr\{\max_{n \ge 0} X_n > \lambda\} \le \frac{E[X_0]}{\lambda}$ for all n
  - The maximal inequality limits the probability of observing a large value anywhere in the time interval 0,...,m, ie. inequality limits the probability of observing a large value at any time in the infinite future of the martingale

Class Examples

1. Conditioning
   - Q: Alice and Bob take turns flipping a coin. First to flip heads wins - first mover has an advantage. What is the probability that Alice wins if she goes first?
   - Let T = # of rounds until a heads is flipped (by A or B). Event $\{T = t\}$ splits into 3 events:
     - 2(t-1) tails (before t^th round) followed by $\begin{cases} HH & A\ wins \\ HT & A\ wins \\ TH & B\ wins \end{cases}$, each has prob $(1/2)^{2t}$ - power of number of flips. $P(A\ wins|T=t) = \frac{P(A\ wins, T=t)}{P(T=t)} = \frac{2(0.5)^{wt}}{3(0.5)^{2t}} = \frac{2}{3}$
     - $P(A\ wins) = \sum_{t=1}^{\infty} P(A\ wins|T=t)P(T=t) = \frac{2}{3}\sum_{t=1}^{\infty} P(T=t) = \frac{2}{3}$
2. Branching
   - Q: Family with N distributed $Pois(\lambda)$. Each child has blue eyes with prob 1/4, brown with prob 3/4, independent of other children and N. Let Z = # of blue eyed children. What is E(Z)?
   - Let $X_i = 1(\text{ith child has blue eyes})$. Then $Z = \sum_{i=1}^{N} X_i$ - we have a sum with a random upper limit. But we can condition on the RV N: $E(Z|N) = E(\sum_{i=1}^{N} X_i|N) = \sum_{i=1}^{N} E(X_i|N)$. Note the summation becomes deterministic conditioned on N, so we can pull out of the expectation. X and N are independent, so $= \sum_{i=1}^{N} E(X_i) = N * \frac{1}{4} = N/4$
   - $E(Z) = EE(Z|N) = \frac{1}{4}EN = \frac{\lambda}{4}$

Markov Chains

- A stochastic process is a family of RVs $(X_t)_{t\in T}$ indexed by a set T (say time) characterized by 1) indexed set T, 2) state space S, the set of possible outcomes X_t, 3) Joint distributions of finite dimensional marginals, $X_{t_1}, \ldots, X_{t_n},\ \forall t_1, \ldots, t_n \in T$
- The joint distribution: need to specify $P(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n)$ for all t in T and all A in S. But generally enough to know what the pmf is. For X_t discrete (S finite) enough to specify the joint PMFs $P_{X_t1,\ldots,X_tn}(x_1, \ldots, x_n)$ for all t in T and all x in S.

Definitions and Theorems

- Chapman-Kolmogorov Equation for m step transition probability: $p^{m+n}(i,j) = \sum_k p^m(i,k)p^n(k,j)$
- Notation: probability of A given initial state x $P_x(A) = P(A|X_0 = x)$
- Time of first return (to y): $T_y = \min\{n \geq 1 : X_n = y\}$
- Probability X_n returns to y when it starts at y: $\rho_{yy} = P_y(T_y < \infty)$. Take powers to see probability of returning twice, etc.
- Stopping time $\{T = n\} = \{X_1 \neq y, \ldots, X_{n-1} \neq y, X_n = y\}$

Markov Processes

- Markov Process: the probability of any particular future behavior of the process, when its current state is known exactly, is not altered by additional knowledge concerning its past behavior. Formally, $\Pr\{X_{n+1} = j|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = \Pr\{X_{n+1} = j|X_n = i\}$ - this is the Markov property.
- One step transition probability: $P_{ij}^{n,n+1} = \Pr\{X_{n+1} = j|X_n = i\}$. When the one-step transition probabilities are independent of the time variable n, we say that the Markov chain has stationary transition probabilities (time homogeneous), ie $P_{ij}^{n,n+1} = P_{ij}$ independent of n. The probability of transitioning from one state to another is independent of time.
- Markov matrix: $\mathbf{P} = \|P_{ij}\|$ where the ith row is the probability distribution of the values of X_{n+1} under the condition that $X_n = i$. All probabilities greater than 0 and all rows sum to 1

- A discrete time **Markov chain** is a stochastic process $(X_t),\ t \in T$ with T = [0,1,2,...] and state space S (generally countable in this class) that satisfies the Markov property, defined as $\forall n \geq 0,\ \forall x_0, \ldots, x_n \in S,\ \Pr\{X_{n+1} = j | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\} = \Pr\{X_{n+1} = j | X_n = i\}$
- Time homogeneous MC when transition probabilities are independent of n. In this case define the transition matrix of entries $P_{xy} = P(x,y) := P(X_{n+1}|X_n = x) = P(X_1 = y|X_o = x)$. Note, if S is infinite then P is an infinite matrix.
- We write $\pi_x = \pi^n(x) = P(X_n = x)$: the pmf of state of chain at time n. Proposition: Distribution of time homogeneous MC is completely determined by P and π_0, the PMF of the initial state. For $n \geq 1$ and any $x_o, \ldots, x_n \in S,\ P(X_n = x_n, \ldots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) P(X_{n-1} = x_{n-1}, \ldots, X_0 = x_0)$. Then applying rules for homogeneous MC $P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P_{X_{n-1}, X_n}$, ie. the additional history does not matter. Repeating $P_{X_{n-1},X_n} P_{X_{n-2},X_{n-1}} \ldots P(X_0 = x_0)$ where $P(X_0 = x_0) = \pi_0(x_0)$.
- In book notation, a markov process is fully defined by its transition matrix and initial state X_0: $\Pr\{X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\} = p_{i_0} P_{i_0,i_1} \cdots P_{i_{n-2},i_{n-1}} P_{i_{n-1},i_n}$
- n-Step Probability Matrices
  - $\mathbf{P}^{(n)} = \left\| P_{ij}^{(n)} \right\|$ denotes the probability that the process goes from state i to state j in n transitions. $P_{ij}^{(n)} = \Pr\{X_{m+n} = j | X_m = i\}$
  - Theorem: $P_{ij}^{(n)}$ satisfies $P_{ij}^{(n)} = \sum_{k=0}^{\infty} P_{ik} P_{kj}^{(n-1)}$ for $P_{ij}^{(0)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$. This gives the conclusion: $\mathbf{P}^{(n)} = \mathbf{P}^n$
  - For distribution of state at time n: $\pi^{(n)}(y) = \sum_{X \in S} \pi^0(x) P^{(n)}(x,y) = (\pi^0 \cdot P^n)(y)$ so $\pi^n = \pi^0 P^n$ as row vectors
- Properties of Transition Matrices
  - Stochastic matrix $P(x,y) \geq 0,\ \forall x, y \in S$
  - $\sum_{y \in S} P(x,y) = 1,\ \forall x \in S$
  - Conversely any N x N stochastic matrix P and a pmf π^0 row vector give rise to a MC on {1,2,...,N} = S

First Step Analysis

- This method proceeds by analyzing the possibilities that can arise at the end of the first transition, and then invoking the law of total probability + the Markov property to establish a characterizing relationship among the unknown variables.
- For a stochastic process $(X_n)_{n=0}^{\infty}$ denote the random times $T_x = min\{n \geq 1, X_n = x\}$ for $x \in S$, $T_A = min\{n \geq 1, X_n \in A\}$ for $A \subset S$. These T's are called **return times**
- $V_x = min\{n \geq 0, X_n = x\}$ for $x \in S$, $V_A = min\{n \geq 0, X_n \in A\}$ for $A \subset S$ - these are **hitting times**. Note return or hitting times could be infinite.
- Only have $V_X \neq T_x$ if $X_0 = x$ in which case $V_x = 0$ and $T_x \geq 1$ is random
- Absorbing: For a MC with transition probabilities P_{XY} a state x is called absorbing iff $P_{xx} = 1$. Then V_X is called an absorbtion time.
- Transitory vs. Absorption Model: Transitory states in which further moves can occur and absorption states from which the model can no longer move
  - Given multiple absorption states, want to determine in which state the process gets trapped and at what time period.
  - Let $\{X_n\}$ be a finite-state Markov chain whose states are labeled 0,1,...,N. Suppose that states 0,1,..., r - 1 are transient in that $P_{ij}^{(n)} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for } 0 \leq i, j < r$ while states r,...,N are absorbing ($P_{ii} = 1 \text{ for } r \leq i \leq N$).
  - The transition matrix has the form $\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$ for $Q_{ij} = P_{ij} \text{ for } 0 \leq i, j < r$
  - The probability of ultimate absorption in state k, as opposed to some other absorbing state, depends on the initial state X0 = i. Let $U_{ik} = u_i$ denote this probability. Starting from state i, with probability $P_i k$ the process immediately goes to state k, thereafter to remain, and this is the first possibility considered. Alternatively, the

process could move on its first step to an absorbing state $j \neq k$, where $r \leq j \leq N$, in which case ultimate absorption in state k is precluded. Finally, the process could move to a transient state j < r. Then $U_{ik} = P_{ik} + \sum_{j=0}^{r-1} P_{ij} U_{jk}, \quad i = 0, 1, \ldots, r-1$

- General Absorbing Markov Chain
  - Random absorption time T: $T = \min\{n \geq 0; X_n \geq r\}$. Let us suppose that associated with each transient state i is a rate g(i) and that we wish to determine the mean total rate that is accumulated up to absorption.
  - w_i, the mean total amount for starting position $X_0 = i$: $w_i = E\left[\sum_{n=0}^{T-1} g(X_n) \mid X_0 = i\right]$
  - The sum $\sum_{n=0}^{T-1} g(X_n)$ always includes first term $g(X_0) = g(i)$. Proceeding from a future transient state j, $w_i = g(i) + \sum_{j=0}^{r-1} P_{ij} w_j \quad \text{for } i = 0, \ldots, r-1$
  - W_{ik}, the mean number of visits to state k prior to absorption: $W_{ik} = \delta_{ik} + \sum_{j=0}^{r-1} P_{ij} W_{jk} \quad \text{for } i = 0, 1, \ldots, r-1$

Stopping Times

- Hitting and return times are examples of a class of RVs for stochastic processes - **stopping times**. Define a random variable $T \in N$ is a stopping time for a stochastic process $(X_n)_{n \geq 0}$ if for any $n \geq 0$ event $\{T_n \leq n\}$ is determined by $X_0, \ldots, X_n$. In other words conditioning on the values of $X_0, \ldots, X_n$ makes the value of the indicator RV $1(T \leq n)$ deterministic.
- Example - return time TA for A in S, by observing $X_0, \ldots, X_n$ we see whether we had $X_k \in A$ for some 1 < k < n
- Non example
  - Last return time: the last time the chain visits A before the sequence stops. We need to know when the sequence stops to say this is the last visit to A, but that is a future event.
  - $L_A = sup\{n \geq 0 : X_n \, inA\} \in \{0, 1, 2, \ldots\} \cup \{\infty\}$. Conditioning on the first n steps doesn't tell you if you ever come back again. (Note, could be deterministic if you knew some information such as A is an absorbing state, but can not generalize to any chain.)
- If we were to observe the values X0,X1,..., sequentially in time and then "stop" doing so right after some time n, basing our decision to stop on (at most) only what we have seen thus far, then we have the essence of a stopping time.
- Recall hitting time $V_A = min\{n \geq 0; X_n \in A\}$. Return time $T_A = min\{n \geq 1; X_n \in A\}$ for A subset S. Both of these are stopping times
- A RV $T \in \{0, 1, 2, 3, \ldots\}$ stopping time for a stochastic $(X_n)_{n \geq 0}$ if for all n greater than 0, the event $\{T \leq n\}$ is determined by $X_0, \ldots, X_n$. Look at $X_0, \ldots, X_n$ tells us whether $X_n \in A$ by that step. For V_A, T_A we can learn whether this time T has happened looking at the X's. Note V_A is 0 if you start there, for return times we are bound by the minimum of 1.
- Kth return time: $T_A^{(k)}$ is a stopping time. $T_A(1) = T_A, \; k \geq 2 \; T_A^{(k)} = min\{n > T_A^{k-1} : X_n \in A\}$.
- Some intuition for stopping time: originates from the gambler's ruin, the gambler has made some decsion to stop after reaching a certain amount of money. It is a stopping rule, once this happens, I will stop. For it to be a stopping rule, you have to be able to tell if it has occurred - ie it must not be random anymore.

Strong Markov Property

- Regular Markov property for $(X_n)_{n \geq 0}$ condtion on $X_{n_0} = x$ then $(X_{n_0+m})_{m \geq 0}$ is a MC with $X_0 = x$.
  - If you have an MC conditioned on reaching x, then once x is reached you have an a new MC starting at x. If time homogeneous $(X_{n_0+m})_{m \geq 0}$ has same distribution as (X_n) conditioned on $X_0 = x$.
- Strong Markov Property - The above remains true even if x is a stopping time.
  - Let T be a stopping time for a MC $(X_n)_{n \geq 0}$. For any $n \geq 1, \; x_0, \ldots, x_{n-1}$ and $x, y \in S$, $P(X_{n+1} = y | T = n, X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_T = x) = P(X_{n+1} = y | T = n, X_t = x)$. Y starts as soon as X hits some state. So putting $Y_m = X_{T+m}$ conditional on $X_T = x, \; (Y_m)_{m \geq 0}$ is a MC conditioned on $Y_0 = x$. If

$(X_n)_{n\geq 0}$ time homogeneous, then $(Y_m)_{m\geq 0}$ has same distribution as $(X_n)_{n\geq 0}|X_0 = x$.
  - It is a stronger property because deterministic times are an example of stopping times. Useful for study of long time behavior of MC, since if you look after some random time, we have a MC with a distribution as if we had started at this random time.
- For example, consider $(X_n)_{n\geq 0}$ on SRW on Z (nearest neighbor transitions). $X_0 = 0$ and stopping time $T_n = min\{n \geq 1 : X_n = 10\}$. Then define $Y_m = X_{T_{10}+m}$ is a SRW on Z started at $Y_0 = 10$.

Special Markov Chains

2-State Markov Chain

- $\mathbf{P} = \begin{Vmatrix} 1-a & a \\ b & 1-b \end{Vmatrix}$ for $0 < a, b < 1$
- The n-step transition matrix: $\mathbf{P}^n = \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b}\begin{bmatrix} a & -a \\ -b & b \end{bmatrix}$. Note $\lim_{n\to\infty} \mathbf{P}^n = \begin{Vmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{Vmatrix}$ - The system, in the long run, will be in state 0 with probability b/(a + b) and in state 1 with probability a/(a + b), irrespective of the initial state in which the system started.
- Independent RV's - all rows of P are identical, since $X_{n+1} \perp X_n$

One-Dimensional Random Walks

- Markov chain with finite or infinite state space. If the particle is in state i, it can either stay in i or move to either i+1, i-1.
- Transition matrix $P = \begin{bmatrix} p_0 & 0 & \cdots & 0 & \cdots & & \\ r_1 & p_1 & & \dots & 0 & \dots & \\ 2 & r_2 & & \dots & 0 & \dots & \\ \vdots & & & & & & \\ & 0 & & q_i & r_i & p_i & 0 \\ & & & \ddots & & & \ddots \end{bmatrix}$ for $p_i > 0, q_i > 0, r_i \geq 0, \text{ and } q_i + r_i + p_i = 1, i = 1, 2, \dots (i \geq 1), p_0 \geq 0, r_0 \geq 0, \ r_0 + p_0 = 1$
- If $X_n = i$ for i > 1, $\Pr\{X_{n+1} = i+1|X_n = i\} = p_i, \ \Pr\{X_{n+1} = i-1|X_n = i\} = q_j, \ \Pr\{X_{n+1} = i|X_n = i\} = r_i$

Success Runs

- Transition matrix $\mathbf{P} = \begin{Vmatrix} p_0 & q_0 & 0 & 0 & 0 & \cdots \\ p_1 & r_1 & q_1 & 0 & 0 & \cdots \\ p_2 & 0 & r_2 & q_2 & 0 & \cdots \\ p_3 & 0 & 0 & r_3 & q_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{Vmatrix}$ for $q_i > 0, \ p_i > 0, \ p_i + q_i + r_r = 1$
- The zero state plays a distinguished role in that it can be reached in one transition from any other state, while state i + 1 can be reached only from state i.
- Useful for applications counting successes in a row, renewal processes like lightbulb age - resets at burnout.

Branching Processes

- Say organisms produce ξ offspring, $Pr\{\xi = k\} = p_k \quad \text{for } k = 0, 1, 2, \dots$
- The process $\{X_n\}$ where X_n is the population size at the nth generation, is a Markov chain of special structure called a branching process.
- In the nth generation, the Xn individuals independently give rise to numbers of offspring $\xi_1^{(n)}, \xi_2^{(n)}, \dots, \xi_{X_n}^{(n)}$ - the cumulative number produced for the (n + 1)st generation is $X_{n+1} = \xi_1^{(n)} + \xi_2^{(n)} + \cdots + \xi_{X_n}^{(n)}$
- Mean: recursively defined $M(n+1) = \mu M(n)$, generally $M(n) = \mu^n \quad \text{for } n = 0, 1, \dots$ where X_n is the population size at time n and M(n) is the mean of X_n

- The mean population size increases geometrically when μ > 1, decreases geometrically when μ< 1, and remains constant when μ = 1.
- Variance: recursively defined $V(n+1) = \sigma^2 M(n) + \mu^2 V(n)$, generally $V(n) = \sigma^2 \mu^{n-1} \times \begin{cases} n & \text{if } \mu = 1 \\ \frac{1-\mu^n}{1-\mu} & \text{if } \mu \neq 1 \end{cases}$
  - The variance of the population size increases geometrically if μ > 1, increases linearly if μ = 1, and decreases geometrically if μ < 1

Extinction

- The random time of extinction N is thus the first time n for which Xn = 0, and then, obviously, Xk = 0 for all $k \geq N$.
- In Markov chain terminology, 0 is an absorbing state, and we may calculate the probability of extinction by invoking a first step analysis.
- Defn $u_n = \Pr[N \leq n] = \Pr\{X_n = 0\}$, be the probability of extinction at or prior to the nth generation, beginning with a single parent X0 = 1. The k subpopulations independent, w/ original statistical properties. Each has probability of dying out in n-1 generations equal to u_{n-1}.
- Probability that all die out in n-1 generations is u_{n-1}^k by independence. Then weighting by probability of k offspring:
- $u_n = \sum_{k=0}^{\infty} p_k (u_{n-1})^k$

Class Examples

1. Stochastic Processes - Simple random walk on Z
   - $T = \{0, 1, 2, \ldots\},\ S = Z$. Given the history of the walk, we derive the n + 1 position: $P(X_{n+1} = k | X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = k | X_n = x_n)$. This is the Markov property, $= \begin{cases} 1/2 & k = X_n + 1 \ or \ X_n - 1 \\ 0 & else \end{cases}$. Gives joint PMFs pf $(X_0, \ldots, X_n)$ for any n, get PMFs $(X_{t1}, \ldots, X_{tn})$ by summing out the extra variables.
   - Ex - $P_{X_0,X_2}(x_0, x_2) = \sum_{x_1} P_{X_0,X_1,X_2}(x_0, x_1, x_2)$ - summing over the probabilities for x1 gives you the marginal desired.
2. Galton Watson Branching Process / Tree
   - X_t = size of the family at generation t. Say X0 = 1, starting with one individual. S= [0, 1, 2, ...] = T.
   - Specifying the distributions: joint distributions of $X_{t1}, \ldots, X_{tn}$ determined recursively by $X_{t+1} = \sum_{k=1}^{X_t} Y_{t,k}$. Define $Y_{t,1}, \ldots, Y_{t,X_t} \sim_{iid} Pois(\lambda)$ conditional on X_t; the Y's are the number of children each person has at a generation, where X's are the state of the family at time t. Note another sum up to a RV up to Xt.
     - Could alternatively draw $\{Y_{t,k}\}_{t=0,k=1}^{\infty,\infty}$ iid Pois beforehand, but would not use a lot of those variables.
3. Gambler's Ruin
   - Game where we win in each round (independent) 1 dollar with probability p=0.4 and lose one dollar with probability 1-p=0.6. Decide ahead of time to quit once we reach N dollars. The game is over once we reach 0 dollars. X_n = amount of money we have after n rounds. State space S = {0,1,...,N} - between and including the absoption states.
   - If still playing $X_n \in [1, N-1],\ P(X_{n+1} = x_n + 1 | X_n = x_n, \ldots, X_0 = x_0) = 0.4$ and $P(X_{n+1} = x_n - 1 | X_n = x_n, \ldots, X_0 = x_0) = 0.6$
   - $P_{0,0} = 1,\ P_{N,N} = 1$ - if you start at either end state the game is immediately over. At every other row, have (0.6,0,0.4) centered at the column equal to the row position. Alternatively, could create a directed graph of the possible state transitions with edge weights equal to the probabilities with that state transition - this makes P an adjacency matrix.
   - Now ask what are the chances of falling into 0 or N and how long does it take to get there? Given $X_0 = x \in \{1, 2, 3\}$

- $P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 \\ 0 & 0.6 & 0 & 0.4 \\ \dots \end{bmatrix}$ Computer calculations suggest $lim_{n\to\infty}(P^n) = \begin{bmatrix} 1 & 0 \\ 57/65 & 8/65 \\ 45/65\dots 0 \dots & 20/65 \\ 27/65 & 38/65 \\ 0 & 1 \end{bmatrix}$ with zeroes for all middle columns.

4. Ehrenfest Chain
   - Stat Physics model for two equal sized containers of gas connected by a small opening. Expect equilibrium eventually with same number of molecules in each container (balls, urns).
   - N = total # of balls (order 10^23), particle exchange modeled as a random process, pick 1 ball uniformly at random a move to the other urn. Let $X_n =$# of balls in the left urn after the nth draw, change +- 1 at each step.
   - $P(X_{n+1} = x+1|X_n = x) = \frac{N-x}{N}$ = # balls in right urn / total N. $P(X_{n+1} = x-1|X_n = x) = \frac{x}{N}$ = # balls in left urn / total N. Note observes the markov property.
   - 1) How long until $x_n \approx N/2$? 2) Does it stay there? 3) How much does it fluctuate? 4) How often does the chain reach endpoints 0,N?
   - For N=4, S = [0,1,2,3,4], $P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$. Similar structure to Gambler's ruin, except:
     - Probability of moving right / left not constant
     - 0,4 are now repelling instead of absorbing states.
     - Long term behavior should be very different - gravitate to the middle rather than the endpoints
5. Simple Random Walk (SRW)
   - Let G = (V, E) be a graph. ||V|| = 6, E=[[1,2], [3,4], [3,5], [4,5], [5,6]].
   - SRW on G: state space S = V, transition probabilities $P(x,y) = \begin{cases} 1/degree(x) & for\ \{x,y\} \in E \\ 0 & otherwise \end{cases}$ - that is at each time step move along an edge at random chosen uniformly at random from set of edges leading out of current state.
   - For V finite, is the distribution of $X_t \approx uniform$ on V after a long time? For our example, certainly not since 1-2 and 3-4-5-6 are separate CCs, whichever component contains X_0 contains $X_t\ \forall t \geq 0$. Additionally, vertices have different degrees, leading to different probabilities of visiting those nodes. Can come up with less intuitive examples as well, such as a connected 2D square - only can visit some vertices on odd states, others on even states.
   - Recalls the drunkard on the street, that is a SRW on G with V = Z and nearest neighbor edges $E = \{\{k, k+1\} : k \in \mathbb{Z}\}$
6. iid Sequence
   - Let $\{X_n\}_{n=0}^{\infty}$ be iid RVs taking values in a countable set S with distribution $p(x) = P(X_n = x)$ then $P(X_{n+1} = x|X_n = x, \dots, X_0 = x_0) = P(X_{n+1} = x|X_n = x) = P(X_{n+1} = x) = p(x)$
   - Transition matrix $P_{X,Y} = P(X_1 = y) = p(y)$ - P has identical rows all equal to the row vector p
7. Deterministic Chain
   - MC with X_t, S = [1,2,3,4] and P(1,2) = P(2,3) = P(3,4) = P(4,1) = 1 and P(x,y) = 0 - think of moving around edges of a box
   - X_t is completely deterministic conditional on X_0 but it is still an MC
   - Color state 1 red, 2,4 green, 3 blue. Define $Y_t =$ color of X_t. Is Y an MC? No, consider if previous color is green, we may be going to red or blue with probability 1 depending on if the color prior to green was red or blue, ie $P(Y_2 = blue|Y_1 = green, Y_0 = red) = 1 \neq 0 = P(Y_2 = blue|Y_1 = green, Y_0 = blue)$ Distribution of the next state depends on both current and previous states. The fact that MC is deterministic did not cause this problem

though.

8. Exit Distribution for Gambler's Ruin
   - $X_n \in [1, N-1],\ P(X_{n+1} = x_n + 1 | X_n = x_n, \dots, X_0 = x_0) = 0.4$ and $P(X_{n+1} = x_n - 1 | X_n = x_n, \dots, X_0 = x_0) = 0.6$ and N=4. 0 and 4 are absorbing states
   - Q: For $X_0 = x \in \{1,2,3\}$ what is the probability we win? What is the probability $P_x(V_4 < V_0) = h(x)$ - hitting time for 4 precedes the hitting time for 0. Notation: for an event E and $x \in S$ write $P_x(E) = P(E|X_0 = x),\ E_X(Y) = E(Y|X_0 = x)))$
   - h(0) = 0 - the probability that we win given the first state is 0
   - h(4) = 1 - the probability that we win given we start at 4
   - $h(1) = P(V_4 < V_0 | V_0 = 1) =$ $P(V_4 < V_0 | X_1 = 0, X_0 = 1)P(X_1 = 0|X_0 = 1) +$ $P(V_4 < V_0 | X_1 = 2, X_0 = 1)P(X_1 = 2|X_0 = 1)$ $= h(0))P(X_1 = 0|X_0 = 1) + h(2)P(X_1 = 2|X_0 = 1) = 0(0.6) + h(2)(0.4)$
   - Similarly, $h(2) = h(1)(0.6) + h(3)(0.4)$, $h(3) = h(2)(0.6) + h(4)(0.4) = h(2)(0.6) + 0.4$. Now have a system of linear equations and can solve for the hitting times. Let $h(1) = a,\ h(2) = b,\ h(3) = c$, then $a = 2b/5,\ b = 3a/5 + 2c/5,\ c = 3b/5 + 2/5$. Solving the system, $c = 38/65,\ b = 20/65,\ a = 8/65$
   - Q: How long does the game take - expected playing time conditional on different starting points?
   - Need to compute $g(x) = E_x V_A$ with A = [1,4] - the set of absorbing states. $g(0) = 0,\ g(4) = 0$ since we are already at the absorbing states (if we used T, they would equal 1).
   - $g(1) = E[V_A | X_0 = 1] = E[V_A | X_1 = 0, X_0 = 1]0.6 + E[V_A | X_1 = 2, X_0 = 1]0.4$ $= (1 + g(0))(0.6) + (1 + g(2))(0.4) = 1 + 0.4g(2)$
   - $g(2) = 1 + 0.6g(1) + 0.4g(3)$, $g(3) = 1 + 0.6g(2) + 0.4g(4) - 1 + 0.6g(2)$. Solving the system we get g(1) = 33/13, g(2) = 50/13, g(3) = 43/13
9. Repeated coin toss
   - We toss a fair coin repeatedly and independently recording results as $X_n \in \{H, T\}, n \geq 1$. What is the expected # of times before we see the pattern HTH?
   - Attempt 1: define $Y_n = (X_n, X_{n-1}, X_{n-2})$, Transition probability P(HHT, HTT) = 1/2 and P(HHT,TTT) = 0 for some examples. Say $X_{-1} = X_{-2} = T$ or Y0 = TTT as a good starting state.
   - Attempt 2: define a MC $(Y_n)_{n \geq 1}$ with S = [0,1,2,3] defined as the largest length l for which the most recent l flips $(X_{n-l+1}, \dots, X_n)$ match the first l letters of HTH. Y measures how far along in the sequence of HTH we are. For instance, if $Y_n = 0\ if\ X_n = T$ and $X_{n-1} = T$, then $Y_n = 1$ if $X_n = H$ but $(X_{n-2}, X_{n-1}) \neq (H, T)$, $Y_n = 2\ if\ (X_{n-1}, X_n) = (H, T)$, and $Y_n = 3$ if $(X_{n-2}, X_{n-1}, X_n) = (H, T, H)$.
     - From 0, we transition to 0 with 1/2 or to 1 with 1/2. (If tails or heads respectively)
     - From 1, we transition to 1 with 1/2 or to 2 with 1/2 (HH, HT respectively)
     - From 2, we transition to 3 with 1/2 or 0 with 1/2 (HTT, HTH respectively)
     - From 3 go to 1 or 2 with equal probability, but doesn't really matter since we have achieved HTH at 3
     - E(time until HTH) = $E_0 V_3$. Let g(x) = $E_X V_3$ then from the first step analysis,
     - g(0) = 1 + g(0)/2 + g(1)/2
     - g(1) = 1 + g(1)/2 + g(2)/2
     - g(2) + 1 + g(0)/2 + g(3)/2
     - g(3) = 0
     - Note the plus 1 ensures we are taking the minimum amount of time required to get to each stage. Solving we get g(0) = E(time until HTH) =10.

Long Term Behavior of MCs

Definitions and Theorems

- Time of kth return: $T_y^1 = T_y$ and k > 1, $T_y^k = \min\{n > T_y^{k-1} : X_n = y\}$. The probability that we return k times is

$P_y\left(T_y^k<\infty\right)=\rho_{yy}^k$

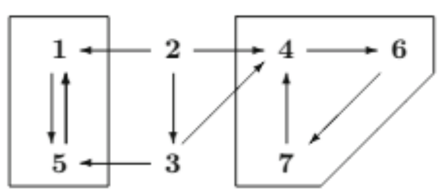
- Transient: $\rho_{yy}<1$ and $\rho_{yy}^k \to 0 \text{ as } k \to \infty$. The number of time periods that the process will be in state y $\sim geom(\frac{1}{1-\rho_{yy}})$
- Recurrent: $\rho_{yy}=1$ and $\rho_{yy}^k=1$ (absorbing state the strongest example of recurrence). Recurrent states visited infinitely often.
- Communication: x communicates with y if $\rho_{xy}=P_x\left(T_y<\infty\right)>0$
- If $\rho_{xy}>0, \text{ but } \rho_{yx}<1$, x is transient. If x recurrent and $\rho_{xy}>0, \text{ then } \rho_{yx}=1$
- Closed: impossible to leave closed set, $p(i,j)=0$ for $i\in A, j\notin A$
- Irreducible: set B irreducible if whenever $i,j\in B$ i and j communicate. The whole MC is irreducible if all states communicate with each other.
- Theorem 1.7: If C is a finite closed and irreducible set, then all states in C are recurrent.
- Lemma 1.9. If x is recurrent and $x\to y$, then y is recurrent.
- Lemma 1.11. $E_xN(y)=\rho_{xy}/\left(1-\rho_{yy}\right)=\sum_{n=1}^{\infty}p^n(x,y)$. Recurrence defined by $E_yN(y)=\infty$, expected # of time periods that process is in state y is infinite.
- Stationary Distribution: If $\pi p=\pi$, then π is called a stationary distribution. If the distribution at time 0 is the same as the distribution at time 1, then by the Markov property it will be the distribution at all times
- Doubly stochastic: transition matrix whose columns sum to 1
- Detailed balance condition: $\pi(x)p(x,y)=\pi(y)p(y,x)$
- Period: The period of a state is the largest number that will divide all the n $\geq$ 1 for which $p^n(x,x)>0$
- Lemma 1.17. If $\rho_{xy}>0 \text{ and } \rho_{yx}>0$, then x and y have the same period - periodicity is a class property. If $p(x,x)>0$ then x has period 1.
- Define I: p is irreducible, A: aperiodic (all states have periood 1), R: all states recurrent, S: stationary distribution π exists
- Convergence Theorem: $p^n(x,y)\to\pi(y)$ as $n\to\infty$ for I, A, S
- Ergodic: State i is positive recurrent if i is recurrent and, starting in i, the expected return time to i is finite. Positive recurrent, aperiodic states are called ergodic.
- Limiting Probabilities: An irreducible, ergodic MC has limit independent of i $\pi_j=\lim_{n\to\infty}P_{ij}^n$, where π_j is the solution to the system $\pi_j=\sum_{i=0}^{\infty}\pi_iP_{ij}, \quad j\geq 0 \text{ s.t. } \sum_{j=0}^{\infty}\pi_j=1$. Note π_j also the long run proportion of time that the process will be in state j

Return Times

- Notation (see Durrett): $\rho_{xy}=P_x(T_y<\infty)=P(T_y<\infty|X_0=x)$ probability you ever return to y
- Time homogeneous case
  - Kth return times: For say K = 2, $P_y(T_y^{(2)}<\infty)$ - probability that if we start at y, we will come back at least twice.
  - $P_y(T_y^{(2)}<\infty)=\sum_{n\geq 1}P(T_y^{(1)}=n,T_y^{(2)}<\infty|X_0=y)=\sum_{n\geq 1}P(T_y^{(2)}<\infty|X_0=y,T_y^{(1)}=n)P_y(T_y^{(1)}=n)$. First time must be finite, then the probability that the second time must also be finite. Conditioning on $X_0=y,T_y^{(1)}$ implies at time n we are at y, so we start a new MC once we have reached n. We can forget about $X_0=y$ since it is now redundant by the SMP.
  - In the language of SMP: $(X_{T_y^{(1)}})_{m\geq 0}=^d(X_m)_{m\geq 0}|X_0=y$
  - So $P_y(T_y^{(2)}<\infty)=\sum_{n\geq 1}P_y(T_y^{(1)}<\infty)P_y(T_y^{(1)}=n)$, where $P_y(T_y^{(1)}<\infty)=\rho_{yy}$. So $=\rho_{yy}\sum_{n\geq 1}P_y(T_y^{(1)}=n)=\rho_{yy}P_y(T_y^{(1)}<\infty)=\rho_{yy}^2$ Therefore to return twice, it just needs to happen once twice in separate MCs - so you get the square of rho.
  - Repeating: $P_y(T_y^{(k)}<\infty)=\rho_{yy}^k$ for time homogeneous MCs
- SMP: interarrival times $\Delta_y^{(k)}=T_y^{(k)}T_y^{(k-1)},\ k\geq 1$ with $T_y^{(0)}=0$, once you return to y, it's as if we are starting over and these are iid variables. So all have distribution of $\Delta_y^{(1)}=T_y^{(1)}-0=T_y$
  - So $T_y^{(k)}=\sum_{j=1}^k\Delta_y^{(j)}$, then $P_y(T_y^k<\infty)=P(\sum_{j=1}^k\Delta_y^{(j)}<\infty)=P_y(\Delta_y^{(j)}<\infty\ \forall\ 1\leq j\leq k)=P_y(T_y<\infty)^k=\rho_{yy}^k$

Classifying States

- Definition: If $y \in S$ such that $\rho_{yy} < 1$, then $\rho_{yy}^k \to 0$ as $k \to 0$ and say y is a transient state. $\rho_{yy} = 1 \implies$ chain returns to y infinitely many times with probability 1 - say y is a recurrent state.
- Example: An absorbing state is recurrent. $P_{X,X} = 1$ - once you are there you stay there.
- Let number of visits to Y equal $N_Y := \sum_{n\geq 1} 1(X_n = y)$ not counting time n=0. Condition on $X_0 = y$ how many time will we return to y, $1 + N_Y \sim \begin{cases} geom(1-\rho_{yy} & \text{if y transient} \\ \infty & \text{if y recurrent} \end{cases}$. Note $1 - \rho_{yy}$ is probability of never returning again. In particular, $E_Y N_Y = \begin{cases} \frac{\rho_{yy}}{1-\rho_{yy}} & \text{y transient} \\ \infty & recurrent \end{cases} = E_y \sum_{n\geq 1} 1(X_n = y) = \sum_{n\geq 1} P_y(X_n = y) = \sum P_{y,y}^n$. What we have shown is Y is transient iff $\sum P_{y,y}^n = E_Y N_Y < \infty$. Serves as another definition of transience.
  - Example: Gambler's Ruin. State space is the number of dollars you have, 0 and N are absorbing states, 0.6 probability of going down, 0.4 to go up taking steps of 1 at a time. 0, N recurrent / absorbing. For $k \in \{1, 2, \ldots, N-1\}$:
    - State 1: $P_1(T_1 = \infty) \geq P_{1,0} = 0.6 > 0 \implies 1$ is transient. At N-1, $P_{N-1}(T_{N-1} = \infty) \geq 0.4 > 0$
    - $P_k(T_k = \infty) \geq P_{k,k-1}P_{k-1,k-2}\ldots P_{1,0} = 0.6^k > 0 \implies$ k is transient.
- Proposition 1: Suppose S is finite, $S \in \{1, 2, \ldots, N\} \to P$ N x N matrix and suppose $P_{X,Y} \geq p_0 > 0,\ \forall x, y \in S$ Then all states are recurrent.
  - Proof: Let $x \in S$ arbitrary, need to show $P_X(T_X = \infty) = 0$, where $P_X(T_X = \infty) = 1 - \rho_{xx}$. Let's consider the probability of starting from X is at least n $P_X(T_X > n) = P_X(X_1 \neq x, \ldots, X_n \neq x) = P_X(X_n \neq x | X_1 \neq x, \ldots, X_{n-1} \neq x)P(X_1 \neq x, \ldots, X_{n-1} \neq x)$. Now we can use the Markov property.
  - Then $P_X(X_n \neq x | X_1 \neq x, \ldots, X_{n-1} \neq x)P(X_1 \neq x, \ldots, X_{n-1} \neq x)$ bounded above (at most) $max_{y\neq x}(1 - P_{y,x}) \leq 1 - p_0$. Applying this inductively, $P_X(T_X > n) \leq (1 - p_0)^n$. Probability than you avoid going to x n-times in a row is going to 0 exponentially fast. OTOH, bounded below by $P_X(T_X = \infty)$ which implies that $P_X(T_X = \infty) = 0$ (by are upper bound converging on an arbitrarily small number).
- Definition: (Following Durrett) Say x communicates with y, written $x \to y$, if starting from X $\rho_{xy} = P_X(T_Y < \infty) > 0$. Say x and y communicate if $x \to y,\ y \to x$ write $x \leftarrow y$. Note this is a transitive property $x \to y,\ y \to z \implies x \to z$. Transitive relation on the state space, allowing us to break the state space into chunks.



  - Motivating example: see drawing (Durrett 17). 2 and 3 are transient - as soon as they leave these states there are no arrows back to them. Very strong case of transience, just need a positive probability that they never return but here that probability is 1 - a guarantee. 1 and 5 communicate with each other. $4 \to 6 \to 7$. $2 \to$ all others, but no other state communicates with 2. $3 \to 1, 5, 4, 6, 7$ (all except 2). $2 \to 3$. In summation, blocks of 1,5 recurrent (R1), 2,3 transient (T), 4,6,7 recurrent (R2). Note also, x does not communicate with y for all x in R1 and all y in R2. Within the components, $x \to y$ for all x,y in R1 or x,y in R2.
- Definition: A set $A \subset S$ is "closed" if it is impossible to get out - $P_{x,y} = 0 \forall x \in A, y \in A^C$.
  - In example, closed sets are $\{1, 5\} = R1,\ \{4, 6, 7\} = R2$. Also $R1 \cup R2$ and $R1 \cup R2 \cup \{3\}$ are closed. Additionally the whole state space S is closed.
- Definition: A set $A \subset S$ is irreducible if $x \to y$ for all x,y in A.
  - In example, R1 and R2 are the irreducible sets.
- **Theorem 1 (1.7)**: If $C \subset S$ is finite, closed, irreducible, then all states in C are recurrent. (Note S may be infinite and theorem still holds)

- Proof follows from following two lemmas:
- Lemma 1.9: If x is recurrent, and $x \to y$, then y is recurrent.
  - Proof: By Lemma 1.6, $y \to x$. Let j,l be powers s.t. $p^j(y,x) > 0,\ p^l(x,y) > 0$. Now let's take $\sum_{k=0}^{\infty} p^{j+k+l}(y,y) \geq p^j(y,x)\left(\sum_{k=0}^{\infty} p^k(x,x)\right)p^l(x,y)$ with left and right terms positive and middle infinite. Recall $E_x N_x = \sum_{n\geq 1} p^n(x,x,) = \infty$ for $N_x = \sum_{n\geq 1} 1(X_n = x)$ and recurrent x (alternate defn of recurrence). Then $\sum_{k=0}^{\infty} p^k(y,y) = \infty \implies$ y is recurrent.
- Lemma 1.10: In a finite closed set there has to be at least one recurrent state.
  - Proof: By contradiction, let's assume all states are transient.
  - Lemma: Expected number of visits to y starting from x = $E_x N_y = \frac{\rho_{xy}}{1-\rho_{yy}}$
  - By this lemma, since all states transient $E_x N_y < \infty,\ \forall x, y \in C$. Since C is finite, $\infty > \sum_{y\in C} E_x N_y = \sum_{y\in C}\sum_{n=1}^{\infty} p^n(x,y)$. Swapping sums, $\sum_{n=1}^{\infty}\sum_{y\in C} p^n(x,y) = \sum_{n=1}^{\infty} 1$ because C is closed, so we will definitely land on a y in C. Then saying $\infty > \sum_{n=1}^{\infty} 1 = \infty$, which is a contradiction.

- **Decomposition Theorem 1.8**: If S is finite, then can express S as a disjoint union $S = T \cup R1 \cup \ldots \cup Rk$ where T is the set of all transient states, and $R1, \ldots, Rk$ closed, irreducible sets of recurrent states.
  - Theorem 1.5: Let $x \in S$ if there exists $y \in S$ s.t. $x \to y, \rho_{xy}$ and $\rho_{yx} < 1$ then x is transient.
    - Proof: Let $m = min\{k; p^k(x,y) > 0\}$, ie. the minimum number of states in which it is possible to go from x to y. There exists $y_1, \ldots, y_{m-1}\ inS$ distinct steps s.t. $p(x,y_1), p(y_1,y_2), \ldots, p(y_{m-1},y_m) > 0$ since otherwise could find a shorter path. $P_X(T_X = \infty)$ (starting from x, the probability that we never return), is at least $p(x,y)\ldots p(y_{m-1},y)P_y(T_x = \infty)$ (lower bound since this is one such way we could never return to x). Then note $p(x,y)\ldots p(y_{m-1},y) > 0$ and $P_y(T_x = \infty) = 1 - \rho_{yx} > 0$ - thus x is transient.
  - Lemma 1.6: If x is recurrent and $x \to y$ then $\rho_{yx} = 1$ - ie. we must eventually return to x from y, since if it were less than 1 x would be transient.
  - Decomposition Proof: Assuming theorem 1. Let $T = \{x \in S\}$ s.t. there exists $y \in S$ with $x \to y$ but y does not communicate with x, then by Thm 1.5, x must be transient for all x in T. Remains to divide $S \setminus T = S \cap T^C$ in closed irreducible sets of recurrent states. Let $x \in S \setminus T$ arbitrary and $C_x = \{y \in S : x \to y\}$. Now find all of the state that x can communicate with, this is C_x and we want to show it is one of these closed irreducible sets.
    - Claim 1: C_x is closed. If it were not closed, then there would be some state with a transition from inside set to outside. For some $y \in C_x \implies x \to y$ and $y \to z$ for $z \notin C_x$, then by transitivity $x \to z$. Then by definition $z \in C_x$, leading to a contradiction.
    - Claim 2: C_x is irreducible - any state communicates with any other in C_x. Let $y, z \in C_x$ arbitrary. By Lemma 1.6, $y \to x$, since $x \notin T$ and $x \to y$. Then $y \to x \to z$, so $y \to z \implies C_x$ irreducible, since y,z arbitrary.
    - Put $R1 = C_x$. Saw R1 closed, irreducible. If $T = T \cup R1$, we are done. Otherwise, pick some $x' \in S \setminus (T \cup R1)$ and repeat to find $R2 = C_{x'}$. Terminates in $S = T \cup R1 \cup \ldots \cup Rk$ since S is finite. All states in $R1 \cup \ldots \cup Rk$ are recurrent by Theorem 1.

Limiting and Stationary Distributions

- Definition: A probability distribution π on S is a **stationary distribution** if $\pi p = \pi$ for π = row vector. In other words, $\pi(y) = \sum_{x\in S} \pi(x)p(x,y)\ \ \forall y \in S$.
  - For a time homogeneous MC with transition matrix P
  - π is a left eigenvector of P with associated eigenvalue 1
- Example: Social Mobility Chain
  - $X_n \in S = \{L, M, U\}$ social class of nth generation of a family / lineage. We have transition matrix $\begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}$. Given initial distribution π^0 (a) does $\pi^n \to$ limiting distribution and (b) what is it?
  - Finding the stationary distribution - just a linear algebra problem of finding an eigenvector. Solve $\pi p = \pi$ for

$\pi = (a, b, c)$. Equations $1) \, a + b + c = 1$. We have more equations than unknowns, but system will be

$$2) \, 10a = 7a + 3b + 2c$$
$$3) \, 10b = 2a + 5b + 4c$$
$$4), 10c = a + 2b + 4c$$

consistent because the rows sum to 1.

  - Solve the system. Could express this as $\pi(p - I) = 0, \; \pi \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 1$, so plugging in $\pi \left(p - I | \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) = (0, 0, 0, 1)$, but one of these 4 columns in p - I is redundent, so can remove the third column say. In our system $(a, b, c) \begin{bmatrix} x & y & 1 \\ a & b & 1 \\ c & d & 1 \end{bmatrix} = (0, 0, 1)$ (plugging in numbers from our equation). Our matrix is invertible since we got rid of the dependent column, so we can solve by taking the last row of M^{-1}. We get $\pi = (\frac{22}{47}, \frac{16}{47}, \frac{9}{47})$ - this is the stationary distribution.

- Periodicity: Given $S = S_1 \cup S_2$ where sets only communicate with each other but not internally - bipartite graph. π will not converge as $n \to \infty$ regardless of starting position.
  - For state $x \in S, \; I_x = \{n \geq 1 : p^n(x, y) > 0\}$, set of possible return times, we say the period of state x is the greater common divisor (gcd) of set.
  - Aperiodic: x is aperiodic if it has period 1. Note that if gcd = 1, chain is aperiodic, even if it cannot return to a given state in 1 step (eg. periods of 2 and 3).
- Convergence Theorem: Suppose S (finite or infinite) is irreducible, aperiodic, and stationary distribution exists (IAS), then for all x,y in S $p^n(x, y) \to \pi(y)$ as $n \to \infty$ and $\pi^n(y) \to \pi(y)$.
- Theorem 1.22 (Ergodic): Suppose S (finite or infinite) and I, S, and $\sum_x |f(x)| \pi(x) < \infty$ then $\frac{1}{n} \sum_{m=1}^{n} f(X_m) \to \sum_x f(x) \pi(x)$
  - Let $f(y) = 1(x = y)$ then $\frac{1}{n} \sum_{m=1}^{n} f(X_m)$ = # of visits to state x = $\frac{N_n(x)}{n} = \pi(x)$
- When does a stationary distribution exist?
- Theorem: If S is **finite and irreducible** then there exists a unique stationary distribution π and moreover $\pi(x) > 0$ for all $x \in S$
  - In the Gambler's ruin not all states were positive in the limit - just the absorbing states have probability. This is not an irreducible chain
- When do we have π^n, the distribution of the pmf of X_n, converging to the stationary distribution π?
  - Irreducibility guarantees existence of stationary distribution but not enough to guarantee a distribution in the limit
  - By decomposition theorem, we know $\pi^n(x) \to 0, \; \forall x \in T$, so it suffices to consider the $lim \, \pi^n$ on the closed irreducible sets R1,...,Rk. Once $X_n \in Rj$ for some j does pi^n converge to the stationary distribution for Rj? No!
  - Example: 2 states 1 and 2, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. This matrix has a left eigenvector with eigenvalue 1 $\pi = (1/2, 1/2)$, the stationary distribution by our theorem above, since finite and irreducible. But the distribution of X_n never converges - $p^n(1, 2) = \begin{cases} 1 & \text{n odd} \\ 0 & \text{n even} \end{cases}$. Conditional on $X_0 = 1, \; \pi^n(1) \begin{cases} 0 & \text{n odd} \\ 1 & \text{n even} \end{cases}$ and this sequence does not converge to $\pi(1) = 1/2$. We have a stationary distribution but not a limiting distribution. But the stationary distribution does tell us the **proportion of time** spent in any one state - π does give limiting proportion of time spent at each state.
- Convergence Thorem: Convergence to stationary distribution if chain is irreducible, all states aperiodic, and has a stationary distribution

- **Ergodic Theorem**: If S is irreducible and chain has stationary distribution π, then for any $f : S \to \mathbb{R}$ with $\sum_{x\in S} |f(x)|\pi(x) < \infty$ (average mod with stationary distribution is finite, note automatric if S is finite), then with probability 1, get a LLN $\frac{1}{n}\sum_{m=1}^{n} f(X_m) \to \sum_{x\in S} f(x)\pi(x)$
    - This looks like the LLN, but this is more general. We have a sequence not iid with Markov dependency, but we are averaging a fixed number of variables that is a function of a MC.
    - $\sum_{x\in S} f(x)\pi(x) = E(f(x))$ for $x \sim \pi$
    - If we take $f = 1_A,\ A \subset S$ we have $\frac{1}{n}\sum_{m=1}^{n} I(X_n \in A)$, the proportion of time spent in A up to time n. $\frac{1}{n}\sum_{m=1}^{n} I(X_n \in A) \to \sum_{x\in A} \pi(x) = \pi(A)$, the stationary measure for that subset of states. The stationary distrbution of $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ does not converge, but we can use this theorem to see its stationary distribution of (1/2,1/2) in a useful way.
- **Theorem 1.21**: Suppose S (finite or infinite) is irreducible and all states are recurrent, the number of visits to y up to time n, $N_n(y) = \sum_{m=1}^{n} 1(X_n = y)$. From ergodic theorem we know $\frac{N_n(y)}{n} \to \pi(y)$ but we also can say $\frac{N_n(y)}{n} \to \frac{1}{E_y T_y},\ \forall y \in S$ converges a.s., the reciprocal of the expected return time. The expected return time is sort of a period, so the reciprocal of a period gives you a frequency. If you take a long time to return on average, you will spend a lower proportion of time in that state.
    - In particular, together with ergodic frequency interpretation, if the chain has a stationary distribution π, then $\pi(y) = \frac{1}{E_y T_y}$. From this we can conclude π is a unique stationary distribution, since we found a unique formula for π.
    - For a MC with S irreducible and all states recurrent, there is a unique stationary distribution given by $\pi(y) = \frac{1}{E_y T_y}$.
    - Idea of proof: Just case $E_y T_y < \infty$. Suppose $X_0 = y$, by the SMP the interarrival times (the time between arrivals to state y) $\tau_1, \tau_2, \ldots \overset{iid}{\sim} T_y$. By strong LLN, $\frac{T_y^k}{k} = \frac{1}{k}\sum_{j=1}^{k} T_y \overset{a.s.}{\to} E_y T_y$. Can check $T_y^{N_n(y)}$, at time n you may be somewhere else but a short time ago you were at y, so $T_y^{N_n(y)} \approx n$ for large n. So we conclude $\frac{T_y^k}{k} \approx \frac{n}{N_n(y)}$ for n,k large, seen by substituting this random time $N_n(y)$ for k.
- Sometimes we are looking at a fully irreducible MC, but we can also look at separated irreducible sets as their own MCs and apply our theorems there

Classes of Markov Chains

Detailed Balance Condition

- Reversible chains satisfy the DBC
- Recall a pmf π on S is a stationary distribution for a MC with transition matrix P if $(1)\ \pi = \pi p$, $\pi(y) = \sum_{x\ inS} \pi(x)p(x,y),\ \forall y \in S$. A pmf π is said to satisfy a DBC if $(2)\ \pi(x)p(x,y) = \pi(y)p(y,x),\ \forall x, y \in S$.
- If π is a mass distribution on the state space, p tells you has this mass moves to other states. Given some amount of sand, the amount moved from x to y is the same amount moved from y to x. For each pair, we exchange the same amount.
- Claim: $(2) \implies (1)$. If (2) holds, for any y, $\sum_{x\in S} \pi(x)p(x,y)$ (the amount of sand coming into y) = $\sum_{x\in S} \pi(y)p(y,x) = \pi(y)\sum_{x\in S} p(y,x)$. By the fact that p is a stochastic matrix $\pi(y)\sum_{x\in S} p(y,x) = \pi(y)$. The total equation here says the amount of sand coming into y is the amount that was there before.
- Example (See Durrett Page 30, Ex 1.27): Ehrenfest Chain (Gas through permeable membrane)
    - With 3 balls, 2 urns $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$. Put $\pi(0) = c$ (TBD). Check DBC $\pi(0)p(0,1) = \pi(1)(1,0)$: we get $c \times 1 = \pi(1)\frac{1}{3} \implies \pi(1) = 3c$.

- System: $\pi(0)p(0,1) = \pi(1)(1,0),$.
  $\pi(1)p(1,2) = \pi(2)(2,1),$
  $\pi(2)p(2,3) = \pi(3)(3,2)$
- We get $\pi(2) = 3c,\ \pi(3) = c$. Take $c = \frac{1}{8}, \pi = (\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8})$
- Note the DBC is a condition for every pair of states for all x,y in S. We only checked it for 3 pairs, but for this chain, for any x,y with $|x-y| > 1$, the DBC is 0=0 which trivially holds. So it just remains to check for (x,y) in $\{(0,1)(1,2),(2,3)\}$.
- For N balls, could just guess and check $\pi(x) = 2^{-N}\binom{N}{x}$ (binomial distribution). Seems binomial from specific case, so then just check the identity for $0 \leq x \leq N-1,\ \pi(x)p(x,x+1) = \pi(x+1)p(x+1,x)$

- **Birth and Death Chains**: a general class of examples with DBCs
  - S is ordered, say as a $\{0,1,2,...\}$, transitions only between neighbors or stay put.
  - Examples: simple random walk on S, Ehrenfest chain, Gambler's ruin
- Simple Random Walks on an **Undirected Graph**
  - G = (V,E), S = V. For x,y in V, $p(x,y) = \begin{cases} \frac{1}{deg(x)} & if\ y \sim x \\ 0 & otherwise \end{cases}$, where y ~ x indicates neighboring vertices. Take $p(1,2) = 1/2, p(3,4) = 1/3, p(4,3) = 1, p(1,4) = 0$
  - Note 3 has the highest degree of 3, so would expect chain to spend the most time at 3 and least at 4. Guess $\pi(x) = c \times deg(x)$ and check if DBC hold.
  - For x,y in V, if x is not connected to y, $\pi(x)p(x,y) = \pi(y)p(y,x) \implies 0 = 0$
  - If x ~ y, then $\pi(x)p(x,y) = c \times deg(x)\frac{1}{deg(x)} = c$. Since this is for arbitrary y and x, switching them also returns c, satisfying the DBC.
  - Now take c st π is a pmf. Then $c = \frac{1}{\sum_{x \in V} deg(x)} = \frac{1}{2|E|}$ since we count every edge twice by summing over degrees.
  - See Durrett Examples 1.33, 1.34 (Knight's Random Walk)
  - Taking the example, $\pi = \frac{1}{8}(2,2,3,1)$. If G were regular, the degree of x equals d (all nodes have same degree), then π is uniform on V.

Doubly Stochastic MCs

- Transition matrix p is doubly stochastic if the columns and rows sum to 1: p, p^T both stochastic.
- For example, SRW on a regular graph. In this case P is actually a symmetric matrix: $p = \frac{1}{d}A$ where A is the adjacency matrix with entries $A(x,y) = \begin{cases} 1 & x \sim y \\ 0 & else \end{cases}$. Since it is an undirected graph, this makes it symmetric - any symmetric matrix is going to be doubly stochastic.
- **D Theorem 1.14**: If P is doubly stochastic N x N, then $\pi(x) = \frac{1}{N}$ (uniformly distributed) is a stationary distribution
  - Proof: Check $\pi p = \pi$. Then $(\pi p)(y) = \sum_{x \in S} \pi(x)p(x,y) = \frac{1}{N}\sum_{x \notin S} p(x,y)\frac{1}{N} = \pi(y)$

Reversibility

- Let X_n be a MC with stationary distribution π and suppose $X_0 \sim \pi$ (so $X_n \sim \pi\ \forall n$). Fix n and put $Y_m = X_{n-m},\ 0 \leq m \leq n$. Time reversed process - just looking at this chain backwards.
- **Theorem 1.15**: Let's assume $\pi(x) > 0\ \forall x \in S$. Then Y_m is a MC with $Y_0 \sim \pi$ and transition probabilities $\hat{p}(x,y) = \frac{\pi(y)p(y,x)}{\pi(x)}$. To go the other direction we are adjusting by the ratio of stationary probabilities.
  - Proof: First, we do not even know it is a MC. NTS $P(Y_{m+1} = y_{m+1}|Y_m = y_m, \ldots, Y_0 = y_0) = \frac{\pi(y_{m+1})\hat{p}(y_{m+1},y_m)}{\pi(y_m)}$ - this shows the markov property since it has forgetten its history until the last step.

$P(Y_{m+1} = y_{m+1} | Y_m = y_m, \ldots, Y_0 = y_0) = \frac{P(X_{n-m+1}=y_{m+1},\ldots,X_n=y_0)}{P(x_{n-m}=y_m,\ldots,X_n=y_0)}$

$= \frac{\pi(y_{m+1}p(y_{m+1},y_m)Pr(X_{n-m+1}=y_{m-1},\ldots,X_n=y_0|X_{n-m}=y_m)}{\pi(y_m)Pr(X_{n-m+1}=y_{m-1},\ldots,X_n=y_0|X_{n-m}=y_m)}$

$\text{Cancelling: } = \frac{\pi(y_{m+1})\hat{p}(y_{m+1},y_m)}{\pi(y_m)}$

- If p, π satisfy DBCs then $\hat{p}(x,y) = p(x,y)$ and so $(Y_m)_{m=0}^n \overset{d}{=} (X_m)_{m=0}^n$ for any n. Say MC is reversible

Metropolis-Hastings Algorithm

- Goal: compute (approximate) $E(f(Y))$ for $Y \sim \pi$ some complicated pmf on set S. We may not have a nice formula st we can compute $\sum_{x\in S} f(x)\pi(x)$.
- Idea: design an irreducible MC on S with π as its stationary distribution. By the ergodic theorem, if we run the MC from time 1 to n $\frac{1}{n}\sum_{m=1}^n f(X_m) \overset{a.s.}{\to} E(f(Y))$
- Sampling algorithm to get at a difficult distribution using the ergodic theorem
- Start with a proposed jump distribution $Q(x,y)$ - this won't be our eventual transition matrix, but we design this to either transition with **jump distribution Q** or do nothing. Accept a transition with **acceptance probability** $R(x,y) = min\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}, 1\}$ (threshold at one to ensure valid probability). Then transition matrix p(x,y), for $x \neq y,\ p(x,y) = Q(x,y)R(x,y)$. We accept a move according to Q with probability R(x,y) else stay put.
- So $p(x,x) = R(x,x)Q(x,x) + \sum_{y\neq x}(1-R(x,y))Q(x,y) = Q(x,x) + \sum_{y\neq x}(1-R(x,y))Q(x,y)$ or simply $p(x,x) = 1 - \sum_{y\neq x} p(x,y)$ to make out probabilies sum to 1. First term - either we accept Q and move, or right term we reject Q and stay put.
- Claim: p, π satisfy DBCs - in particular π is the stationary distribution for p.
    - Proof: Pick arbitrary pair of states, $x, y \in S$. WLOG we may assume $\pi(y)Q(y,x) > \pi(x)Q(x,y)$. We then have $\pi(x)p(x,y) = \pi(x)Q(x,y)R(x,y) = \pi(x)Q(x,y)$. The reverse $\pi(y)p(y,x) = \pi(y)Q(y,x)R(y,x) = \pi(y)Q(y,x)\frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)} = \pi(x)Q(x,y)$ The whole art of the algorithm is in designing Q.
- Example: M-H for Geometric distribution. S = [0,1,2,...], $\pi(x) = \theta^x(1-\theta)$, for some $\theta \in (0,1)$. Since the distribution is easy, we wouldn't use M-H so this is just illustrative. View π as a pmf on $\mathbb{Z}$ by putting $\pi(x) = 0, \forall x < 0$
    - Take Q to be the transition matrix SRW on $\mathbb{Z}$ (can also take it to be the reflecting random walk). $Q(x,y) = \begin{cases} 1/2 & \text{for } (x-y) = 1 \\ 0 & else \end{cases}$, and acceptance probability $R(x,y) = min\{1, \frac{\pi(y)}{\pi(x)}\}$ since Q is symmetric so we get cancellation in the ratio.
    - For x > 0 transition to the left, $p(x,x-1) = \frac{1}{2}R(x,x-1) = \frac{1}{2}$ since $\pi(x)$ decreases with x. Transition to the right $p(x,x+1) = \frac{1}{2}\theta$
    - $p(x,x) = \frac{1-\theta}{2}$. For x = 0, $\pi(-1) = 0 \implies p(0,-1) = 0,\ p(0,1) = \theta/2,\ p(0,0) = 1 - \frac{\theta}{2}$
    - Is this irreducible? Yes, see transitions to neighbors are all positive. $p(x,y) > 0,\ \forall x, y$ neighbors ($|x-y| = 1$), so we can get from x to y in |x-y| steps with positive probability, ie. all states communicate with all others. By claim (accept / reject condition satisfies DBC) and ergodic theorem: $\frac{1}{n}\sum_{m=1}^n f(X_m) \overset{a.s.}{\to} E(f(Y))$ for $Y \sim geom(\theta),\ X_n$ MC with transition matrix p.
    - The point is to get to a p we can compute and is irreducible, something we can work with more easily.
- Physics Spin Aligment: More complicated π. Say V is a large finite set, say atoms in a crystal. S is set of all configurations of +1,-1 on V: $S = \{+1,-1\}^V = \{\sigma : V \to \{+1,-1\}\}$, size 2^V - enormous state space. We have a Hamiltonian function $H : S \to \mathbb{R}$, with $H(\sigma)$ giving the energy of this configuration σ. We have pmf $\pi(\sigma) = \frac{1}{Z}exp(-\beta H(\sigma)),\ \beta > 0$ parameter, inverse temperature. Z = normalizing constant, given by $\sum_{\sigma\in S} e^{-\beta H(\sigma)}$, $Z(\beta)$, partition function.
    - Gibbs measure: Have an inverse temperature parameter $\beta > 0$. Simplified model of atom spin up or down, tend to align with their neighbors. A configuration will be more likely if it has lower energy. For β very small, close to

the uniform distribution for π, while large β makes it important to minimize the energy.

  - Generally for Gibbs measures on Graphs: Graph G=(V,E), S = $\{+1,-1\}^V$ = configurations of spins (± 1) on sites V, with $|S| = 2^{|V|}$. Hamiltonian $H : S \rightarrow]R$ is the energy at configuration σ. Given this data, Gibbs measure is $\pi_\beta(\sigma) = \frac{1}{Z(\beta)} e^{-\beta H(\sigma)}$. Note we cannot compute normalizing constant $Z(\beta)$ - it is simply over too large a space..
  - To model a 2D ferromagnet, have a grid with $\Lambda = \{-L, -L+1, \ldots, L\}^2$ defining its sites. Each site gets a label +1 or -1. Our state space is the set of all possible labelings. $H(\sigma) = -\sum_{x\ simy} \sigma(x)\sigma(y)$, since $\sigma(x)\sigma(y) = +1$ for $\sigma(x) = \sigma(y)$ else -1.
  - Problem: $Z(\beta) = \sum_{\sigma \in \{+1,-1\}^V} e^{\beta \sum_{x\sim y} \sigma(x)\sigma(y)}$ is a gigantic sum - cannot compute
  - M-H: For $\sigma \in S$ and $x \in V$ write σ^x for the configuration you get for taking sigma and flipping the value at site x to the opposite sign, so they only disagree on one atom. $\sigma^x(y) = \begin{cases} \sigma(y) & x \neq y \\ -\sigma(y) & x = y \end{cases}$. Draw $X \in V$ uniformly at random and flip spin at X. Proposed jump distribution $Q(\sigma, \sigma^y) = \frac{1}{(2L+1)^2},\ y \in V$.

- Ising Model of a Ferromagnet
  - Problem: explain why nearnest neighbor preferences to align with each other explain how globally we get alignment across most atoms. Aim is to explain long range order in spin alignments emerging from short-range interactions.
  - High energy when spins disagree among many neighbors. $H(\sigma) = -\sum_{i,j \in V} \sigma(i)\sigma(j)$ with $i \sim j$ (${i,j} \in E). If they agree this number is quite negative, if disagree then we sum over a positive quantity.
  - Sample from π_β with M-H: we need to propose a jump distribution. Pick $i \in V$ uniformly at random and flip the bit $\sigma(i)$ For $\sigma, \tau \in S$, we have $Q(\sigma,\tau) = \begin{cases} \frac{1}{|V|} & \text{for } \tau = \sigma^i \\ 0 & else \end{cases}$ $\text{some } i \in V$ where $\sigma^i(j) = \begin{cases} \sigma(j) & i \neq j \\ -sigma(i) & i = j \end{cases}$ is probability of flipping the bit.
  - So $p(\sigma,\tau) = Q(\sigma,\tau)\ min\ \{\frac{\pi_\beta(\tau)}{\pi_\beta(\sigma)}, 1\},\ \sigma \neq \tau$ - the $Z(\beta)$ cancels! Only nonzero for $\sigma = \tau or$\tau=\sigma^i$ for some i in V.
  - We only changed the spin at i, so only pairs at i are involved. $\frac{\pi_\beta(\sigma^i)}{\pi_\beta(\sigma)} = e^{\beta(\sum_{j\sim i} \sigma^i(i)\sigma^i(j) - \sigma(i)\sigma(j))} = e^{-2\beta \sum_{j\sim i} \sigma(i)\sigma(j)}$ from $\sigma^i(i) = -\sigma(i),\ \sigma^i(j) = \sigma(j)$
  - Specfically on G = (V,E) 2D lattice, $V = \{-L, -L+1, \ldots, L\}^2$, E = nearest neighbor edges = points differing by 1 in vertical or horizontal (no diagonals). #V = $(2L+1)^2$. On the boundary, you have three neighbors rather than 4, so instead we view this as being on the infinite lattice and set boundary conditions that are all (+) spins outside of our lattice range.
    - In this case, ratio is $e^{2\beta(2k-4)}$ where $\sigma(i)$ agrees with $\sigma(j)$ for k of the neighbors $j \sim i$ and $0 \leq k \leq 4$ (since 2k - 4 is then between -4 and 4). Our sum $e^{-2\beta \sum_{j\sim i} \sigma(i)\sigma(j)}$ will sum +1 or -1 four times.
    - We accept this move with probability 1 when $\frac{\pi_\beta(\sigma^i)}{\pi_\beta(\sigma)} \geq 1$. Then $\frac{\pi_\beta(\sigma^i)}{\pi_\beta(\sigma)} \begin{cases} > 1 & for\ k = 0,1 \\ = 1 & for\ k = 2 \\ < 1 & for\ k = 3,4 \end{cases}$ and we accept for the first two cases. If you agree with less than or equal to half of your neighbors, we flip your bit.
    - If k = 4, say accept with probability $e^{-8\beta}$. If beta very large, we are not going to accept with very high probability. We really want a low energy orientation so we have very little tolerance for flipping bits.
    - In our 2D lattice, the lowest energy configuration is all + (since boundary condition is (+)). This model has a phase transition, β critical = $\frac{1}{T_c}$ critical temperature (Curie temp), where all + with high probability if you are below the Curie temperature (so high beta $\beta > \beta_c$) and looks random for $\beta < \beta_c$. Heating this lattice will cause it to go from high alignment to random assignment. Below the Curie temperature, spins align and we have a ferromagnet. Note in 1D there is not a phase transition, but in 2D we have a phase transition.

Branching Processes

- $S = \{0, 1, 2, 3, \dots\}$ countably infinite. $X_n \in S$ is the number of individuals in generation n.
- Offspring distribution: each individual independently gives birth to a number of children with pmf f on S.
  - For example $f(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & x > 2 \end{cases}$
- At generation n, each of the X_n individuals independently gives birth to random number of children with distribution f. Say $Y_{n,m}$ = # of children of person m in generation n where $1 \leq m \leq X_n$.
- Formally, let $\{Y_{n,m}\}_{n=0,m=1}^{\infty}$ array, iid with distribution f. Then $X_{n+1} = \sum_{m=1}^{X_n} Y_{n,m}$ - sum over individuals in the nth generation of the number of children they have. Sum holds for $X_n \geq 1$ else it is 0 for $X_n = 0$. The $Y_{n,m}, 1 \leq m \leq X_n$ are independent of $X_0, \dots, X_{n+1}$.

Fig.1.

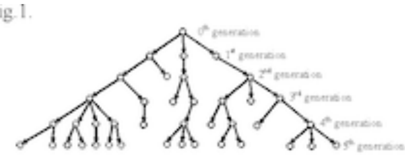


Figure 1. Multiplication tree of the branching process.

Extinction

- Given $X_0 = 1$, what is the probability of extinction, $P_1(T_0 < \infty)$? Probability of starting from 1, time we reach 0 is finite, where 0 is an absorbing state. Can see this since we are summing from 1 to 0, which by definition is 0.
- Let the mean number of children be $\mu = E(Y) = \sum_{k=0}^{\infty} k f(k)$ with $Y \sim f$. This is the mean of the offspring distribution.
- Let $\phi(r) = \sum_{k=0}^{\infty} r^k f(k)$ be the generating function of the distribution f. Note: r = 1, we get $\phi(1) = \sum_{k=0}^{\infty} f(k) = 1$, $\phi(0) = f(0)$, the probability that any person has 0 children.
- The expected population size. For $n \geq 0$, $E(X_{n+1}|X_n) = E(\sum_{m=1}^{X_n} Y_{n,m}|X_n) = \sum_{m=1}^{X_n} E(Y_{n,m}|X_n) = \mu X_n$. So we see after iterating $E(X_n) = \mu^n E(X_0)$, with the mean determining whether we go extinct in expectation.
- Outside of expectation (taking $X_0 = 1$ throughout), we can use Markov's inequality:
  - If $\mu < 1$: $P(X_n \geq 1) \leq \frac{E(X_n)}{1} = \mu^n E(X_0) \overset{n\to\infty}{\to} 0$ if $\mu < 1$. This implies $P(T_0 = \infty) = P(X_n \geq 1 \; \forall n) \leq P(X_n \geq 1) \to 0$. So in the case $\mu < 1$, the probability of extinction is 1, $P(T_0 < \infty) = 1$.
  - If $\mu > 1$, let's do a first step analysis. Let $\rho = \rho_{10} = P_1(T_0 < \infty)$, then want to find ρ. At time 0 $X_0 = 1$, they give birth to random number of children, $Y_{0,1}$ children = X_1. From the 0 node, what is the probability that the entire progeny eventually disappears. This mean each subtree must go to zero, and each sibling subtree is independent of each other. If there are k children in the 1st generation, to have process reach state 0 (extinction), need each of these k independent lineages to die out. Starting from a node on step 1 looks the same as starting from the 0 node - each of these events also has probability ρ.
    - So we have probability ρ^k that they all die out.
    - We get $\rho = \sum_{k=0}^{\infty} \rho^k P(Y_{0,1} = k) = \sum_{k=0}^{\infty} \rho^k f(k) = \phi(\rho)$. So ρ is a fixed point of function ϕ - a number ρ st $\phi(\rho) = \rho$. We saw 1 is a fixed point, but perhaps there are others.
    - In particular if f(0) = 0, $\rho = 0$ by theorem below. That is somewhat obvious since, f(0) = 0 means everyone has at least 1 child, so clearly the branching process does not die out.
    - We have properties $\phi(1) = 1, \; \phi(0) = f(0)$. The derivative of ϕ, $\phi'(r) = \sum_{k=1}^{\infty} k r^{k-1} f(k) \geq 0 \implies \phi$ is increasing. Additionally, $\phi'(1) = \sum_{k=1}^{\infty} k f(k) = \mu$. The second derivative: $\phi'(r) = \sum_{k=2}^{\infty} k(k-1)\rho^{k-2} f(k) \geq 0 \implies \phi$ is convex on [0,1]. Can see all derivatives are positive and continuous for $0 < r < 1$ (so no jump discontinuities).

        - Critical case $\mu = 1$, we get $\rho = 1$, it will extinguish eventually. We get the same result as a $\mu < 1$ and the population dies out with probability 1. In this case we cannot have $\phi(r) = r$ for $r < 1$ as then ϕ would have some derivative discontinuous (since once it touches the extinction line, it must stay on that line and all higher derivatives must be 0). <- This statement is not totally true, more explicitly we can say: We cannot have $\phi(r) = r$ for $r < 1$ as then ϕ would not be analytic (but outside the scope of the class to show).

- Theorem: ρ is the smallest solution of $\phi(r) = r$ in [0,1].
    - Extinction probability $\rho = P_1(T_0 < \infty)$ os the smallest solution of $\phi(r) = r$ with $0 \leq r \leq 1$
- Corrolary: If f(0) > 0, and $\mu = 1$, then $\rho = 1$. If $\mu > 1$, $\rho \in (0, 1)$.
    - Proof of theorem: let $\rho_n = P_1(X_n = 0)$ (0 absorbing state), so we are saying by time n it hits zero, though also could have happened before n. Absorbing at 0 implies $X_n = 0 \implies X_{n+1} = 0$, so sequence ρ_n is monotone non-decreasing.
        - $\lim_{n\to\infty} \rho_n = \rho$. We can establish the recursion for ρ_n. For descendents of Eve to die out within n generations, need descendents of each of her children to die out within n - 1 generations. In math, by independence of descendents of each child, the probability that they die out $\rho_n = \sum_{k=0}^{\infty} f(k)\rho_{n-1}^k = \phi(\rho_{n-1})$. K independent events in which a whole tree needs to die out, probability of each tree if ρ_{n-1} and since independent we multiply over k times. We get a recursion using our generating function.
        - $\rho_n = \phi(\rho_{n-1})$. Starting at ρ_1, we get $\rho_1 = P_1(X_1 = 0) = f(0)$, the probability that Eve has no children. Then $\rho_2 = \phi(\rho_1),\ \rho_3 = \phi(\rho_2) = \phi(\phi(\rho_1))$. Call $\rho_* = min\{r \in [0, 1] : \phi(1) = r\}$. Note: for $0 \leq r \leq \rho_*,\ r < \phi(r) < \rho_*$. Let's assume $f(0) > 0$, otherwise we have proved the theorem in this case.
        - Sequence ρ_n increasing and bounded by ρ_*. See chart of bouncing ρ toward ρ_*. It's limit $\rho \leq \rho_*$. Remains to show that it is equal to ρ_*. Taking $n \to \infty$ in $\rho_n = \phi(\rho_{n-1})$, $\rho = \lim_{n\to\infty} \rho_n = \lim_{n\to\infty} \phi(\rho_{n-1}) = \phi(\lim_{n\to\infty} \rho_{n-1})$ (can take lim internal to phi since continuous) $= \phi(\rho)$. So ρ is a fixed point of ϕ
        - ρ_* was defined to be the smallest fixed point, so $\rho \geq \rho_*$. Since we also showed $\rho \leq \rho_*$, we conclude $\rho = \rho_*$.
- We had easy arguments for cases $\mu < 1$ (using Markov's inequality) and f(0) = 0 (all people have at least 1 child, so no extinction possibility).
- Example: Suppose $f(k) = \begin{cases} 1/4 & k = 0 \\ 1/4 & k = 1 \\ 1/2 & k = 2 \\ 0 & else \end{cases}$.
    - $\mu = \sum_{k=0}^{\infty} kf(k) = 1/4 + 2(1/2) = \frac{5}{4} > 1$. Expectation of $X_n = E(X_n) = \mu^n \frac{E(X_0)}{1} = \left(\frac{5}{4}\right)^n$, which is exponentially large.
    - Find probability of extinction. Using this theorem, we just need to find the fixed point of our polynomial method $\phi(r)$: $\phi(r) = \sum_{k=0}^{\infty} r^k f(k) = 1/4 + r(1/4) + r^2(1/2)$
    - Solve $\phi(r) = r$: $0 = \frac{1}{4}(1 - 3r + 2r^2)$, then $r = \frac{3\pm\sqrt{9-8}}{4} = \frac{3\pm1}{4} = \{1/2, 1\}$
    - Recall 1 is always a fixed point so can help factor higher degree polynomial cases. Here we get $\rho = P_1(T_0 < \infty) = 1/2$, since this is the smaller root of $\phi(r)$ We have a 50/50 chance of eventual extinction.

Reflecting Random Walk

- Let S = $\{0, 1, 2, 3, \ldots\}$, we move to the left with probability 1-p and right with probability p. At 0 we transition to 0 with probability 1-p. This is a birth and death chain and is clearly irreducible. We can walk straight from any state to any other state.
- Cases: p < 1/2: DBC: $\pi(x) = \left(\frac{p}{1-p}\right)^x \pi(0)$ is solution to $\pi P = \pi$. Normalizable to get stationary distribution if p < 1/2. Then we get $\pi(0) = \frac{1-2p}{1-p}$
    - Expected return time to 0 - $E_0 T_0 = \frac{1}{\pi_0} = \frac{1-p}{1-2p}$ when P < 1/2. All states are recurrent.

- Case: p > 1/2: Claim 0 is transient. From Gambler's ruin 1st step analysis we have performed, for 0 < x < N, $P_x(V_N < V_0) = \frac{\left(\frac{1-p}{p}\right)^x - 1}{\left(\frac{1-p}{p}\right)^N - 1}$. Plugging in $P_x(T_0 < \infty) = P_x(V_0 < \infty) = \lim_{n\to\infty} P_x(V_0 < V_N) = \left(\frac{1-p}{p}\right)^x$. Then $1 - \rho_{00} = P_0(T_0 = \infty) \geq P(0,1)P_1(T_0 = \infty) = p\frac{1-p}{p} = 1 - p > 0$. Positive probability of never coming back, so it is transient.
    - This is a common proof technique in infinite spaces - imagine a large finite space and then take to infinity. The last step is essentially a Gambler's ruin calculation. See Durrett for this example as well.
    - Similarly, x is transientl $\forall x \geq 0$. See this from : $P_x(T_x = \infty) \geq P(x, x+1)P_{x+1}(T_X = \infty) = P(x, x+1)P_1(T_X = \infty) = 1 - p > 0$. If you are at 3 wondering if you will ever get back to 2, this is the same question as if you are at 1 wondering if you will get back to 0.
- Critical case: p = 1/2. Can tie back to our branching process where we had critical case $\mu = 1$, again this is the more interesting case. For some N and $1 \leq x \leq N-1$, $P_x(V_N < V_0) = \frac{x}{N}$ (Similar to SRW on a clock). Taking $N \to \infty$ for any fixed $x \geq 1$, the probability from x that we run away to infinity before coming back to 0, $P_x(V_0 < V_N) = 1 - x\frac{x}{N} \to 1$, so $P_x(V_0 < \infty) = 1$, tend towards probability 1 of hitting 0 at some finite time.
    - Looks recurrent, so let's try to show using ρ_{00} and first step analysis: $\rho_{00} = P_0(T_0 < \infty) = p(0,0)(1) + p(0,1)P_1(T_0 < \infty) = 1 - p + (p)P_1(T_0 < \infty) = 1 - p + p(1) = 1$. Therefore 0 is recurrent. Similarly, can show any state x is recurrent. So in the critical case all states are recurrent.
    - But this is not our typical recurrence - we are null recurrent, not positive recurrent. The expected return time to a state is infinite.
    - First step analysis shows: $E_1 T_{x_{\{0<N\}}} = N - 1 \overset{N\to\infty}{\to} \infty$. This means $E_1 T_0 = \infty$, or looking at return time $E_0 T_0 = 1 + p(0,0)(1) + p(0,1)E_1 T_0 = 1 + 1/2 + (1/2)\infty = \infty$.
- Definition: A state x is **positive recurrent** if the expected return time if finite $E_x T_x < \infty$, **null recurrent** if recurrent but $E_x T_x = \infty$.
    - For reflecting random walk, state 0 is positive recurrent if p < 1/2, null recurrent if p = 1/2, transient if p > 1/2. While we performed example on state 0, this holds for any state x in the RRW.

Other Topics in Discrete Time

- Mixing time of MC's. See Bayer - Diaconis, it takes 7 shuffles to randomize a deck of cards. State space is all 52! orderings of the deck. After about 7 shuffles, the distribution is pretty close to the stationary distribution of the riffle shuffle transition matrix, which is the uniform distribution.
- MCMC Markov Chain Monte Carlo. Extensions in sampling of MH.

Poisson Processes

- We will model random times as exponential variables. For $T \sim Exp(x) : F_T(t) = P(t \leq T) = 1 - e^{-\lambda t}$ and $f_T(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$. By integration by parts: $ET = \int_0^\infty \lambda t e^{-\lambda t} dt = 1/\lambda$ and $E(T^2) = \int_0^\infty \lambda t^2 e^{-\lambda t} dt = 2/\lambda$, so $Var(T) = 2/\lambda - (1/\lambda)^2 = 1/\lambda^2 = (ET)^2$
- Memoryless property: $P(T > s + t | T > t) = \frac{P(T > t+s, T > t)}{P(T > t)} = \frac{P(T > t+s)}{P(T > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(T > s)$. Given a waiting period, the probability of waiting some time more is the same as if we hadn't waited at all.
- Theorem (**Exponential Races**): Let $T_1, \ldots, T_n$ independent $T_i \sim Exp(\lambda_i)$. Set $S = min\{T_1, \ldots, T_n\}$ (first arrival time, say arrival of the first bus). Let $I = argmin_{1 \leq i \leq n}\ (T_i)$ be the index of the min arrival time. Then 1) $S \sim exp(\lambda_1 + \ldots + \lambda_n)$ and 2) $P(I = i) = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_n}$ and 3) S and I are independent.
    - Proof: Starting with 1), probability that min is larger than t, $P(S > t) = P(T_1 > t, \ldots, T_n > t)$ (want to use a condition that can bring in all of the comparisons then use independence, would use less than if S were a max). Then $P(T_1 > t, \ldots, T_n > t) = e^{-\lambda_1 t} \ldots e^{-\lambda_n t} \implies F_S(t) = P(S \leq t) = 1 - e^{-(\lambda_1 + \ldots + \lambda_n)t}$

- Proving 2), First case n =2: $P(T_1 < T_2 = \int_0^\infty f_{T_1}(t)P(T_2 > t)dt = \int_0^\infty \lambda_1 e^{-\lambda_1 t}e^{-\lambda_2 t}dt = \frac{\lambda_1}{\lambda_1+\lambda_2}$. For the general case n we write $S_i min\{T_j; j \neq i\}$, $P(I=i) = P(T_i < S_i) = \frac{\lambda_i}{\lambda_i + \sum_{j\neq i}\lambda_i} = \frac{\lambda_i}{\lambda_1+\ldots+\lambda_n}$
- Proving 3) $S \perp I$: We will show conditional density, given the min index happens at i, $f_{S|I}(t|i) = f_S(t)$ (conditioning on I has no effect, so independent). Indeed, taking joint over marginal $f_{S|I}(t|i) = \frac{\lambda_i e^{-\lambda_i t}P(T_j > t\ \forall j\neq i)}{P(I=i)} = \frac{\lambda_i e^{-\lambda_i t}\prod_{j\neq i} e^{-\lambda_j t}}{\lambda_i/(\lambda_1+\ldots+\lambda_n)}$ by 2), then $= (\lambda_1 + \ldots + \lambda_n)e^{-(\lambda_1+\ldots+\lambda_n)t} = f_S(t)$ by (1).

- Gamma Distribution: Let $T_n = \tau_1, \ldots, \tau_n \overset{iid}{\sim} exp(\lambda)$, then $\tau_1 + \ldots + \tau_n \sim Gamma(n, \lambda)$. (Taus will be interarrival times and T is an arrival time). ie, $f_{T_n}(t) = \begin{cases} \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!} & t \geq 0 \\ 0 & t < 0 \end{cases}$.
- Poisson Distribution: $X \sim Pois(\lambda)$ with pmf $p_X(n) = e^{-\lambda}\frac{\lambda^n}{n!},\ n \geq 0$. Proposition: If $X_1, \ldots, X_n \overset{\perp}{\sim} Pois(\lambda_i)$, then $Y = X_1 + \ldots X_n \sim Pois(\lambda_1 + \ldots + \lambda_n)$.
  - Proof: Take case n = 2. $P(X_1 + X_2 = n) = \sum_{m=0}^n P(X_1 = m)P(X_2 = n-m) = \sum_{m=0}^n e^{-\lambda_1}\frac{\lambda_1^m}{m!}e^{-\lambda_2}\frac{\lambda_2^{n-m}}{(n-m)!}$ (For sum of RV, the distribution is the convolution of the two RVs), then $= \frac{e^{-\lambda_2-\lambda_1}}{n!}\sum_{m=0}^n \lambda_1^m\lambda_2^{n-m} = \frac{e^{-\lambda_2-\lambda_1}}{n!}(\lambda_1\lambda_2)^n$
  - Case of general $n \geq 2$ follows by induction: $X_1 + \ldots + X_n = (X_1 + \ldots + X_{n-1}) + X_n \sim Pois(\lambda_1 + \ldots + \lambda_{n-1}) \perp Pois(\lambda_n) \sim Pois(\lambda_1 + \ldots + \lambda_n)$
- **Definition**: Let $\lambda > 0$ A **Poisson Process** of rate λ is a continuous time stochastic process $(N(t))_{t\geq 0}$ with state space $\mathbb{Z}_{\geq 0}$ satisfying:
  1. N(0) = 0 almost surely
  2. $N(t+s) - N(s) \sim Pois(\lambda t)\ \forall s \geq 0, t > 0$
  3. (Independent increments) For any $0 \leq t_0 < t_1 < \ldots < t_n$, then $N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$ are independent.
- Remark: We will show that processes that meet these properties exist. Additionally, note that N(t) is a non-decreasing function of t - it is like a random staircase never going down.
- **Theorem**: Let infinite sequence of $\tau_1, \tau_2, \ldots \overset{iid}{\sim} exp(\lambda)$, set $T_0 = 0$, and for each $n \geq 1$, put $T_n = \tau_1 + \ldots + \tau_n$. For $t \geq 0$, let $N(t) = max\{n : T_n \leq t\}$. Then $(N(t))_{t\geq 0}$ is a rate λ Poisson process ($PP(\lambda)$).
- Lemma: $\forall t > 0,\ N(t) \sim Pois(\lambda t)$.
  - Proof: We have T_n before t and T_{n+1} after t and we have $\tau_{n+1} =$ the interval from T_n to T_{n+1}. Say for any $n \geq 0,\ P(N(t) = n) = P(T_n \leq t < t_{n+1}) = P(T_n \leq t < T_n + \tau_{n+1})$. Then conditioning on the event that $T_n = s$ for some s (removing randomness) $= \int_0^t f_{T_n}(s)P(\tau_{n+1} > t - s)ds = \int_0^t \lambda e^{-\lambda s}\frac{(\lambda s)^{n-1}}{(n-1)!}e^{-\lambda(t-s)}ds = \frac{\lambda^n}{(n-1)!}e^{-\lambda t}\int_0^t s^{n-1}ds = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$.
- Lemma: For any $s, t > 0$, $N(s+t) - N(s) \overset{d}{=} N(t)$ and $N(s+t) - N(s) \perp (N(\delta))_{0\leq\delta\leq s}$
  - Proof: Want to see why process of arrivals for s to s+t is the same as if we had started from 0. Say we have 4 arrivals in this time increment. Between our arrivals, we have our interarrival times τ_i. Time s lies within an interarrival time. Starting from 0, we simply added iid exponentials to get our sequence, and it is almost the same starting from s - we just need to examine the distribution of the first waiting time, from s to our first arrival in s to s+t. Let $\tau_1^{(s)}$ be the interval to first arrival from s. Write time of arrivals after s, $T_0^{(s)} = 0,\ T_j^{(s)} = T_{N(s)+j} - s,\ j \geq 1$. Then $T_j^{(s)} = T_j^{(s)} - T_{j-1}^{(s)},\ j \geq 1$. For $j \geq$, $T_j^{(s)} \sim exp(\lambda)$. So we need to show $\tau_1^{(s)} \sim exp(\lambda)$ and independent of $(N(r))_{0\leq r\leq s}$. Say $P(\tau_1^{(s)}|N(s) = k)$, given k arrivals before time s, $= P(\tau_{k+1} \geq s - T_k + u|N(s) = k)$, the probability that the k+1st interval from 0 is at least s + u less the previous arrival time. Note $N(s) = k$ is the event $T_k \leq s \leq T_k + tau_{k+1}$. So we have $= P(\tau_{k+1} \geq s - T_k + u|\tau_{k+1} > s - T_k,\ T_k \leq s)$. Now we can use the memoryless property: T_k and τ_{k+1} are independent and can treat T_k as fixed (since conditioning on it does not affect things). So $= P(\tau_{k+1} \geq s - T_k + u|\tau_{k+1} > s - T_k) = P(\tau_{k+1} > u) = e^{-\lambda u}$, with the last part from memoryless property. The answer in the end is independent of the arrivals before time s. So $\tau_1^{(s)} \sim exp(\lambda)$ and independent of N(s).
  - Comments: N(t) = 0 starting at $T_0 = 0$, then jumps up to next value 1 at T_1. Repeat for each interval. The interval

- from T_0 to T_1 is τ_1, etc.
  - Call T_j the jth arrival time and τ_j is the jth interarrival time. To prove theorem, must show properties (1), (2), (3).
- Proof applied to Theorem:
  - Proof of (1): $P(\tau_1 > 0) = 1$ since exponental RV, so $P(N(0) \geq 1) = P(max\{n : T_n \leq 0\} \geq 1) \leq P(\tau_1 = 0) = 0$. So N(0) = 0 almost surely.
  - For (2) with s=0, we use the first lemma above. $N(t) \sim Pois(\lambda t)$, so we have checked for the interval from 0 to t. From lemma 2, for all s,t $N(s+t) - N(s) \sim Pois(t\lambda)$, $N(s+t) - N(s) \stackrel{d}{==} N(t)$ and $N(s+t) - N(s) \perp (N(r))_{0 \leq r \leq s}$. Lemma 1 and 2 gives us property 2.
  - For (3), independence of increments, $\forall 0 \leq t_0 < \ldots < t_m$, $N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$ are independent. We get this from Lemma 2 by induction.
- Example: Let $(N(t))_{t \geq 0} \sim PP(2)$. Arrival time of 8th person, $ET_8 = E(\tau_1 + \ldots + \tau_8) = 8E(\tau_1) = \frac{8}{2} = 4$. What about $E(T_8|N(1) = 3)$ - we take the same expectation conditioned that by time 1 we have seen 3 arrivals? We have seen more than we expected to see (3 instead of 2), so the 8th arrival will probably happen sooner, so should be smaller than by time 4. Since $(N(1+t) - N(1))_{t \geq 0} \sim PP(\lambda)$, independent of $(N(s))_{0 \leq s \leq 1}$. So $E(T_8|N(1) = 3) = 1 + ET_5 = 1 + 5/2 = 3.5$.

Compound Poisson Processes

- Cars arriving to Dumbarton bridge toll according to a $PP(\lambda)$. Car i pays Y_i dollars according to its type, say 5 types. Model car types as an iid sequence of RVs Y. The amount of money paid up to time t is a new process $S(t) = \sum_{i=1}^{N(t)} Y_i$. Say $Y_i \in [5, 10, 15, 20, 25]$ with some probabilities $p_1, \ldots, p_5$ that sum to 1.
- We call $(S(t))_{t \geq 0}$ a compound poisson process. Relates to our earlier discussion of sums of random variables.
- Proposition: Let infinite sequence $Y_1, . Y_2, \ldots$ iid, $N \in \{1, 2, 3, \ldots\}$ a random variable independent of Y_i. Set $S = \sum_{i=1}^{N} Y_i$. Then:
  1. If $EN,\ E|Y_1| < \infty$, then $ES = EN \times EY_1$
  2. If $E(N^2), (Y_1^2) < \infty$, then $Var(S) = (EN)Var(Y_1) + Var(N)(EY_1)^2$
  3. In particular, if $N \sim Pois(\lambda)$, then (2) becomes $Var(S) = \lambda Var(Y_1) + \lambda(EY_1)^2 = \lambda E(Y_1^2)$

Thinning

- Theorem (Poisson Thinning): Let $N(t) = PP(\lambda)$ and $(Y_i)_{i \geq 1} \stackrel{iid}{\sim} P(Y_i = k) = P_k,\ k \in \mathbb{Z}$ (integer valued RVs) and independent of N(t). Let us only count arrivals of type k, $N_k(t) = \sum_{i=1}^{N(t)} 1(Y_i = k)$. Then the $N_k(t)$ are independent $PP(\lambda P_k)$.
  - Every time a point comes in, we give it a label k according to the distribution P_k. I then get a bunch of different arrivals processes for each k that are subset of the total arrivals process, independent of each other. The joint distribution of each of the types are independent.
  - Proof: Let s,t > 0, $S_k = N_k(s+t) - N_k(s)$. We need to check that the S_k's are independent $Pois(\lambda t P_k)$. The joint pmf: (case Y_1 takes values $\{1, \ldots, m\}$)
  - $P(S_1 = n_1, \ldots, S_m = n_m) = P(N(s+t) - N(s) = n_1 + \ldots + n_m) P(S_1 = n_1, \ldots, S_m = n_m | N(s+t) - N(s) = n_1 + \ldots + n_m)$
  - Then plugging in distributions: $e^{\lambda t} \frac{(\lambda t)^{n_1 + \ldots + n_m}}{(n_1 + \ldots + n_m)!} \times \frac{(n_1 + \ldots + n_m)!}{n_1! + \ldots + n_m!} p_1^{n_1} \ldots p_m^{n_m}$
  - (the second part being a multinomial). This equals $\prod_{k=1}^{m} e^{-\lambda P_k t} \frac{(\lambda P_k t)^{n_k}}{n_k!}$

Superposition

- Theorem: Let $(N_1(t))_{t \geq 0}, \ldots, (N_m(t))_{t \geq 0}$, independent $PP(\lambda_k),\ 1 \leq k \leq m$, then $N(t) = N_1(t) + \ldots + N_m(t)$ is a $PP(\lambda_1 + \ldots + \lambda_k)$.
  - Take a bunch of independent PP's with their own rates, then we can sum them to get another PP.
  - In particular, $N(1) = N_1(1) + \ldots + N_m(1)$ is sum of $Pois(\lambda_k)$ has distribution $Pois(\lambda_1 + \ldots + \lambda_k)$ which we

knew from our work on exponential races. The first arrival time for $N(t)$ is $T_1 = min\{T_1^{(1)}, \ldots, T_1^{(m)}\} \sim exp(\lambda_1 + \ldots + \lambda_k)$ where $T_1^{(k)}$ is first arrival for $N_k(t)$.

- Example: Poisson races. Trucks arrive at the bridge according to $N_{truck}(t) \sim PP(\lambda)$ and cars arrive according to $N_{cars}(t) \sim PP(\mu)$. These are independent processes. Starting at some fixed time, what is the probability that 6 trucks arrive before 4 cars?
    - We can view the trucks process as a thinning of the total process $N(t) = N_{truck}(t) + N_{cars}(t)$. Our theorem tells us $N(t)$ is a $PP(\lambda + \mu)$ and the subprocesses are thinning of the overall $N(t)$. Each vehicle is independently a truck with probability $\frac{\lambda}{\lambda+\mu}$. It's as if everytime a vehicle comes we flip a weighted coin to determine whether it will be a truck or a car.
    - Rewrite our problem as $P(\geq 6 \text{of 1st 9 vehicles are trucks})$ (9 since we don't want the last to be a truck, otherwise 4 cars could arrive before 6 trucks). $Y =$ number of trucks in 1st 9 vehicles is binomial of 9 with parameter $\frac{\lambda}{\lambda+\mu}$, since each vehicle is either truck or car with this probability. Then plugging in the binomial distribution: $\sum_{k=6}^{9} \binom{9}{k} (\frac{\lambda}{\lambda+\mu})^k (\frac{\mu}{\lambda+\mu})^{9-k}$.

Conditioning

- Conditioning on the number of arrive up to some time t
- Theorem: Conditional on $N(t) = n$, vector of arrival times $(T_1, \ldots, T_n)$ (a random, increasing list of numbers between 0 and t), has the same distribution as a different list of ordered RVs $(V_1, \ldots, V_n)$, where $0 \leq V_1 < \ldots < V_n \leq t$ is ordered sample $\{U_1, \ldots, U_n\}$ of iid uniform([0,t]) variables.
    - V is just a relabeling of U's to be in increasing time. The point is the arrival times have the same distribution as a uniform sampling (reordered) when we condition on a Poisson process
    - In particular, (note brackets are unordered sets) $\{T_1, \ldots, T_n\} \stackrel{d}{=} \{U_1, \ldots, U_n\}$ - random unordered sets of n points in [0,t]. If we just look at the process and arrival times, it is the same as if we had sampled iid points and then we can order them in increasing order of time. Fixing any interval, we can throw out points uniformly at random.
    - Proof: Fix $0 \leq t_1 < \ldots < t_n \leq t$, conditional on $N(t) = n$ for some fixed n, the probability density $T_1 = t_1, \ldots, T_n = t_n$ is same as $\tau_1 = t_1, \tau_2 = t_2 - t_1, \ldots, \tau_{n+1} = t - t_n$ (when calculating arrival times often switch to interarrival times since those are iid). Take density $f_{T_1,\ldots,T_n|N(t)}(t_1, \ldots, t_n|n) = \frac{(\lambda e^{-\lambda t_1})\ldots(\lambda e^{-\lambda(t-t_n)})}{e^{-\lambda t}\frac{(\lambda t)^n}{n!}} = \frac{n!}{t^n}$, constant in $t_1, \ldots, t_n$. This means that $(T_1, \ldots, T_n)|N(t)$ is uniform over $\{(t_1, \ldots, t_n) : 0 \leq t_1 < \ldots < t_n \leq t\}$.
    - For 2nd part, density for unordered set $\{T_1, \ldots, T_n\}$ is $\frac{n!}{t^n}\frac{1}{n!} = \frac{1}{t_n}$, the density for given order times probability of observing a particular order. This equals the density for $\{U_1, \ldots, U_n\}$.
- Corollary: Distribution in a subinterval of N(t). For $0 < s < t$ and $n \geq 1$, conditional on $N(t) = n$, $N(s) \sim Bin(n, \frac{s}{t})$.
- Example: Arrivals of customers between 9 am and 10 am, students and professors pass Coupa at rates 2 and 1 per minute. 1/2 of students and 1/3 of professors stop to buy coffee. (a) P(6 professors between 9 and 9:15)? (b) Given 6 profs stopped in 9-9:15, what is the probability that 2 arrived between 9:05 and 9:10? (c) Students get 1 shot of espresso with prob 1/2, 2 shots with prob 1/2. Profs get 1,2,3 shots with probs 1/4, 1/2, 1/4. Find the mean and variance of the total number of shots served between 9 and 9:15.
    - (a) Stopping profs is a PP with rate 1/3 (by thinning), so the number between 9 and 9:15 is $Pois((15)(1/3)) = Pois(5)$. Then P(6 professors between 9 and 9:15) $= e^{-5}\frac{5^6}{6!} \approx 0.1462$
    - (b) Apply the corollary - throwing profs uniformly at random in this time interval. P(2 professors between 9:05 and 9:10) $= \binom{6}{2}(\frac{1}{3})^2(\frac{2}{3})^4 \approx 0.3292$
    - (c) Compound thinned process. Two arrival processes -> thinned -> compounded with shot probabilities. Expected num of shots is E(# served to students) + E(# served to professors). For students, from thinning theorem, let X_i be number of shots for students, then E(# served to students) = E(# students served) x $E(X_i) = 2\frac{1}{2}15\frac{3}{2}$ $(EX_i = \frac{3}{2})$. Then E(# served to professors) $= 1\frac{1}{3}15EY_i$, for Y_i be number of shots for profs,

$EY_i = 2$. Then total E(# served to students) + E(# served to professors). For variance, use the formula from the thinning theorem.

Continuous Time Markov Chains (4.1-4.3)

- Definition: a CTMC is a stochastic process $(X_t)_{t \in [0,\infty]}$ with (countable for this course) state space S satisfying Markov property: $P(X_{t+s} = y | X_s = x, X_{s_0} = x_0, \ldots, X_{s_n} = x_n) = P(X_t = y | X_0 = x)$ for $\forall x, y, x_0, \ldots, x_n \in S, \ \forall 0 \leq s_0 < \ldots < s_n < s, t > 0$
- Time homogeneous built into this definition - a jump from s to s + t the same from a jump from 0 to t.
- Denote $p_t(x, y) = P(X_t = y | X_0 = x)$
- Example 0: Randomly shifted Poisson process. Let $N(t) \sim PP(\lambda), \ X_0 \in \{0, 1, 2, \ldots\}$ an independent RV. Let $X_t = X_0 + N(t)$, so $(X_t)_{t \geq 0}$ is a CTMC.
    - Checking the Markov Property: use the independence of increments of the PP. For $0 \leq s_0 < s_1 < \ldots < s_k < s < s + t$, $P(X_{s+t} = n | X_{s_0} = m_0, \ldots, X_{s_k} = m_k, X_s = m)$. Conditioning on where we were at time zero, $= \sum_{n_0=0}^{\infty} P(N(t+s) - N(s) = n - m | N(s) - N(s_k) = m - m_k, \ldots, N(s_0) - n_0 = m_0, X_0 = n_0)$, $P(X_0 = n_0 | N(s) - N(s_k) = m - m_k, \ldots, N(s_0) - n_0 = m_0)$ rephrasing in terms of gaps to get to independence, have to take expectation over possible starting positions. By independence $= \sum_{n_0=0}^{\infty} P(N(t+s) - N(s) = n - m) P(X_0 = n_0) = P(N(t+s) - N(s) = n - m)$, since $P(N(t+s) - N(s) = n - m) \perp n_0$. Then $= P(N(t) = n - m)$.
    - Now check $P(X_t = n | X_0 = m)$: This equals $P(X_0 + N(t) = n | X_0 = m) = P(N(t) = n - m | X_0 = m) = P(N(t) = n - m)$, with the last step given by independence.
- Proposition (Ex 4.1): Let $(N(t))_{t \geq 0} \sim PP(\lambda)$ an independent discrete time MC $((Y_n)_{n \geq 0})$ with transition matrix $p(x, y)$. Put $X_t = Y_{N(t)}$, making the discrete time MC continuous by have random interarrival times, then $(X_t)_{t \geq 0}$ is a CTMC with $p_t(x, y) = \sum_{n=0}^{\infty} P(N(t) = n) p^n(x, y) = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} p^n(x, y)$
    - Note: with $p(x, y)$ given by state space of non-negative integers $S = \{0, 1, 2, \ldots\}$ and transition probabilities $p(x, x+1) = 1$, we get example 0 above. We are going to go to the next step up and it is just a question of when.

Jump Rates

- CTMCs are typically specified in terms of **jump rates**: $\lim_{t \to 0} \frac{1}{t} p_t(x, y) =: q(x, y), \ x \neq y$. The derivative of the transition probabilities evaluated at 0.
    - When t is very small, $P(N(t) = n)$ should be very small when we compute the derivative at 0. Take X_t as in the proposition, $x \neq y \in S$. Then (if we haven't made any jumps, the probability of jumping from x to y is 0), $p_t(x, y) = (n = 0 \text{ term }) + \lambda t e^{-\lambda t} \frac{(\lambda t)^n}{n!} p(x, y) + \sum_{n=2}^{\infty} p^n(x, y) P(N(t) = n)$. Let (1) = $\lambda t e^{-\lambda t} \frac{(\lambda t)^n}{n!} p(x, y)$, (2) = $\sum_{n=2}^{\infty} p^n(x, y) P(N(t) = n)$.
    - Then $(2) \leq \sum_{n=2}^{\infty} P(N(t) = n) = P(N(t) \geq 2) = 1 - P(N(t) = 0 \text{ or } 1) = 1 - e^{-\lambda t} - \lambda t e^{-\lambda t} = 1 - e^{-\lambda t}(1 + \lambda t)$ Then plugging in the taylor expansion, $= 1 - (1 - \lambda t + O(t^2))(1 + \lambda t) = O(t^2)$. So $q(x, y) = \lim_{t \to 0} \frac{1}{t} p_t(x, y) = \lim_{t \to 0} [\lambda t e^{-\lambda t} p(x, y) + O(t)] = \lambda p(x, y)$
- Example 1 (M / M / k queue): M = memoryless (exponential here), customer interarrivals, M = memoryless service times, k is # of servers. Customers arrive according to $PP(\lambda)$, service times are $Exp(\mu)$, independent, k servers, $X_t =$ # in bank.
    - Interarrivals are memoryless exponentials from PP. The service times are also independent $exp(\mu)$. X_t is a CTMC, the # of customers still in sevice or waiting.
    - Jump rates: $q(n, n+1) = \lambda$, another person enters the queue. This is just the PP rate λ since this governs

customer arrivals. To go down one, $q(n,n-1) = \begin{cases} n\mu & 0 \le n \le k \\ k\mu & k \le n \end{cases}$. The first case is $n\mu$ because all n enter service, ie. there is no line / wait because n is less than the number of servers. The time until first service is completed is an exponential race of $exp(n\mu)$, the minimum of the n independent service times of the n servers. Similarly, when n exceeds k, n-k customers wait in line and we have a race among the k servers at capacity.

- Example 2 (Continuous Time Branching Process): Also allowing a death rate parameter. Let X_t = population size at time t, just like in discrete time. Here we have individuals give birth to one at a time instead of a certain number of offspring at a given discrete time. People independently die at rate μ and give birth at rate λ.
  - Given n people, what is rate of the appearance of one more.$q(n,n+1) = \lambda n \; n \ge 1$, since it is an exponential race between n independent $exp(\lambda)$'s.
  - Similarly, first death is an exponential race $q(n,n-1) = \mu n, \; n \ge 1$. All other rates are zero, we just have transitions to neighboring states with state space equal to the non-negative integers. $\mu = 0$ is the Yule process, all births and no deaths, more similar to the Galton-Watson branching process.

Chain Construction

- **Chain Construction**: Given jump rates, $q(x,y), \; x, y \in S$ distinct, how can we construct the chain?
  - For a state x, $\lambda(x) =$ rate of leaving state x $= \sum_{y \ne x} q(x,y)$.
  - **Routing Matrix**: A transition matix for a discrete time MC. For $x \in S, \; \lambda(x) > 0$ (positive rate of leaving) we set $R(x,y) = \frac{q(x,y)}{\lambda(x)}$ for $x \ne y, \; R(x,x) = 0$. For $x \in S$ with $\lambda(x) = 0$ (rate of leaving is 0), we have an absorbing state, $R(x,x) = 1, \; R(x,y) = 0$ for $x \ne y$. Note R is a stochastic matrix.
  - We then build a continuous transition matrix from our discrete routing matrix, add continuous waiting times between transitions. X walks around the state space in continuous time, and Y walks around the space in discrete time. Let t be the continuous time parameters and n be the discrete time parameter.
  - Let $(Y_n)_{n \ge 0}$ be a discrete time MC on S, with transition probabilities $R(x,y)$ and $Y_0 = X_0$. Let $\tau_0, \tau_1, \ldots$ be iid $\sim exp(1)$, independent of the chain $(Y_n)_{n \ge 0}$. If $\lambda(Y_0) = 0$, then $X_t = Y_0 = X_0, \; \forall t \ge 0$ (we started in an absorbing state and make no moves). Otherwise X_t stays at Y_0 for time $\frac{\tau_0}{\lambda(Y_0)}$, so X_t leaves at rate $\lambda(Y_0)$. It then jumps to Y_1 after this waiting time $\frac{\tau_0}{\lambda(Y_0)}$.
  - X_t stays at Y_1 for time $\frac{\tau_1}{\lambda(Y_1)}$, unless $\lambda(Y_1) = 0$ in which case $X_t = Y_1, \; \forall t \ge \frac{\tau_0}{\lambda(Y_0)}$. It then jumps to Y_2 if not absorbing.
  - Continuing in this way, get a sequence of arrival times $T_0 = 0$ and $T_n = \frac{T_0}{\lambda(Y_0)} + \frac{T_1}{\lambda(Y_1)} + \ldots + \frac{T_{n-1}}{\lambda(Y_{n-1})}, \; n \ge 1$, time of arrival to state Y_n.
  - $X_t = Y_n$ for $T_n \le t < T_{n+1}$, we constructed the same sequence as Y, but now know the time sequence X spends in each state. X_t is a CTMC with initial state $X_0 \overset{d}{=} Y_0$ and jump rates $q(x,y)$.

Markov Chain Generators

- Let $Q(x,y) = \begin{cases} q(x,y) & x \ne y \\ -\lambda(x) & x = y \end{cases}$. Q is the generator of the MC.
- Lets recall $\lambda(x) =$ rate of leaving state x $= \sum_{y \ne x} q(x,y)$ and $q(x,y) = \lim_{h \to 0} \frac{p_h(x,y)}{h} = p_t'(x,y)|_{t=0} = \frac{d}{dt} p_t(x,y)|_{t=0}$
- Now we will redefine in terms of matrix exponentials. Let $e^{Qt} := \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n$, an infinite series.
- Theorem A: For all $t \ge 0$, $p_t = e^{Qt}$
- Theorem B: (1) (**Kolmogorov Backward Equation**) For all $t \ge 0$, $p_t' = Qp_t$ ($\forall x, y \in S, \; \frac{d}{dt} p_t(x,y) = (Qp_t)(x,y)$). (2) (**Kolmogorov Forward Equation**) $p_t' = p_t Q$ (the matrices commute).
  - From the matrix exponential, we can see why they commute. If p_t is the infinite series $e^{Qt} = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n$, multiplying by Q on left or right, get Q^{n+1}.
  - Theorem B implies Theorem A, since Thm A is the solution to these differential equations. The solution for the

backward equation is $p_t = e^{Qt}p_0$, and $p_0(x,y) = I$, since it is the probability of instantaneous transition from x to y. So we get $p_b = e^{Qt}p_0 = e^{Qt}$

- Lemma (Chapman-Kolmogorov Identities): $\forall s, t$ pairs of times and $\forall x, y \in S,\ p_{s+t}(x,y) = \sum_{z\in S} p_s(x,z)p_t(z,y) = p_{s+t} = p_s p_t \text{ (as matrices)}$.
    - Similar to what we had in discrete time
    - Proof: $p_{s+t}(x,y) = P(X_{s+t} = y|X_0 = x) = \sum_{z\in S} P(X_{s+t} = y|X_0 = x, X_s = z)P(X_s = z|X_0 = x)$. By defn $P(X_s = z|X_0 = x) = p_s(x,z)$. Now using the Markov property, conditioning on $X_0 = x, X_s = z$ is the same as conditioning on $X_s = z$, and using our time homogenous defn $P(X_{s+t} = y|X_s = z) = P(X_t = y|X_0 = z) = p_t(z,y)$.
- Proof of Theorem B
    - Backward equation: Let $t > 0,\ h > 0$ small. By C-K identities lemma, $p_{t+h}(x,y) - p_t(x,y) = (\sum_z p_h(x,z)p_t(z,y)) - p_t(x,y)$. We have broken the interval t + h into the short interval h and the rest, t. Now $\sum_z p_h(x,z)p_t(z,y) = \left(\sum_{z\neq x} p_h(x,z)p_t(z,y)\right) + (p_h(x,x) - 1)p_t(x,y)$. Note $p_h(x,x) - 1 = -\sum_{z\neq x} p_h(x,z)$. Dividing by h, $\implies p_t' = \frac{p_{t+h}(x,y)-p_t(x,y)}{h} = \left[\sum_{z\neq x}\left(\lim_{h\to 0}\frac{p_h(x,z)}{h}\right)p_t(z,y)\right] - p_t(x,y)\left(\sum_{z\neq x}\lim_{h\to 0}\frac{p_h(x,z)}{h}\right)$. Then $\lim_{h\to 0}\frac{p_h(x,z)}{h} = q(x,z)$. So $= \left(\sum_{z\neq x} q(x,z)p_t(z,y)\right) - p_t(x,y)\lambda(x) = (Qp_t)(x,y)$
    - The forward equation, similar lines. Instead writing $p_{t+h}(x,y) = \sum_z p_t(x,z)p_h(z,y))$

Stationary Distributions & Limiting Behavior

- Definition: A probability distribution π on S is a stationary distribution for a CTMC $(X_t)_{t\geq 0}$ with transition probabiltiies $p_t(x,y)$ if $\pi p_t = \pi,\ \forall t \geq 0$.
- Proposition: Let $(X_t)_{t\geq 0}$ a CTMC with generator Q. Then a probability distribution π on S is a **stationary distribution** for the chain if and only if $\pi Q = 0$
    - Written out, $\forall y\ inS,\ \sum_{z\neq y}\pi(z)q(z,y) = \pi(y)\lambda(y)$ Equivalent expression to $\pi Q = 0$. If we imagine π is a distribution of mass on the state space, the LHS is saying the total amount of mass entering state y, RHS is the total amount leaving state y. The amount entering is the amount leaving. After rerouting everything according to the q's, we have the same mass distribution as before.
    - Proof: Starting with necessary condition. If $\pi p_t = \pi$ for all t, then $\forall y \in S$, since π does not depend on time $0 = \frac{d}{dt}\pi(y) = \frac{d}{dt}(\pi p_t)(y)$ using the identity given. Then this derivative is equal to $= \frac{d}{dt}\sum_{x\in S}\pi(x)p_t(x,y) = \sum_{x\in S}\pi(x)p_t'(x,y)$. Now using the Kolmogorov forward equation, $= \sum_{x\in S}\pi(x)\sum_{z\in S}p_t(x,z)Q(z,y) = \sum_{z\in S}\left(\sum_{x\in S}\pi(x)p_t(x,z)\right)Q(z,y) = \sum_{z\in S}\pi(z)Q(z,y) = (\pi Q)(y)$.
    - Sufficient condition: If $\pi Q = 0$, $\frac{d}{dt}(\pi p_t)(y) = \sum_x \pi(x)p_t'(x,y)$. Using the backward equation, $= \sum_x \pi(x)\sum_z Q(x,z)p_t(z,y) = \sum_z\left(\sum_x \pi(x)Q(x,z)\right)p_t(z,y) = \sum_z\left(\sum_x(\pi Q)(x,z)\right)p_t(z,y)$ and since $\pi Q = 0$, whole expression equals 0. So $(\pi p_t)(y)$ is constant $\implies \pi p_t = \pi p_0,\ \forall t \geq 0$. p_0 is the identity matrix as seen previously, $\implies \pi p_t = \pi$.
- One nice thing is we do not have to assume aperiodicity; here because everythng happens at exponential waiting times, we have different time intervals between events and fixes the periodic issues from discrete time. Just need irreducibility assumptions
- Definition: A CTMC $(X_t)_{t\geq 0}$ with jump rates q(x,y) is **irreducible** if there exists $x = x_0, x_1, .., x_n = y$ such that the rates $q(x_{i-1}, x_i) > 0,\ \forall 1 \leq i \leq n$.
    - A path from x to y with positive rates all the way
- **Theorem (Convergence)**: If a CTMC $(X_t)_{t\geq 0}$ is irreducible and has a stationary distribution π, then $\forall x, y \in S$, $p_t(x,y) \to \pi(y)$ as $t \to \infty$.
    - The continuous time case of Theorem 1.19, without aperiodicity requred.
    - Forgets about the starting position as t goes to infinity.
    - π gives the limiting proportion of time spent in any given state

- Example: Weather. Sunny for $exp(1/3)$ amount of time, then cloudy and stays cloudy for $exp(1/4)$, then rains for $exp(1)$ amount of time. This would be completely trivial in discrete time, since each state transitions to the next with probability one - deterministic. Now we assign rates to those transitions; while in discrete time we label transitions with probabilities, now label with rates. Can view the rates as the amount of time spent waiting at a state before transitioning to the next, or can see it as the transition rate out of that state. With multiple possible transitions, the time spent will be the sum of rates across possible transition rates. Forming Q matrix, the diagonal entries are negative the row sums: $Q = \begin{bmatrix} -1/3 & 1/3 & 0 \\ 0 & -1/4 & 1/4 \\ 1 & 0 & -1 \end{bmatrix}$.
    - Note in general $Q(x, y)$ is the transition rate off the diagonal, ie. $x \neq y$, but on the diagonal is $-\lambda(x)$, negative the rate of leaving our current state. This is how we originally defined Q.
    - Solve for π SD: $\pi Q = 0$, solving this directly will give infinitely many solutions, since the row sums are zero. Rate of stuff in is rate of stuff out, which is what $\pi Q = (0, 0, 0)$ tells us. We need to employ $\pi 1_S = 1 \implies$ solve \pi A = (0,0,1) replacing one redundant equation with a column of 1's: $A = \begin{bmatrix} -1/3 & 1/3 & 1 \\ 0 & -1/4 & 1 \\ 1 & 0 & 1 \end{bmatrix}$, ie $\pi = (0, 0, 1)A^{-1}$
    - Here get $\pi = (3/8, 4/8, 1/8)$.
    - On average how much time does it take to do a whole weather cycle? 1 + 3 + 4 = 8 - sum of the means of the waiting times. Easy check on our answer for π - expected time to complete cycle is the sum over the reciprocal of rates, then on average 3, 4, and 1 days waiting at S, C, R respectively.
- **Exit distributions and times** with the Q matrix: For $A \subset S$, let $V_A = \min\{t \geq 0 : X_t \in A\}$ first time that you are in set A. Then let $T_A = \min\{t \geq 0 : X_t \in A \text{ and exists } s \in [0, t) \text{ st } X_s \notin A\}$; there is no smallest non zero number in continuous time, so T is the smallest time for which we left A and returned to A between 0 and t. Let $V_x = V_{\{x\}}, T_x = V_{\{x\}}$.
    - Let $A, B \subset S$, $\hat{V} = V_A \wedge V_B = min(V_A, V_B)$. Want $P_X(V_A < V_B)$ for $x \in C = S \backslash (A \cup B)$. Picture: A and B separate in larger space S, X a point in S not in A,B and C all space in S not in A,B. Can use the embedded discrete time MC Y_n; recall Y_n has transition matrix $R(x, y) = \frac{q(x,y)}{\lambda(x)}$ for $x \neq y,\ R(x, x) = 0$.
    - $P_x(V_A < V_B) = P_(Y_n \text{ reaches A before B} | Y_0 = x)$. From the DTMC theory (Thm from Durrett), if C is finite and $P_x(\hat{V} < \infty) > 0,\ \forall x \in C$ then if $h : S \to [0, 1]$, assignment of probabilities to the state space, such that $h(a) = 1\ \forall a \in A,\ h(b) = 0\ \forall b \in B$, and $h(x) = \sum_{y \neq x} R(x, y)h(y)\ \forall y \in C$ then $h(x) = P_x(V_A < V_B)$. This is first step analysis pulled from discrete time theory.
    - Rephrasing $h(x) = \sum_{y \neq x} R(x, y)h(y)$ in terms of the Q matrix; using that $\lambda(x) = -Q(x, x)$ we get $h(x) = \sum_{y \neq x} \frac{Q(x,y)}{\lambda(x)} h(y) \implies -Q(x, x)h(x) = \sum_{x \neq y} Q(x, y)h(y) \implies \sum_{y \in S} Q(x, y)h(y) = 0\ \forall x \in C$.
    - Since h = 1 on A and 0 on B, this simplifies further. Let Q_C be the $(|C| \times |C|)$ submatrix with rows and columns corresponding to $x \in C$. Then we have $h = -Q_C^{-1} v$ where $v(x) = \sum_{y \in A} Q(x, y)$ for $x \in C$. The point here is the matrix Q is bigger than what you need, you only need to look at the submatrix for states in C.
- Example (Queueing): X_t = # of customers in a shop with 2 servers, each having service rate 3 per hour. Customers arrive at rate 2 per hour, but there are only two chairs in the waiting room, so they leave if chairs are occupied. First find the probabilty that starting from x, the shop is devoid of customers before it is full $P_x(V_0 < V_4)$ for x = 1,2,3.
- Note: $Q(x, x-1) = \begin{cases} 3 & x = 1 \\ 6 & x = 2, 3, 4 \end{cases}$ because then there is an exponential race between the two servers for x > 1.
- $Q = \begin{array}{c|ccc|c} -2 & 2 & 0 & 0 & 0 \\ \hline 3 & -5 & 2 & 0 & 0 \\ 0 & 6 & -8 & 2 & 0 \\ 0 & 0 & 6 & -8 & 2 \\ \hline 0 & 0 & 0 & 6 & -6 \end{array}$ Then $Q_C = \begin{array}{ccc} -5 & 2 & 0 \\ 6 & -8 & 2 \\ 0 & 6 & -8 \end{array}$, $v = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} \implies h = (-Q_C)^{-1} \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} = 1/41 \begin{pmatrix} 39 \\ 36 \\ 27 \end{pmatrix}$

Probability Reference

Probabilities of Events

- Intersection - probability that both A and B occur
- Complement - A^c event that A does not occur, all events in the sample space that are not A
- Disjoint - A and C are disjoint if $A \cap C = \emptyset$
- Probability Axioms: 1) $P(\Omega) = 1$, 2) If $A \subset \Omega$ then $P(A) \geq 0$ 3) If A, B disjoint then $P(A \cup B) = P(A) + P(B)$
- Addition Law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Permutation: ordered arrangement of objects
- Binomial coefficients: $(a+b)^n = \sum_{k=0}^{n} a^k b^{n-k}$
- # of ways n objects can be grouped into r classes with n_I in the i^{th} class: $\binom{n}{n_1 n_2 \ldots n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$
- Bayes, multiplication law: $P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum P(A|B_i)P(B_i)}$
- Law of total probability: $P(A) = \sum P(A|B_i)P(B_i)$
- Independence for sets: $P(A \cap B) = P(A)P(B)$. Mutual independence implies pairwise independence

Probability Algebra / Calculus

- Change of variables: $Y = g(X),\ F_Y(y) = \Pr\{Y \leq y\} = \Pr\left\{X \leq g^{-1}(y)\right\} = F_X\left(g^{-1}(y)\right)$. Then $f_Y(y) = \frac{1}{g'(x)} f_X(x),\ \text{where } y = g(x)$

Hazard + Survival Functions

- Survival Function: $S(t) = P(T > t) = 1 - F(t)$. Simply a reversal of the CDF for data consist of time until death or failure, chance of surviving past t.
- Hazard: As the instantaneous death rate for individuals who have survived up to a given time. $h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt} log(S(t))$
- May be thought of as the instantaneous rate of mortality for an individual alive at time t. If T is the lifetime of a manufactured component, it may be natural to think of h(t) as the instantaneous or age-specific failure rate.

Convolutions

- X, Y independent RVs, Z = X + Y, then: $F_Z(z) = \int_{-\infty}^{+\infty} F_X(z-\xi) \mathrm{d}F_Y(\xi) = \int_{-\infty}^{+\infty} F_Y(z-\eta) \mathrm{d}F_X(\eta)$
- With PDFs: $f_Z(z) = \int_{-\infty}^{\infty} f_X(z-\eta) f_Y(\eta) \mathrm{d}\eta = \int_{-\infty}^{+\infty} f_Y(z-\xi) f_X(\xi) \mathrm{d}\xi$
- With nonnegative X and Y: $f_Z(z) = \int_0^z f_X(z-\eta) f_Y(\eta) \mathrm{d}\eta = \int_0^z f_Y(z-\xi) f_X(\xi) \mathrm{d}\xi \quad \text{for } z \geq 0$

Expectations, Moments, Variances

- Expectation: $E(X) = \begin{cases} \sum_x x p_X(x) & \text{discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{continuous} \end{cases}$. Weighted average of x values by probabilities
- Linearity of expectation: $E(aX + bY) = aE(X) + bE(Y)$
- $\operatorname{Var}(X) = E\left\{[X - E(X)]^2\right\} = E\left(X^2\right) - [E(X)]^2$
- MGF uniquely determines a probability distribution - same MGF means same distribution
- $M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$
- The rth moment: $M^{(r)}(0) = E(X^r)$
- $\text{If } X \text{ has the mgf } M_X(t) \text{ and } Y = a + bX, \text{ then } Y \text{ has the mgf } M_Y(t) = e^{at} M_X(bt)$
- If X, Y independent and Z= X + Y, $M_z(t) = M_x(t) M_y(t)$
- mth central moment - the mth moment of the RV $X - \mu x$. Variance is the second central moment. The first central moment is 0

Markov Inequality

- $X \geq 0$, then $P(X \geq t) \leq \frac{E(X)}{t}$
- This result says that the probability that X is much bigger than E(X) is small.

Chebyshev's Inequality

- $P(|X-\mu|>t) \leq \frac{\sigma^2}{t^2}$
- Plug in t = K x sigma to get alternate form. Not necessary to bound X to be positive.

Probabilty Distributions

Discrete Mass Functions

- Bernoulli: $p\left(x\right)=\left\{\begin{matrix} p^x\left(1-p\right)^{1-x}, & if\; x=0\; or\; x=1 \\ 0, & otherwise \end{matrix}\right\},\; 0\leq p\leq 1$
- Binomial: $p(k)=\binom{n}{k}p^k(1-p)^{n-k}$
  $n\in\mathbb{N}, k=0,1,\ldots,n,\; 0\leq p\leq 1$
  - $Bin(n,p)=\sum_{i=1}^n X_i$, for X Bern(P)
  - $E(X)=np, Var(x)=np(1-p)$
  - Can view in terms of Bern: $X=^d \sum_{i=1}^n X_i,\; X_i\; iid\; Ber(p)$
- Geometric - # trials until success
  - $p(k)=p(1-p)^{k-1}, \quad k=\mathbb{N}$
  - $$E(X)=\frac{1}{p}$$ $$\mathrm{Var}(X)=\frac{1-p}{p^2}$$
  - From Bernoulli: $X_1, X_2, \ldots\; iid\; Ber(p)$, then $X=^d\; min\{n: X_n=1\}$ (min of n st X = 1 / heads)
- Geometric - # failures prior to success
  - $p(k)=p(1-p)^k, k=0\cup\mathbb{N}$
  - $E[Z]=\frac{1-p}{p};\quad \mathrm{Var}[Z]=\frac{1-p}{p^2}$
- Negative Binomial Using Geom 1: $P(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r} \qquad ,0\leq p\leq 1, k=r, r+1,\ldots, r=1,2,\ldots,k$
- Negative Binomial Using Geom 2 (failures before rth success): $p(k)=\Pr\{W_r=k\}=\frac{(k+r-1)!}{(r-1)!k!}p^r(1-p)^k$, for $k\in 0\cup\mathbb{N}$
- Hypergeometric
  - *n*: population size; *n*∈N
  - *r*: successes in population; *r*∈{0,1,...,*n*}
  - *m*: number drawn from population; *m*∈{0,1,...,*n*}
  - *X*: number of successes in drawn group
  - $P(X=k)=\frac{\binom{r}{k}\binom{n-r}{m-k}}{\binom{n}{m}}\; \max(0,m+r-n)\leq k\leq \min(r,m)\; 0\leq p(k)\leq 1$
- Poisson: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}\; k=0,1,2,3,\ldots,\;\; \lambda>0$
  - $E(X)=\lambda, Var(X)=\lambda$
  - $e^-\lambda$ is the normalizing factor, since $\sum\frac{\lambda^k}{k!}$ is the power series for e^λ
  - Law of small numbers: The binomial distribution with parameters n and p converges to the Poisson with parameter λ if $n\to\infty$ and $p\to 0$ in such a way that $\lambda=np$ remains constant. In words, given an indefinitely large number of independent trials, where success on each trial occurs with the same arbitrarily small probability, then the total number of successes will follow, approximately, a Poisson distribution. $Bin(n,\frac{\lambda}{n})\to_d\; Pos(\lambda)$
- Multinomial: $\Pr\{X_1=k_1,\ldots,X_r=k_r\}=\begin{cases}\frac{n!}{k_1!\cdots k_r!}p_1^{k_1}\cdots p_r^{k_r} & \text{if } k_1+\cdots+k_r=n \\ 0 & \text{otherwise}\end{cases}$
  - $E\left[X_i\right]=np_i, \mathrm{Var}[X_i]=np_i\left(1-p_i\right),\; \mathrm{Cov}[X_iX_j]=-np_ip_j$

Continuous Density Functions

- Uniform: $f(x)=\begin{cases}1/(b-a) & a\le x\le b\\ 0 & x<a\ or\ x>b\end{cases}$, $x\in[a,b]$, $a<b$, $f:\mathbb{R}\mapsto[0,\infty)$
  - $E(X)=\frac{1}{2}(a+b)$
  - $Var(X)=\frac{1}{12}(b-a)^2$
- Exponential $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$, $\lambda>0$
  - $E(X)=\frac{1}{\lambda}$
  - $Var(X)=\frac{1}{\lambda^2}$
- Normal: $f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $f:\mathbb{R}\to(0,\infty),\mu\in\mathbb{R},\sigma>0$
- Log Normal $V=e^X, X\sim N$: $f_V(v)=\frac{1}{\sqrt{2\pi}\sigma v}e^{-\frac{1}{2}\left(\frac{\ln v-\mu}{\sigma}\right)^2},\quad v\ge 0$
  - $E[V]=e^{\mu+\frac{1}{2}\sigma^2}$
  - $\mathrm{Var}[V]=\exp\{2\left(\mu+\frac{1}{2}\sigma^2\right)\}\left[\exp\{\sigma^2\}-1\right]$
- Gamma:
  - $\Gamma(x)=\int_0^\infty u^{x-1}e^{-u}\,du, x>0$
  - $g(t\mid\alpha,\lambda)=\begin{cases}\frac{\lambda^\alpha}{\Gamma(\alpha)}t^{\alpha-1}e^{-\lambda t}, t\ge 0\\ 0, t<0\end{cases}$
  - $g:\mathbb{R}\to[0,\infty),\alpha>0,\lambda>0$
  - $$E(X)=\frac{\alpha}{\lambda}$$ $$\mathrm{Var}(X)=\frac{\alpha}{\lambda^2}$$
- Beta: $f(u)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}u^{a-1}(1-u)^{b-1}, 0\le u\le 1,\ a,b>0$
  - $$E(X)=\frac{a}{a+b}$$ $$\mathrm{Var}(X)=\frac{ab}{(a+b)^2(a+b+1)}$$
- Cauchy: $f_Z(z)=\frac{1}{\pi(z^2+1)}$ for $z\in(-\infty,\infty)$

Weak Law of Large Numbers

- $X_1\ldots X_i\sim iid$, $E(X_i)=\mu$ and $Var(X_i)=\sigma^2$. Let $X_n=n^{-1}\sum_{i=1}^n X_i$ and
- $lim_{n\to\infty}P(|\bar{X}_n-\mu|>\epsilon)=0$
- Convergence in probability

Central Limit Theorem

- For X_i iid, $E(X)=\mu$, $\ Var(X)=\sigma^2$
- $\lim_{n\to\infty}\Pr\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\le x\right)=\Phi(x)$
- Convergence in distribution

Useful Functions and Integrals

- Gamma
  - $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x,\quad \text{for } \alpha>0$
  - $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$
  - $\Gamma(k)=(k-1)!\quad \text{for } k=1,2,\ldots$
  - $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$
- Stirling's Formula:

    - $n! = n^n \mathrm{e}^{-n} (2\pi n)^{1/2} \mathrm{e}^{r(n)/12n}$ in which $1 - \frac{1}{12n+1} < r(n) < 1$
    - More loosely, $n! \sim n^n \mathrm{e}^{-n} (2\pi n)^{1/2} \quad \text{as } n \to \infty$
    - Implies binomial coefficients $\begin{pmatrix} n \\ k \end{pmatrix} \sim \frac{(n-k)^k}{k!} \quad \text{as } n \to \infty$
- Beta
    - $B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} \mathrm{d}x = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)} \quad \text{for } m > 0, n > 0$
    - $B(m+1, n+1) = \int_0^1 x^m (1-x)^n \mathrm{d}x = \frac{m!n!}{(m+n+1)!}$
- Geometric Series
    - $\sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots = \frac{1}{1-x} \quad \text{for } |x| < 1$
- Sums
    - $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$
    - $1 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$
    - $1 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4}$