Math 340: Lec 2 (Combinatorics)

Asa Royal (ajr74)

January 11, 2024

"There is not much theory involved in combinatorics. For each new situation, we have to think again how the counting should be done. Often there are various approaches leading to the correct answer. THe only way to get acquainted to combinatorics is to train yourself by doing exercises"

Definition 1 (permutation)**.** A **permutation** is an ordering of objects. The number of permutations of size k from a set with n elements is denoted nPk.

$$nPk = \frac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)$$

Remark. If we're selecting an ordered subset of size k, there are n choices for the first element, $n-1$ for the second, and so on, down to $n-k-1$ for the kth element.

Example (Permutations of cards)**.** How many ways can 4 cards be dealt from a 52-card deck if order matters? There are 52 choices for the first card, 51 for the second, 50 for the second, and 49 for the fourth. Thus, there are $52 * 51 * 50 * 49 = 6,497,400$ 4-card permutations.

Corollary 2. There are $n!$ unique permutations of a set of n objects. This is nPn.

Definition 3 (combination)**.** A **combination** is an order-blind subset of objects from a set. The number of subsets of k elements from a set with n elements is denoted $\binom{n}{k}$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Remark. This draws on the formula for permutations, but accounts for the fact that permutations are ordered subsets. If we're selecting a subset of k objects from a set of n objects, there are n choices for the first element, $n-1$ for the second, and so on, down to $n-k-1$ for the kth element. If we multiply those terms, there are $k!$ different orderings of k elements in the factors. If we divide by $k!$, we get the number of unordered combinations.

One can consider the $(n-k)!$ to represent the number of orderings of objects not in our subset (objects we don't select). If $n!$ tells us how many orderings of n objects there are, and we only care about k we select, we need to divide by $(n-k)!$ to get rid of ordering outside our subset selection, then divide by $k!$ again to get rid of orderings in our selection.

Example (Combinations of cards)**.** How many combinations of 4 cards can be dealt from a 52-card deck? There are 52 choices for the first card... 49 choices for the $4th$ card. But we don't care which of the four cards is first, so we divide the product of those factors by 4. We don't care about which of the cards is second, so we divide again by 3, and then 2...

Math 340: Lec 3 (Conditional probability)

Asa Royal (ajr74)

January 18, 2024

Conditional probability

Definition 1 (conditional probability)**.** Given two events $A, B \subset \Omega$ with $\mathbb{P}(B) > 0$,

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

This is the probability that an outcome in A and B occurs divide by the probability that an outcome in B occurs. In a manner of thinking, B has become the updated sample space, and the event we're calculating the probability of is the set of outcomes in A that overlap with B.

Note: conditional probability requires that $\mathbb{P}(B) \neq 0$, because it doesn't make much sense to think about the probability of A given an event that cannot happen.

Conditional probabilities let us update our models with new information. For example, imagine we have some prior expectation about $\mathbb{P}(A)$. If we know B has occurred, an updated forecast for A is $\mathbb{P}(A|B)$.

Bayes' Theorem

Derivation

A corollary of the formula for conditional probability is

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

Since $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{1}$$

Equation (1) is called Bayes' Theorem: it allows us to express the conditional probability of A on B in terms of the conditional probability of B on A, once again giving us a way to update models.

Partition rule

Definition 2 (partition)**.** Events $B_1, \ldots, B_n$ are a partition of Ω if the B_k'sQ are mutually disjoint and their union is equal to Ω.

Theorem 3 (Partition rule)**.** The partition rule states that for any event $A \subset \Omega$, the sets $\{(A \cap B_K)\}^n$ are a partition of A. i.e. the set of those intersections is disjoint and their union is A. Mathematically, by the additivity proposition,

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A|B_k)\mathbb{P}(B_k)$$

p>

Remark. Note 2: We can think think about the calculations involved in the partition formula as cascading probabilities in a decision tree.

Remark. The partition rule is used when a problem's setup includes randomization in the first stage, then selection. The randomization forms partitions! Problem examples: Boxes and marbles.

Theorem 4 (Baye's theorem w/ partition rule)**.** We can express the denominator in Baye's theorem as a sum of disjoint partition intersections. I.e.

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\sum_{k=1}^{n} \mathbb{P}(A|B_k)\mathbb{P}(B_k)}$$

Note 3: We can use the partition rule to calculate the denominator in the Baye's theorem formula by summing up the weighted conditional probability of the denominator event. See e.g., medical diagnosis problem.

Math 340: Lec 4 (Independence)

Asa Royal (ajr74)

January 28, 2024

Intuition/definitions

Remark. Intuitively, we might call two events A and B independent if knowing that one happened gives you no additional information about whether the second will happen. I.e., A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P})(B|A) = \mathbb{P}(B)$. The mathematical definition of independence follows from this.

Definition 1 (Independence)**.** Events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Proof. For two events A and B,

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

Which means

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) * \mathbb{P}(B)$$

And since $\mathbb{P}(A \cap B) = \mathbb{P}(A)$ for independent events, $\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B)$. □

Remark. $\mathbb{P}(A|B)$ only makes sense to think about if $\mathbb{P}(B) \neq 0$. Otherwise, we're considering the probability of A given an event that cannot happen.

Remark. Conditional probabilities let us update our models with new information. For example, imagine we have some prior expectation about $\mathbb{P}(A)$. If we know B has occurred, an updated forecast for A is $\mathbb{P}(A|B)$.

0.1 p/q-coins (repeated coin tosses)

Definition 2 (p-coin)**.** A **p-coin** is a coin that lands head with probability p.

Theorem 3 (Independent coin tosses)**.** Consider a p-coin. Let A_k be the event that k heads are tossed among N independent tosses.

$$\mathbb{P}(A_k) = \sum_{\omega \in A_k} \rho(\omega) = \binom{n}{k} p^k (1-p)^{n-k}$$

In this formula, $\binom{n}{k}$ is the number of sequences ω with exactly k heads, p^k is the probability of seeing k heads, and $(1-p)^{n-k}$ is the probability of seeing $n-k$ heads.

Example. Imagine we toss a fair coin three times. What's the chance we see two heads?

The chance of seeing 2 heads in 2 tosses is 1/4. Now we need to factor in the outcome of the third toss and note that the chance of seeing a tails is 1/2. Why? If we simply said $\mathbb{P}(2 \text{ heads}) = 1/4$, and we dont' care about the third toss, we'd be considering an experiment with two tosses. Instead, we consider the third tails toss another independent event and find that the probability of tossing, e.g., $\{HHT\} = 1/4 * 1/2 = 1/8$.
Now we want to calculate not just the probability of seeing $\omega = \{HHT\}$, but of any combination of tosses that has two heads. So we note that there are $\binom{3}{2}$ ways to arrange the heads in an ω here, so our probability is $1/8 * \binom{3}{2} = 3/8$.

Remark. Note: The assumption of independence is embedded in our assertion that the probability of throwing k heads is p^k, and that we can calculate the probability of seeing k heads and $n-k$ tails by multiplying the probabilites of each individual event.

Math 340: Lec 6 Big Ideas Journal (Law of large numbers)

Asa Royal (ajr74)

January 30, 2024

Remark. The law of small numbers and law of large numbers are both applied in the context of coin tossing when we study A_k, the event where k heads are tossed. **The law of small numbers is applied when $n \to \infty$ and p is very small.** (p equals λn for a constant λ, so it actually approaches 0).

Remark. The law of large numbers is applied when p is fixed and $n \to \infty$. It stems from the intitution is that we expect that when a fair coin is tossed n times, the average proportion of heads will be $1/2$, and the average number of heads will be $n/2$.

Theorem 1 (Law of large numbers (weak))**.** For any $\varepsilon > 0$ and any $n \geq 1$, the probability that more than $(1/2 + \varepsilon)n$ heads are tossed (or that the proportion of heads is greater than $1/2 + \varepsilon$ is bounded by:

$$\mathbb{P}\left(\bigcup_{k \geq (1/2+\varepsilon)n} A_k\right) \leq e^{-\varepsilon^2 n}$$

Corollary 2. We can use the law of large numbers to apply a lower bound, too. The probability that fewer than $(1/2 = \varepsilon)n$ heads are tossed (or that the proportion of heads is less than $1/2 - \varepsilon$ is bounded by:

$$\mathbb{P}\left(\bigcup_{k \leq (1/2-\varepsilon)n} A_k\right) \leq e^{-\varepsilon^2 n}$$

Corollary 3. Unifying the two and thinking about complements, we can bound the probability that we see a proportion or number of heads inside a given range:

$$\mathbb{P}\left(\bigcup_{(1/2-\varepsilon)n < k < (1/2-\varepsilon)n} A_k\right) \geq 1 - 2e^{-\varepsilon^2 n}$$

Corollary 4. Consider $\mathbb{P}(\frac{1}{2} - \delta \leq f \leq \frac{1}{2} + \beta)$ for $\delta \neq \beta$.

$$\mathbb{P}(\frac{1}{2} - \delta \leq f \leq \frac{1}{2} + \beta) = 1 - \mathbb{P}\left(\bigcup_{k \geq (1/2+\beta)n} A_k \cup \bigcup_{k \leq (1/2-\delta)n} A_k\right)$$

Since both sets of events on either side of the union are disjoin, the probability of their union (call it H) is the sum of their probabilities, and is thus, per the law of large numbers, bounded by $H \leq e^{-\beta^2 n} + e^{-\delta^2 n}$. Then

$$\begin{aligned}\mathbb{P}(\frac{1}{2} - \delta \leq f \leq \frac{1}{2} + \beta) &= 1 - H \\ &\geq 1 - e^{-\beta^2 n} + e^{-\delta^2 n}\end{aligned}$$

Math 340: Lec 7 Big Ideas Journal (Law of large numbers)

Asa Royal (ajr74)

February 1, 2024

Definition 1 (random variable)**.** A **random variable** X is a mapping from sample space Ω into R.

Remark. $\mathbb{P}(X = 2$ is essentially the probability that the random variable X outputs 2.

Math 340: Lec 8 Big Ideas Journal (Independence of random variables)

Asa Royal (ajr74)

February 6, 2024

Definition 1 (Independence of random variables)**.** Two discrete random variables X and Y are independent if $\forall a \in \text{range}(X), \forall b \in \text{Range}(Y)$,

$$\mathbb{P}(X = a, Y = b) = \mathbb{P}(X = a)\mathbb{P}(Y = b)$$

Or, assuming $\mathbb{P}(Y = b) \neq 0$,

$$\mathbb{P}(X = a | Y = b) = \mathbb{P}(X = a)$$

Remark. If X and Y are independent random variables, the event that X takes on some value a should give us no information about the value Y takes on.

Theorem 2 (Independence of many random variables)**.** Discrete random variables $X_1, X_2, \ldots, X_n$ are independent iff

$$\mathbb{P}(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n) = \mathbb{P}(X_1 = a_1) * \mathbb{P}(X_2 = a_2) * \ldots * \mathbb{P}(X_n = a_n)$$

where $a_k \in \text{Range}(X_k)$.

Corollary 3. TFAE:

1. $X_1, X_2, \ldots, X_n$ are independent random variables
2. For any intervals $I_1, \ldots, I_n \subset R$,

$$\mathbb{P}(X_1 \in I_1, X_2 \in I_2, \ldots, X_n \in I_n) = \mathbb{P}(X_1 \in I_1) * \mathbb{P}(X_2 \in I_2) * \ldots * \mathbb{P}(X_n \in I_n)$$

3. For any $a_1, \ldots, a_n$

$$\mathbb{P}(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n) = \mathbb{P}(X_1 \leq a_1) * \mathbb{P}(X_2 \leq a_2) * \ldots * \mathbb{P}(X_n \leq a_n)$$

Proof. 2 $\iff$ 3 is trivial. Proved 2 $\implies$ 1 in hw. □

Independence when N is chosen randomly

Example (Trivial non-independence of t/f in N flips)**.** Imagine we toss N p-coins. Let $X = \#$ heads, $Y = \#$ tails. Clearly $X + Y = N$, so we imagine the two r.v.s are not independent. Indeed, observe $\mathbb{P}(X = N, Y = N) = 0$, since we cannot flip n heads and n tails in n tosses. But $\mathbb{P}(X = n) * \mathbb{P}(Y = n) \neq 0$.

Example (Independence of t/f counts when N is Poisson random)**.** Let $N \sim \text{Poisson}(\lambda)$. If we toss N p-coins and let $X = \#$ heads, $Y = \#$ tails, then:

1. X and Y are independent
2. $X \sim \text{Poisson}(\lambda p), Y \sim \text{Poisson}(\lambda(1-p))$

Proof. Proofs are a bitch. Check notes. □

Remark. The difference between the two examples above where X, Y are dependent/independent is that in example 2, N is randomly chosen, so knowing $X = a$ doesn't give us information about Y.

Math 340: Lec 9 Big Ideas Journal (Expectation)

Asa Royal (ajr74)

February 8, 2024

Definition 1 (Expectation of a random variable)**.** If X is a discrete random variable, its expectation (mean) is

$$\mathbb{E}(X) = \sum_{x \in R(x)} x\mathbb{P}(X = x)$$

so long as the series converges absolutely.

Remark. There are a few ways to interpret $\mathbb{E}(X)$.

1. $\mathbb{E}(X)$ is a **weighted average** of the outcomes of X or the "center of probability mass". $\mathbb{E}(X)$ would be the the position of the fulcrum if you graphed the distribution of X) with the x-axis as a lever
2. $\mathbb{E}(X)$ is a **long-run average**. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with the same distribution as some random variable X. For large n, $\frac{1}{n}(X_1 + \ldots + X_n) \approx \mathbb{E}(X)$. Imagine you play n slot machines. Let X_j be your winnings from a slot machine j. $\frac{1}{n}(X_1 + \ldots + X_n)$ is your average slot machine winnings over n plays. We expect this sum will $\approx \mathbb{E}(X)$.

   Remark. The above follows from the weak law of large numbers, which essentially says that the probability that the difference between the average of the X_js and $\mathbb{E}(X)$ is greater than any $\varepsilon > 0$ approaches zero as $n \to \infty$.
3. $\mathbb{E}(X)$ is the **"fair price"** for a random prize.

Properties of expectation

Remark. Consider using the below properties of expectation to calculate expectation when direct calcuation is tricky.

1. **Linearity:** for any two random variables X and Y and any $\alpha, \beta \in \mathbb{R}$,

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$$

2. **Method of indicators:** Let $A \subset \Omega$ be any event. Consider $X(\omega) = \chi_A(\omega)$.

$$\mathbb{E}(X) = \mathbb{E}(\chi_A) = \sum_{x \in R(\chi_a)} x\mathbb{P}(\chi_a = x) = (0)\mathbb{P}(A^c) + (1)\mathbb{P}(A) = \mathbb{P}(A)$$

   Remark. $\mathbb{E}(\chi_A)$ for any indicator χ_A is always $\mathbb{P}(A)$.

3. **Functions:** Let X be a random variable and $g(x) : \mathbb{R} \mapsto \mathbb{R}$. Consider the random variable $g(X(\omega))$.

$$\mathbb{E}(g(X)) = \sum_{x \in R(X)} g(x)\mathbb{P}(X = x)$$

4. **Tail sum formula:** Suppose X has range $R(X) = \{0, 1, 2, \ldots\}$. Then

   Theorem 2 (tail sum formula)**.**

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{j=0}^{\infty} \mathbb{P}(X > j)$$

Math 340: Lec 12 Big Ideas Journal (Variance)

Asa Royal (ajr74)

February 20, 2024

Definition 1 (Variance)**.** For a random variable X, the variance of X is

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{1}$$

or alternatively, if $\mathbb{E}[X] = \mu$,

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] \tag{2}$$

Remark. We can think of variance as quantifying deviation from or closeness to the mean.

Properties of variance

1. $\mathrm{Var}(x) \geq 0$
2. $\mathrm{Var}(X)$ can be ∞ even if $\exists[X] < \infty$
3. $\mathrm{Var}(X) < \infty \iff \mathbb{E}[X^2] < \infty$
4. Scaling: For any $\alpha, \beta \in \mathbb{R}, \mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X)$

Variance of an indicator variable

Proposition 2. Let $A \subset \Omega$ be any event. Consider $X = \chi_A$. $\mathrm{Var}(X) \leq 1/4$

Proof. Generally, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Because X is an indicator variable, $X^2 = X$ and $\mathbb{E}[X] = \mathbb{P}(X)$. Thus $\mathrm{Var}(X) = \mathbb{P}(X) - \mathbb{P}(X)^2 = \mathbb{P}(X)[1 - \mathbb{P}(X)]$. $\mathrm{Var}(X)$ is clearly maximized when $\mathbb{P}P(X) = 1/2$, so $\mathrm{Var}(X) \leq 1/4$. □

Covariance

Definition 3 (Covariance)**.** The covariance of two random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{3}$$

Correlation is a normalized measure of covariance.

Proposition 4. Cov is bilinear because it is an inner product on certain vector spaces (see endof notes). Thus

1. $\mathrm{Cov}(cX, Y) = c\mathrm{Cov}(X,Y)$
2. $\mathrm{Cov}(X, Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$
3. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
4. $\mathrm{Cov}(\sum_{i=1}^m a_i X_i, \sum_{j=1}^n b_j Y_j) = \sum_{i,j} a_i b_j \mathrm{Cov}(X_i, Y_j)$

Proof. $\mathrm{Cov}(X, Y+Z) = \mathbb{E}[X(Y+Z)] - \mathbb{E}[X]\mathbb{E}[Y+Z] = \mathbb{E}[XY] + \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Z] = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$ □

Variance of sums

Proposition 5.

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{j<k} \mathrm{Cov}(X_j, X_k) \tag{4}$$

Remark. Note that

$$\begin{aligned}\mathrm{Var}(X_1 + X_2) &= \mathrm{Cov}(X_1 + X_2, X_1 + X_2) \\ &= \mathrm{Cov}(X_1, X_1) + \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_2, X_1) + \mathrm{Cov}(X_2, X_2) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)\end{aligned}$$

Proof. Applying prop 4.4,

$$\begin{aligned}Var(X_1 + \ldots + X_n) &= \mathrm{Cov}(X_1 + \ldots + X_n, X_1 + \ldots + X_n) \\ &= \mathrm{Cov}(X_1 + X_1) + \mathrm{Cov}(X_2 + X_2) + \ldots + \mathrm{Cov}(X_n + X_n) + \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_2, X_1) + \ldots \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{j<k} \mathrm{Cov}(j, k)\end{aligned}$$

□

Corollary 6. If $X_1, \ldots, X_n$ are independent (or just uncorrelated)

$$\mathrm{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

Vector spaces of random variables

Remark. Given $\Omega, \mathbb{P}$, let S be the set of all random variables X on Ω s.t. $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] < \infty$ (which means $\mathrm{Var}(X) < \infty$). Then S is a vector space. Thus, $X_1, X_2 \in S \implies \alpha X_1 + \beta X_2 \in S$ for any $\alpha, \beta \in R$.

Remark. We can think of Cov as an inner product on S. $\|X_1\| = \mathrm{Cov}(X_1, X_1) = \sqrt{\mathrm{Var}(X)}$, so $\|X_1\|^2 = \mathrm{Var}(X)$.

Remark. Independent vectors in S are orthogonal to each other, so for independent vectors X_1, X_2, $\|X_1 + X_2\| = \|X_1\| + \|X_2\|$. This makes sense, because the $\cos\theta$ term we'd see when calculating out $\|X_1 + X_2\|$ would be obliterated for orthogonal vectors.

Math 340: Lec 13: Variance, Weak LLN, Chebyshev/Markov)

Asa Royal (ajr74)

February 22, 2024

0.1 Variance of sum of iids

Remark. Recall that generally, $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j)$

Proposition 1. If random variables $X_1, \ldots, X_n$ have equal variance,

$$\text{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^n \sigma^2 = n\sigma^2$$

and

$$\text{SD}(X_1 + \ldots + X_n) = \sigma\sqrt{n}$$

Remark. The standard deviation of a sum of iid random variables grows more slowly than the sum of the variables

Example (Compute the vriance of $X \sim \text{Bin}(n, p)$)**.** We can express X as a sum of indicators. Each has teh same variance, $p(1-p)$, so $\text{Var}(X) = np(1-p)$

0.2 Inequalities

Theorem 2 (Markov's inequality)**.** Intuition: Extreme values of a random variable are unlikely, because we need to balance probability mass around the mean. For a positive random variable Y,

$$\mathbb{P}(Y > t) \leq \frac{1}{t}\mathbb{E}[Y]$$

The tail probability of a positive random variable is bounded by its expected value.

Theorem 3 (Chebyshev's inequality)**.** Intuition: If the $\text{Var}(X)$ is small, then X is unikely to be far from its mean. The amount a random variable X with $\text{Var}(X) < \infty$ can vary from its mean, μ is bounded:

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\text{Var}(x)}{t^2}, \forall t > 0$$

Remark. Markov's inequality leads to the weak law of large numbers

Theorem 4 (Weak law of large numbers)**.** Let X be the sum of n iid random variables. When $\text{Var}(X) < \infty$, we can bound the probability that the average of the $X_1, \ldots, X_n$ deviates from the mean of the random variables:

$$\mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| < \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2 n}$$

And even if $\text{Var}(X)$ is infinite, if $\mathbb{E}[X] < \infty$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| > \varepsilon\right) = 0$$

Math 340: Lec 14 Big Ideas Journal (Discrete joint, marginal, and conditional distributions)

Asa Royal (ajr74)

February 27, 2024

Definition of distributions

Definition 1 (joint distribution)**.** Let X and Y be two discrete random variables. Their joint distribution is

$$\mathbb{P}(X = k, Y = j) \text{ for } k \in R(X), j \in R(Y)$$

Remark. The joint distribution can be considered a mass function with domain in the 2D-plane.

Remark. If X, Y are independent, $\mathbb{P}(X = k, Y = j) = \mathbb{P}(X = k)\mathbb{P}(Y = j)$, and we say that the joint probability distribution has a product structure.

Definition 2 (marginal distribution)**.** Let X be a discrete random variable. The marginal distribution of X is $\mathbb{P}(X = k)$.

Remark. Even if a set of three random variables have the same marginal distributions, they may have different joint distributions if one of the random variables is dependent on another.

Definition 3 (Conditional distribution)**.** Let X and Y be discrete random variables. A conditional distribution looks like $\mathbb{P}(X = k|Y = j)$. The conditional distribution is given by:

$$\mathbb{P}(X = k|Y = j) = \frac{\mathbb{P}(X = k, Y = j}{\mathbb{P}(Y = j)}$$

Moving between distributions

Remark. We can find the marginal distribution $\mathbb{P}(X = k|Y = j)$ given the joint probability distribution $\mathbb{P}(X = k, Y = j)$ by normalizing the joint probabilities of X and Y for a given $Y = j$. We can use the same process to find $\mathbb{P}(Y = j|X = k)$.

Remark. We can relate the joint and marginal distributions using the partition rule. In particular, if we know the joint distributions of X and some some other r.v. Y, we can recover the marginal distribution of X (and/or Y). But if we know the marginal distribution of X, we cannot recover the joint distribution of X and Y without add'l info.

Example (Calculating expected value of the marginal distribution using the conditional)**.** By the partition rule,

$$\mathbb{E}[X] = \sum_{j \in R(Y)} \mathbb{E}[X|Y = j]\mathbb{P}(Y = j)$$

This lets us calculate $\mathbb{E}[X]$ when it's difficult to think directly about X but easier to think about $X|Y$. And note

$$\mathbb{E}[X|Y = j] = \sum_{k \in R(X)} k\mathbb{P}(X = k|Y = j)$$

Math 340: Lec 15 (Convolutions, random walks)

Asa Royal (ajr74)

February 29, 2024

0.1 Convolutions

Theorem 1. If we know the marginal distributions of two discrete random variables, X and Y, we can find the marginal distribution of their sum:

$$\mathbb{P}(X+Y=k)=\sum_{y\in R(Y)}\mathbb{P}(X=k-y)\mathbb{P}(Y=y)$$

We refer to that quantity as

$$\rho_X * \rho_Y(k)$$

0.2 Random walks

Definition 2 (Random walk)**.** Intuition: Consider a path on Z with S_n = position at time n and $a\in\mathbb{N}$ as the path's starting point. At each point, fli pa coin. If heads, move in the positive direction. If tails, move in the negative. Mathematically, fix n. $\Omega=\{\omega-(\omega_1,\ldots,\omega_n)|\omega_k\in\{-1,1\}\}$. ω_k is the step taken at time k. $S_k(\omega)$ is the position after the kth step.

$$S_k(\omega)=\alpha+\sum_{j=1}^{k}\omega_j$$

Remark. If coin tosses are fair, $\mathbb{P}(\{\omega\}=2^{-n}$. All paths are equally likely.

Example (Probability of ending at b if starting at a)**.**

$$\mathbb{P}^a(S_n=b)=\frac{\#\text{ paths from a to b}}{\text{total }\#\text{ paths}}=\frac{\binom{n}{\frac{n+b-a}{2}}}{2^n}$$

Theorem 3 (Reflection principle)**.** Let ${N_n}^0(a,b)$ be the number of paths $a\mapsto b$ of length n that touch or cross 0.

$$\begin{aligned}{N_n}^0(a,b)&=N_n(-a,b)\\&=\binom{n}{\frac{n+b+a}{2}}\end{aligned}$$

Remark. Intuition for reflection principle: Imagine the set of paths $X:{S_n}^a=b$ that cross zero. Assume that the paths in X have $S_k=0$. Now reflect each path in X over the x axis before step k. Keep the paths the same after step k. This partially-reflected set represents every path $Y:{S_n}^{-a}=b$; each path in Y is guaranteed to cross zero. Because there is a 1:1 correspondence between paths in X and Y, we can say ${N_n}^0(a,b)=N_n(-a,b)$

See the picture below for an example of the reflection principle. Note the 1:1 correspondence betwen the solid paths and the dashed paths before the x-axis crossing

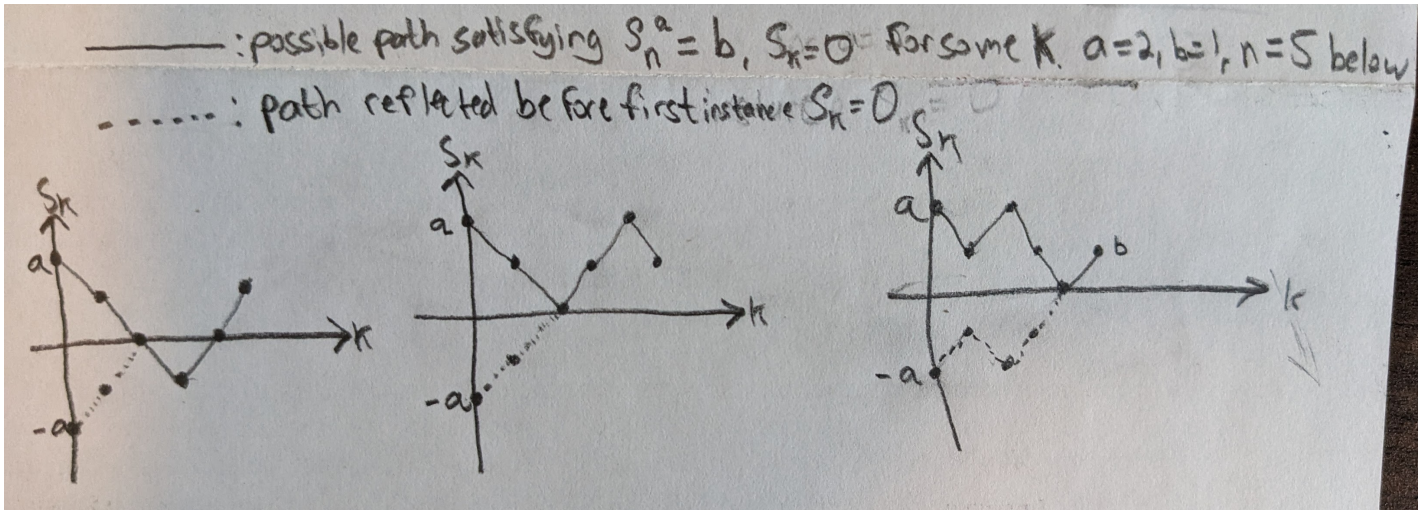


Figure 1: Demonstration of reflection principle. For each solid path that crosses the x-axis, there is a corresponding path reflected before crossing

Example (Applying reflection principle)**.** Compute the probability of going from a to b without hitting 0.

$$\mathbb{P}^a(S_n = b, S_k \neq 0 \text{ for all } k) = 1 - \frac{N_n{}^0(a,b)}{2^n} = 1 - \frac{\binom{n}{\frac{n+b+a}{2}}}{2^n}$$

Math 340: Lec 16 Big Ideas Journal (Random walks continued; central limit theorem)

Asa Royal (ajr74)

March 5, 2024

Random walks

Theorem 1 (Ballot theorem)**.** Consider $N_n^+(0,b)$: paths from $S_0 = 0$ to $S_n = b$ for which $S_k > 0 \forall k \in \{1, \ldots, n\}$. We can think of S_k as how many votes ahead candidate X is over candidate Y on election night.

$$\forall b \neq 0, {N_n}^+(0,b) = \frac{|b|}{n} N_n(0,b)$$

$|b|/n$ is the fraction of paths that don't touch the x-axis (never go negative).

Central limit theorem

CLT applied to random walks

Example (CLT random walks)**.** Consider the following probability measure (bakes in equal parity)

$${\mathbb{P}_n}^0(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-(2n)}$$

Let X_k be a random variable denoting how much we move on step k.

$$\mathbb{E}[S_n] = \mathbb{E}[\alpha + \sum_{k-1}^{n} X_k(\omega) = \sum_{k=1}^{n} \mathbb{E}[X_k] = 0$$

$$\mathrm{Var}(S_n) = \sum_{k=1}^{n} \mathrm{Var}(X_k) = \sum_{k=1}^{n} \mathbb{E}(|X_k - \mu|^2) = \sum_{k=1}^{n} \mathbb{E}(|X_k - 0|^2) = \sum_{k=1}^{n} 1 = n$$

$$SD(S_n) = \sqrt{n}$$

Remark. The CLT suggests us that we shouldn't be surprised if $S_n \approx O(\sqrt{n})$.

Theorem 2 (Central Limit Theorem for random walks)**.** For any α, β

$$\lim_{n\to\infty} \mathbb{P}_n^{\alpha}\left(\alpha \leq \frac{S_n}{\sqrt{n}} \leq \beta\right) = \int_{\alpha}^{\beta} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy$$

Remark. The CLT lets us bound how far away we expect a random walker to wander from the mean of the walk.

General CLT

Theorem 3 (Central Limit Theorem)**.** Suppose $X_1, \ldots, X_n$ is a sequence of i.i.d random variables with $\mu = \mathbb{E}[X_i], \sigma^2 = \mathrm{Var}(X_i)$ and $\mathbb{E}[{X_i}^4] < \infty$. For any α, β

$$\lim_{n\to\infty} \mathbb{P}\left(\alpha \leq \frac{(X_1 + \ldots + X_n) - \mu n}{\sqrt{n\sigma^2}} \leq \beta\right) = \int_{\alpha}^{\beta} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy = \Phi(\beta) - \Phi(\alpha)$$

Remark. $\sqrt{n\sigma^2}$ is the standard deviation for a sum of iid $X_1, \ldots, X_n$. If $X_1 + \ldots + X_n$ were replaced by some other random variable that is a sum of components (e.g. a Poisson arrival process variable), the denominator of the CLT would reflect the SD of that random variable.

Remark. Note that the CLT for random walks is a particular case of the general CLT where S_n is a sum of random step variables with mean zero and variance 1.

Example (Using CLT to bound p-coin head count)**.** Let Z_n be the nubmer of heads we see in n tosses. The marignal distribution of Z_n is given by the binomial distribution. As we know, $Z_n = X_1 + \ldots + X_n$ where X_j is a bernoulli random variable. $\mathbb{E}[X_j] = p$ and $\text{Var}(X_j) = p(1-p)$. Per the CLT:

$$\mathbb{P}\left(\alpha \leq \frac{(X_1 + \ldots + X_n) - np}{\sqrt{np(1-p)}} \leq b\right) = \int_\alpha^\beta \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy$$

Also, note that the CLT expression gives us

$$\begin{aligned} &\mathbb{P}\left(\alpha\sqrt{(np(1-p)} \leq (X_1 + \ldots + X_n) - np \leq \beta\sqrt{np(1-p)}\right) \\ =&\mathbb{P}(Z_n \in (np + \alpha\sqrt{np(1-p)}, np + \beta\sqrt{np(1-p)} \end{aligned}$$

i.e. the probabiliy that Z is within α and β standard deviations of its mean.

Math 340: Lec 17 Big Ideas Journal (Continuously distributed random variables)

Asa Royal (ajr74)

March 7, 2024

Proposition 1. For any continuously-distributed r.v. X, $\forall z, \mathbb{P}(X = z) = 0$.

Remark. Defining outcome spaces and probability measures for continuous random variables is a bit odd. Consider, e.g., the fact that because intervals for continuous r.v. uncountable unions,

$$1 = \mathbb{P}(X \in (\ell, r)) = \mathbb{P}\left(\bigcap_{z \in (\ell,r)} \{X = z\}\right) \neq \sum_{z \in (\ell,r)} \mathbb{P}(X = z) = 0$$

Definition 2 (Cummulative distribution function (CDF))**.** Let X be any real-valued r.v.. The CDF of X is

$$F(z) = \mathbb{P}(X \leq z) = \mathbb{P}(X \in (-\infty, z))$$

Note that $F : \mathbb{R} \mapsto [0, 1]$

Definition 3 (continuous r.v.)**.** Formally, a random variable is **continuous** if its CDF is continuous (and can thus be integrated to recover a density function).

Definition 4 (density function)**.** A continuously-distributed r.v. X has a density $f(x)$ if $\forall a < b \in R$,

$$\mathbb{P}(X \in (a, b)) = \int_a^b f(x)dx$$

Remark. Note that by the fundamental theorem of calculus, $F(b) - F(a) = \int_a^b f(x)dx$ and $F'(z) = f(z)$.

Examples of continuously-distributed r.v.s

1. Uniform distribution
   $X \sim \text{Unif}(\ell, r)$ if
   $$\mathbb{P}(X \in (a, b)) = \frac{b - a}{r - \ell}$$
   if $l < a < b < r$.

2. Exp(λ): exponential distribution w/ paramter λ.
   $$\mathbb{P}(X > t) = e^{-\lambda t}$$
   $$F(t) = \mathbb{P}(X \leq t) = \begin{cases} 1 - e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$
   $$f(t) = F'(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$
   The exponential distribution is used, e.g., to model time until an arrival, akin to how we use the geometric distribution with discretized time.

3. Gaussian (normal) dist
   Let $\mu \in R, \sigma^2 > 0$. $X \sim \text{Normal}(\mu, \sigma^2)$ if

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

   Remark. $N(0,1)$ with $\mu = 0, \sigma^2 = 1$ is called the standard normal.

Proposition 5. Any random variable can be simulated as a function of a uniformly-distributed r.v.
Let $U \sim \text{Unif}(0,1)$. Let F be the CDF of a functino we wish to find an RV for. Assume that F is strictly increasing and continuous. Let $X = F^{-1}(U)$. Then $X \sim F$.

Proof.

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U)) \leq x = \mathbb{P}(U \leq F(X)$$

So X has the same CDF as F. □

Math 340: Lec 18 Big Ideas Journal (Expectation of continuously-distributed random variabless)

Asa Royal (ajr74)

March 19, 2024

Theorem 1 (Expectation of continuously-distributed r.v.s)**.** For a a continuously-distributed random variable X with density $f(z)$,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} zf(z)dz$$

Assuming the integral converges

Properties of expectation for continuous r.v.s

Remark. Many of the properties of expectation of discrete r.v.s carry over to continuous r.v.s, including

1. Linearity: for $\alpha, \beta \in \mathbb{R}$,
$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$$
2. Expectation of functions of r.v.s. In particular, if $g : \mathbb{R} \mapsto \mathbb{R}$,
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(z)f(z)dz$$
   (a) In particular, $\mathbb{E}[X] = \int_{\mathbb{R}} zf(z)dz$ and $\mathbb{E}[X^2] = \int_{\mathbb{R}} z^2 f(z)dz$, and $\mathrm{Var}(X)$ is still $\mathbb{E}[X^2] - \mathbb{E}[X]^2$.
3. Tail sum formula. If X is a positive random variable,
$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X \geq t)dt = \int_0^{\infty} (1 - F(t))\, dt$$
4. Markov's inequality and Chebyshev's inequality hold for positive random variables:
$$\mathbb{P}(X \geq r) \leq \frac{1}{r}\mathbb{E}[X]$$

Remark. When trying to compute the variance of symmetric continuously-distributed random variables, we can try to recenter them around 0 so that $\mathbb{E}[X]^2 = 0$. Then we only need to calculate $\mathbb{E}[X^2] = \int_{-\infty}^{\infty} z^2 f(z)dz$

Math 340: Lec 19 Big Ideas Journal (Joint distributions of cont. r.v.s)

Asa Royal (ajr74)

March 21, 2024

Joint density

Definition 1 (joint density)**.** X and Y have the joint density $f(x, y)$ if

$$\mathbb{P}((x, y) \in B) = \int \int_B f(x, y)dxdy$$

for all open $B \subset \mathbb{R}^2$.

Remark. $\int \int_{\mathbb{R}^2} f(x, y)dxdy = 1$

Theorem 2 (marginal density from joint density)**.** The marginal density of X is

$$f_x(x) = \int_{\mathbb{R}} f(x, y)dy$$

This calculation of the marginal density from the joint density is similar to the one we did in the discrete case. Here, instead of summing over all the discrete values Y can take, we take an infinite sum across the domain of Y, which is $\mathbb{R}$.

Independence

Definition 3 (independence of continuous random variables)**.** Suppose X and Y have densities f_X, f_Y respectively. They are **independent** iff $\forall x, y$, their joint density is

$$f(x, y) = f_X(x)f_Y(y)$$

Remark. Even if X and Y have a constant joint density on some region B, their marginal densities need not be constant.

Examples: moving between joint and marginal densities

Example (Obtain marginal density from joint)**.** Consider the right triangle with vertices at $(0, 0), (1, 0), (1, 1)$. Assume density is uniformly distributed across the triangle.

Note that the marginal density of X is not constant! the density between $(0, \varepsilon)$ and $(1 - \varepsilon, 1)$ is not equivalent! The latter is clearly larger.

We can find $\mathbb{P}(X < a)$ for some $0 < a < 1$ by integrating $f(x, y)$ across the triangle formed by the existing hypotenuse, the x axis, and the line $x = a$.

Example (Obtain joint density from marginals)**.** Imagine $T_1 \sim \text{Unif}(1, 4)$ and $T_2 \sim \text{Unif}(2, 5)$ are independent and represent arrival times. What is $\mathbb{P}(T_1 < T_2)$?

Because the random variables are independent, we can derive their joint distribution from their marginals:

$$f(t_1, t_2) = f_1(t_1)f_2(t_2)$$

Both marginals have density $1/3$ in their respective non-zero areas, and thus $1/9$ on their overlapping non-zero areas. The overlapping area (a square) has corners $(1, 2), (4, 2), (1, 5), (4, 5)$ and area 9. The joint density is thus $1/9$ on the overlapping

region, and the $\mathbb{P}(T_1 < T_2 =$

$$\int\int_B \frac{1}{9} dt_1 dt_2 = \frac{1}{9} B$$

where B is the region in the overlapping areas above the lime $t_1 = t_2$.

Math 340: Lec 20: Conditional Probability Distributions)

Asa Royal (ajr74)

March 26, 2024

Remark. In general, marginal density alone does not determine joint density. Knowing how X is distributed and how Y are distributed is not enough to know how they are jointly distributed.

Conditional probability: discrete and continuous cases

Theorem 1 (Conditional density (discrete))**.** Note that in the discrete case,

$$\begin{aligned}\mathbb{P}(X=k|Y=j) &= \frac{\mathbb{P}(X=k,Y=j)}{\mathbb{P}(Y=j)} \\ &= \frac{\mathbb{P}(X=k,Y=j)}{\sum_{x\in R(X)}\mathbb{P}(Y=j|X=x)\mathbb{P}(X=x)} \\ &= \frac{\mathbb{P}(X=k,Y=j)}{\sum_{x\in R(X)}\mathbb{P}(X=x,Y=j)}\end{aligned}$$

Theorem 2 (Conditional density (continuous))**.** If X, Y have joint density $f(x,y)$ then the conditional density of X given $Y=y$ is

$$f(x|Y=k) = \frac{f(x,k)}{\int_{\mathbb{R}} f(\ell,k)d\ell} = \frac{f(x,k)}{f_Y(y)}$$

Remark. The conditional density is found by taking the joint density with one variable fixed and dividing that by the marginal of the non-fixed variable. Logically, this is very similar to how we find conditioanl probability in the discrete case!

Corollary 3.

$$\begin{aligned}f(x,y) &= f(x|Y=y)f_Y(y) \\ f_X(x) &= \int_{\mathbb{R}} f(x,y)dy = \int_{\mathbb{R}} f(x|Y=y)f_Y(y)dy\end{aligned}$$

Remark. Given iid exponentially-distributed RVs $T_1,\dots,T_n$,

$$\min(T_1,\dots,T_n) \sim \text{Exp}(n\lambda) \tag{1}$$

Math 340: Lec 21 Poisson processes)

Asa Royal (ajr74)

March 28, 2024

0.1 Basics of Poisson arrival processes

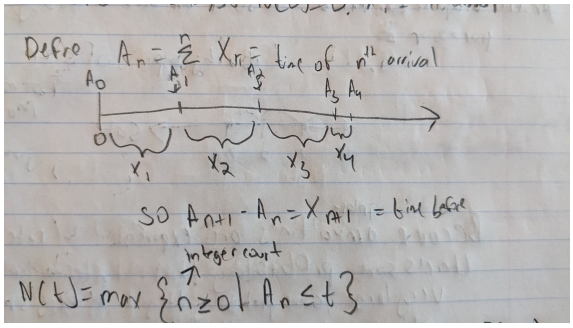
Definition 1 (Poisson arrival proces). $N(t)$ represents a Poisson arrival process, a random function denoting the number of arrivals that occur before some time t.

Let $\lambda > 0, X_1, X_2, X_3, \ldots$ be independent with $X_i \sim \text{Exp}(\lambda)$. X_i represents the ith inter-arrival time (aka waiting time). Since $X_i \sim \text{Exp}(\lambda)$,

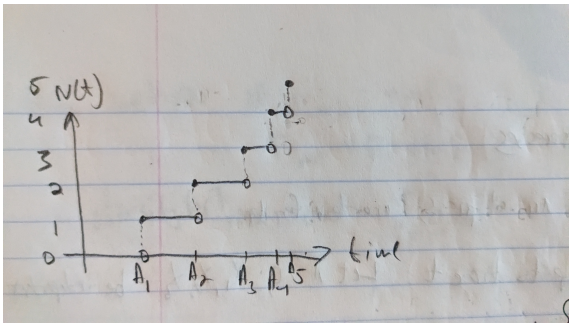
$$\mathbb{P}(X_i > s) = e^{-\lambda s}, \forall s > 0$$

Moreover, if A_n is the nth arrival time, $A_n = \sum_{k=1}^{n} X_k$, so

$$\begin{aligned} N(t) &= \max\{n \geq 0 | A_n \leq t\} \\ &= \max\{\sum_{k=1}^{n} X_k \leq t\} \end{aligned}$$



(a) Inter-arrival and arrival times



(b) N(t) piecewise relationship w/ arrivals

Remark. $N(t)$ is:

1. Non-negative and non-decreasing
2. piecewise constant
3. has jumps of size 1 at arrival times

0.2 Distribution of $N(t)$

Definition 2 (Modified Poisson Arrival Process). For an interval $I = (a, b] \subset (0, \infty)$, define $N(I) = \#$ arrivals in interval I. i.e.

$$N(I) = N(b) - N(a)$$

Note that this means

$$N(t) = N((0, t]) = N(t) - N(0)$$

Theorem 3. Let N be a Poisson arrival process w/ parameter λ.

1. If $I = (a, b]$ is any interval then $N(I)$ has the Poisson($\lambda|I|$) distribution where $|I| = |b - a|$. Thus,

$$\mathbb{P}(N(I) = k) = \frac{(\lambda|I|)^k}{k!} e^{-\lambda t}$$

   And

$$\mathbb{E}[N(I)] = \lambda|I|$$

   (a) In particular, $N(t) \sim \text{Poisson}(\lambda t)$, so $\mathbb{E}[N(t)] = \lambda n$

2. For any disjoint intervals $I_j = (a_j, b_j], j = 1, \ldots, n$, the random variables $N(I_1), N(I_2), \ldots, N(I_n)$ are independent.

0.3 Distribution of nth arrival time

Definition 4 (Gamma distribution)**.** $G(n, \lambda)$ is a continuous distribution on $[0, \infty)$ with density:

$$g_n(t) = \begin{cases} \frac{(\lambda t)^{n-1}(n-1)!}{\lambda} e^{-\lambda t} & \text{if} t \geq 0 \\ 0 & \text{if} t < 0 \end{cases}$$

Remark. Gamma(n, λ) is a distribution of the sum of n independent Exp(λ) random variables.

Proposition 5. The nth arrival time A_n has the Gamma(n, λ) distribution

Math 340: Lec 23 Markov Chains (1)

Asa Royal (ajr74)

April 9, 2024

Remark. Markov chains are useful because they reduce conditional probability calculations to matrix operations

0.1 Overview of Markov Chains

Definition 1 (Markov chain)**.** A **Markov Chain** is a sequence of rnadom variables $X_1, \ldots, X_n$ that takes values in some "state space" S and satisfy the Markov property. The Markov Property states that the future is independent of the past but conditioned on the present. Formally,

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n)$$

Definition 2 (time-homogenous Markov Chain)**.** If we assume that $\mathbb{P}(X_{n+1} = y | X_n = x)$ does not depend on n, we say that the chain is time-homogenous. If a Markov Chain is time-homogenous, we can define a transition matrix, P, which includes the transition probability between all states in the state space.

Theorem 3 (Properties of the transition matrix P)**.** Let P be a transition matrix. Then

1. $0 \leq P(x, y) \leq 1$
2. $\forall x, \sum_{y \in S} P(x, y) = 1$. (Interpretation: If we're at state x, we must move somewhere)

0.2 Examples of markov chains

Example (Simple random walk)**.** A simple random walk on $\mathbb{Z}$ is a Markov Chain with $S = \mathbb{Z}$ and

$$P(x, y) = \begin{cases} 0 & \text{if } |x - y| \neq 1 \\ p & \text{if } |x - y| = 1 \end{cases}$$

Example (Other examples)**.** Simple walk on graph, random process with urn of red/blue marbles where num of given color of marbles change when we pick one of its kind.

0.3 n-step transitions

Motivating question: What is the distribution of X_n given we're at some current state X_0?

Proposition 4.

$$\mathbb{P}(X_n = y | X_0 = x) = P^{(n)}(x, y)$$

Where $P^{(n)}$ is the nth power of the transition matrix P.

Example (Finding two step transition probabilities)**.**

$$
\begin{aligned}
\mathbb{P}(X_2 = x_2 | X_0 = x_0) &= \sum_{x_1 \in S} (X_2 = x_2, X_1 = x_1 | X_0 = x_0) && \text{partitioning} \\
&= \sum_{x_1 \in S} \frac{\mathbb{P}(X_2 = x_2, X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_0 = 0)} && \text{cond. prob} \\
&= \sum_{x_1 \in S} \frac{\mathbb{P}(X_2 = x_2 | X_1 = x_1, X_0 = x_0)\mathbb{P}(X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_0 = 0)} && \text{cond. prob} \\
&= \sum_{x_1 \in S} \frac{\mathbb{P}(X_2 = x_2 | X_1 = x_1)\mathbb{P}(X_1 = x_1, X_0 = x_0)}{\mathbb{P}(X_0 = 0)} && \text{Markov property} \\
&= \sum_{x_1 \in S} \frac{\mathbb{P}(X_2 = x_2 | X_1 = x_1)\mathbb{P}(X_1 = x_1 | X_0 = x_0)\mathbb{P}(X_0 = 0)}{\mathbb{P}(X_0 = 0)} && \text{cond. prob} \\
&= \sum_{x_1 \in S} \mathbb{P}(X_2 = x_2 | X_1 = x_1)\mathbb{P}(X_1 = x_1 | X_0 = x_0) && \text{cond. prob} \\
&= \sum_{x_1 \in S} P(x_0, x_1)P(x_1, x_2) && \text{trans. matrix} \\
&= P^{(2)}(x, y) && \text{def. matrix mult}
\end{aligned}
$$

Math 340: Lec 24 Markov Chains (2)

Asa Royal (ajr74)

April 11, 2024

Remark.

0.1 N-step probability distributions with random start

Proposition 1. Assume $X_0 \sim \nu$ (i.e. $\mathbb{P}(X_0 = x_0) = \nu_{x_0}$). Then the distribution of X_n is given by

$$\mathbb{P}(X_n = y) = \nu(P^n)_y = \sum_{x \in S} \nu_x (P^n)_{x,y}$$

Note that ν is a $1 \times m$ row vector where $m = |S|$. P is obviously $m \times m$.

Proof.

$$\begin{aligned}
\mathbb{P}(X_n = x_n) &= \sum_{x_0 \in S} \mathbb{P}(X_n = x_n | X_0 = x_0)\mathbb{P}(X_0 = x_0) \\
&= \sum_{x_0 \in S} P^{(n)}(x_0, x_n)\nu(x_0) && \text{n-step prob} \\
&= \sum_{x_0 \in S} \nu(x_0)P^{(n)}(x_0, x_n) \\
&= \nu P^{(n)}(x_n) && \text{matrix-vec mult.}
\end{aligned}$$

□

0.2 Stationary distributions

Definition 2 (Stationary distribution)**.** A distribution π on S is stationary if $\pi P = \pi$. This means

$$\mathbb{P}(X_n = y | X_0 \sim \pi) = \mathbb{P}(X_{n-1} = y | X_0 \sim \pi) = \ldots = \mathbb{P}(X_1 = y | X_0 \sim \pi) = \pi(y)$$

Or in English, the chance of hopping to state $y \in S$ is the same regardless of our current state. Also, note that P can be thought of as a linear transformation so

$$\pi P = \pi \implies \pi P^{(n)} = \pi$$

Remark. If the distribution π on S is stationary, π is a left eigenvector of P with eigenvalue 1.

Example (stationary distribution)**.** Consider

$$\pi = \begin{bmatrix} 0.54 & 0.41 & 0.05 \end{bmatrix}, P = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.4 & 0.6 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Think about $(\pi P)_1$. This is the probability that given $x_0 \sim \pi$, after one jump in the Markov chain, we end up at state 1. To find $(\pi P)_1$ we consider the probability of every path to $X_1 = 1$ (i.e. $P_{x,1} * \pi_x$ for $x \in S$).

$$(\pi P)_1 = \sum_{x \in S} \pi_x P_{x,1} = (0.54)(0.7) * (0.41)(0.4) * (0.05)(0) \approx 0.54$$

Math 340: Lec 25 Markov Chains (3)

Asa Royal (ajr74)

April 16, 2024

0.1 Chain properties

Definition 1 (irreducible/reducible)**.** A Markov chain is **irreducible** if it is possible with positive probability to get from any state to any other state. If a chain is not irreducible, it is **reducible**. An irreducible Markov chain is kind of like a connected graph.

Two states $x, y \in S$ **communicate** ($x \longleftrightarrow y$ if it is possible to navigate from either state to the other. i.e. $P^n_{x,y} > 0$ and $P^m_{y,x} > 0$ for $m, n > 0$. If all states communicate, a graph is irreducible.

0.2 State properties

Definition 2 (recurrent/transient states)**.** A state is **recurrent** if $\mathbb{P}(X_n = x \text{ for some } n \geq 1 | X_0 = x) = 1$. That is, a state is recurrent if we are guaranteed to eventually return to it.
If a state is not recurrent, it is **transient**. That means that there is some probability that after visiting it, we may never return to it: $\mathbb{P}(X_n = x \text{ for some } n \geq 1 | X_0 = x) < 1$

Definition 3 (Absorbing state)**.** An **absorbing state** $x \in S$ is a state with $P_{x,x} = 1$. Once the Markov chain reaches an absorbing state, it never moves from it (think of species extinction in a ecosystem population model).

0.2.1 Periodicity

Definition 4 (periodicity)**.** The **period** of a state $x \in S$ is

$$d(x) = \gcd\{n \geq 1 | (P^n)_{x,x} > 0\}$$

This is the gcd of length of all paths that loop from x to x.

Corollary 5. If a chain is irreducible and $P_{x,x} > 0$, then $d(x) = 1$ because we can go from x to x in one step. Thus, any irreducible chain with a self-loop is aperiodic.

Proposition 6. If a chain is irreducible, all its states have the same period. We then define the common period to be the period of the chain. We call a chain **aperiodic** if the period is 1. To show aperiodicity, we can show that the lengths of two return paths to a node are relatively prime.

0.3 Connection between state and chain properties

Theorem 7 (Markov chain <-> state properties)**.** If a Markov chain is irreducible, either

1. All of its states are transient
2. All of its states are recurrent

Corollary 8. If a Markov chain is irreducible and $|S| < \infty$, there must be at least 1 recurrent state, which means all states are recurrent.

0.4 stationary distributions

Theorem 9 (limit converges to stationary)**.** Assume $|S| < \infty$. If a chain is irreducible, then there is a unique invariant (stationary) probability distribution π. Furthermore, if the chain is aperiodic, for any initial distribution ν,

$$\lim_{n\to\infty} \nu P^n = \pi$$

i.e.

$$\lim_{n\to\infty} \mathbb{P}(X_n = y | X_0 \sim \nu) = \pi(y)$$

The distribution of the nth step of the Markov Chain is given by π, no matter our starting place.

Theorem 10 (Ergodic theorem)**.** If a Markov chain is aperiodic and irreducible, for any function $F : S \mapsto \mathbb{R}$ (function on a state), the following holds with probability 1:

$$\underbrace{\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(X_k)}_{\text{temporal avg}} = \underbrace{\sum_{x\in S} F(x)\pi(x)}_{\text{spacial average}}$$

Remark. We can think of F as a cost or reward function that tells us how much it costs or how much we're rewarded for being at some state k. Over time, the average cost will be $F(\mathbb{E}[\pi])$, where π is the stationary distribution. Note that the RHS of theorem 11 looks like an expected value of F on the state space (because it is one...)

Math 340: Lec 26 Markov Chains (4)

Asa Royal (ajr74)

April 18, 2024

0.1 Expected return time

Definition 1 (Return time). Fix $x \in S$. The **first arrival/return time** for S is defined as

$$T_x = min\{n \geq 1 | X_n = x\}$$

Corollary 2. State X is recurrent iff

$$\mathbb{P}(T_x < \infty | X_0 = x) = 1$$

Definition 3 (Expected time of firrst return to x). The epxected time to first return is

$$\mu_x = \mathbb{E}[T_x | X_0 = x]$$

Definition 4 (positive/null recurrence). We say that x is positive recurrent if $\mathbb{E}[T_x | X_0 = x] < \infty$. I.e. we expect to return to x in a finite amount of time. We say that x is null recurrent if $\mathbb{E}[T_x | X_0 = x] = \infty$.

Remark. Example of a null-recurrent markov chain: A simple random walk on $\mathbb{Z}$.

Theorem 5 (Stationary distribution relation to expected first return). An irreducible chhain has a stationary distribution iff all states are positive recorrent. Additionally,

$$\pi_x = \frac{1}{\mathbb{E}[T_x | X_0 = x]}$$

0.2 Q's about random walks

Example (Given two states A and B, what's the probability of reaching A before B?). Imagine we start at state X_0. Define $h(x)$ as the probability of reaching A before B when starting at x. Note that $h(A) = 1, h(B) = 0$ and additionally that $\forall x \in S \backslash \{A, B\}$,

$$h(X) = \sum_{y \in S} h(y) P_{x,y}$$

$h(A), h(B)$, and the $h(x)$ equations for a system of linear equations that can be solved for each $h(x)$.

Example (Expected return time for a simple random walk). For a simple random walk, $h(x) = x/B$. If we've hit B, it would take $2B + 1$ steps to return to zero, so note that

$$\mathbb{P}(T_0 > B | X_0 = 1) \geq \frac{1}{B}$$

And thus by the tail sum formula,

$$\begin{aligned} \mathbb{E}[T_0 | X_0 = 1] &= \sum_{k=1}^{\infty} \mathbb{P}(T_0 \geq t | X_0 = 1) \\ &\geq \sum_{k=1}^{\infty} \frac{1}{k} = \infty \end{aligned}$$

So the expected time of return for a simple random walk is ∞. The Markov Chain is null recurrent!

Math 340: Lec 27 (Markov Chain Monte Carlo algorithms)

Asa Royal (ajr74)

April 23, 2024

1 Motivation for Markov Chain Monte Carlo Algorithms

Remark. We might want to sample a probability distribution

$$\pi(x) = \frac{f(x)}{c} \text{ for } x \in S$$

where we know $f(x)$ but cannot calculate $c = \sum_{x \in S} f(x)$ because the state space is so large.

To efficiently sample from the distribution π, we can try to generate a Markov chain that has π as its stationary distribution.

Remark. Examples of applications:

1. In Bayesian statistics, when we try to calculate $\mathbb{P}(Y = y|X = x)$, the normalizing denominator $\mathbb{P}(X = x)$ can be very expensive to calculate because it requires us to sum over all possible values of the random variable Y.
2. In cryptography, if we have a substitution cipher, we might create a mapping σ from the cipher alphabet to our normal alphabet. We could then decode an encrypted message using σ and measure how much the decrypted message mimics English letter patterns with some function $f(\sigma)$. But assuming the cipher alphabet has 26 letters, there are 26! possible σ mappings. So to normalize the score of any σ, we'd need to calculate all 26! $f(\sigma)$s. Expensive!!

2 Markov Chain Monte Carlo Algorithms

2.1 Metropolis-Hastings

Theorem 1 (Metropolis-Hastings). Objective: sample from $\pi(x) = \frac{f(x)}{c}$ using a proposal function $q(x, y)$. Metropolis-Hastings generates a Markov Chain X_n on S. Given $X_n = x$, M-H generates X_{n+1} as follows:

1. Propose a new state $y \in S$ according to the probability transition kernel $q(x, y)$
2. Accept or reject the Proposition
   y is accepted with probability

$$\min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y))}\right) = \min\left(1, \frac{f(y)q(y,x)}{f(x)q(x,y))}\right)$$

If we accept, $X_{n+1} = y$. Otherwise, $X_{n+1} = X_n = x$.

Remark. π is stationary for this Markov Chain, and with an appropriate kernel q, the chain is irreducible + aperiodic.

Example (Example of accept/reject stage of MH). Imagine we have

$$X_n = \sigma = (1, 3, \ldots, 7, 9, 12)y \qquad = \sigma' = (1, 12, \ldots, 7, 9, 3)$$

We check whether $f(\sigma') > f(\sigma)$. If so, we transition to $X_{n+1} = \sigma'$ with a decent probability.

2.2 Gibbs sampling

Theorem 2 (Gibbs sampling)**.** Imagine we have a a graph (V, E) where the m vertices are pictures in an image recognititon dataset. Edges represent shared features between images. z represents the image of a label . We want to calculate $\pi = f(x)/c$, where $f(x)$ is some function involving the degree of a vertex. But there are so many edges and vertices that calculating c is impractical. Instead, we find π as follows:

1. Pick an index $i \in \{1, \ldots, m\}$ uniformly at random
2. Resample its label according to

$$\mathbb{P}(z_i = c) = \frac{f(z_1, \ldots, z_{i-1}, c, z_{i+1}, \ldots, z_m)}{\sum_{j=1}^{k} f(z_1, \ldots, z_{j-1}, c, z_{j+1}, \ldots, z_m)}$$

   Basically, we try to identify the probabiliy that a vertex's label should be z_k given its neighbors have the labels they do.