Mathematical Tools in Computer Science

Based on lectures by Prof. Nati Linial and Elad Romanov
Huji Fall 2020

Written by Nir Lavee
Thanks to Inbal Magar for transcribing lectures

Last update: 2021-01-07

Contents

Notation

$[N] = \{1, \ldots, N\}$

$f \sim g$ means $\lim_{n\to\infty} f(n)/g(n) = 1$

$f = O(g)$ means there is a constant $c > 0$ such that $|f(n)| \leq cg(n)$ for sufficiently large n

$f = o(g)$ or $f \ll g$ means $\lim_{n\to\infty} f(n)/g(n) = 0$

$f = \omega(g)$ or $f \gg g$ means $\lim_{n\to\infty} f(n)/g(n) = \infty$

$f = \Theta(g)$ means we have both $f = O(g)$ and $g = O(f)$

$\log n$ = logarithm base 2 of n

$\ln n$ = logarithm base e of n

Part I

Lectures

Lecture 1
2020-10-20

1 Probability

1.1 Basics

Definition 1.1.1. A *discrete probability space* is a finite set Ω and a probability function $\Pr : \Omega \to [0,1]$. The elements of Ω are called *elementary events*. Every element $\omega \in \Omega$ has a non-negative number $\Pr[\omega] \geq 0$ such that $\sum_{\omega \in \Omega} \Pr[\omega] = 1$. A subset $A \subseteq \Omega$ is called an *event* and we define $\Pr[A] = \sum_{\omega \in A} \Pr[\omega]$.

Remark 1.1.2. Generally Ω is not necessarily finite, and not every $A \subseteq \Omega$ is necessarily an event, but there is a collection $\mathcal{A} \subseteq 2^{\Omega}$ of the events. It needs to be closed under intersection, countable union and complement. We can take as an example the interval $\Omega = [0,1]$ and with $\Pr[[a,b]] = b - a$. To develop this fully we would need a course in *measure theory*, which we won't go into here.

Definition 1.1.3. Let $(\Omega, \Pr)$ be a probability space and let $A, B \subseteq \Omega$ be events. A and B are called *independent* if $\Pr[A \cap B] = \Pr[A]\Pr[B]$.

Example 1.1.4. Suppose we roll a fair die and flip a fair coin. Define $\Omega = [6] \times [H, T]$, and assign every elementary event probability $\frac{1}{12}$. Let A be the event that the die outcome is even. We have:

$$\Pr[A] = 6 \cdot \frac{1}{12} = \frac{1}{2}$$

because there are 6 elementary events in A, with uniform probability $\frac{1}{12}$. Similarly, let B be the event that the coin outcome is tails. We have:

$$\Pr[B] = \frac{1}{2}$$

Observe that there are 3 elements in $A \cap B$, so:

$$\Pr[A \cap B] = \frac{3}{12} = \frac{1}{2} \cdot \frac{1}{2} = \Pr[A]\Pr[B]$$

Hence A and B are independent. In this case, there is an intuitive reason for independence (the die and the coin don't affect each other), but in other cases it can happen "coincidentally".

Definition 1.1.5. Let A, B be events with $\Pr[A] \neq 0$. We define the probability of B conditional on A as:

$$\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

Intuitively, we are considering the event B as if the probability space was A: we only look at elements in B that are also in A, and we normalize the result by the size of A.

Remark 1.1.6. Notice that A, B are independent iff $\Pr[B \mid A] = \Pr[B]$. Intuitively, knowing that A happened does not affect the probability of B.

Definition 1.1.7. A *random variable* (RV) is a function from Ω. The range of an RV can be any set, usually we will see $\mathbb{R}$, $\mathbb{Z}$, $\{0,1\}$, etc.

Example 1.1.8. We can define Ω to be the set of all courses we have taken, and $X : \Omega \to \mathbb{R}$ to match each course to our grade.

Definition 1.1.9. Let $X, Y : \Omega \to S$ be RVs with finite range, $|S| < \infty$. We say X, Y are *independent* if for all $s, t \in S$ we have

$$\Pr[X = s,\ Y = t] = \Pr[X = s] \Pr[Y = t]$$

Equivalently we can define X, Y to be an independent if for all $P, Q \subseteq S$ we have

$$\Pr[X \in P,\ Y \in Q] = \Pr[X \in P] \Pr[Y \in Q]$$

Remark 1.1.10. This is a slight abuse of notation. Here $X = s$ denotes an event: it is the set of all $\omega \in \Omega$ such that $X(\omega) = s$ (and similarly $X \in P$ is the event $X^{-1}(P) \subseteq \Omega$ and so on). We also used a comma for intersection ("and").

Definition 1.1.11. For a real RV X we define the expected value ("average"):

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \Pr[\omega]\, X(\omega)$$

Example 1.1.12. If X maps courses to grades, then $\mathbb{E}[X]$ is the average grade.

Definition 1.1.13. For a real RV X we define its variance:

$$\begin{aligned}\operatorname{Var}[X] &= \mathbb{E}\left[(\mathbb{E}[X] - X)^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

The variance measures the concentration of values of X around its average. We prove the second equality below.

Proposition 1.1.14. *Linearity of expectation: for RVs X, Y and $a, \alpha, \beta \in \mathbb{R}$ we have:*

$$\mathbb{E}[\alpha X + \beta Y + a] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y] + a$$

Proposition 1.1.15. *We have* $\operatorname{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Proof. We use linearity of expectation:

$$\begin{aligned}\operatorname{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

□

Definition 1.1.16. We often denote $\mathbb{E}[X] = \mu$ and $\operatorname{Var}[X] = \sigma^2$, where σ is called the standard variation.

Theorem 1.1.17 (Markov's inequality)**.** *Let $Y : \Omega \to \mathbb{R}_{\geq 0}$ be a non-negative RV and $\alpha \geq 1$. Then we have:*

$$\Pr\left[Y \geq \alpha \mathbb{E}\left[Y\right]\right] \leq \frac{1}{\alpha}$$

Example 1.1.18. Suppose Ω is a set of courses, and X is a student's grades. Assume $\mathbb{E}\left[X\right] = 70$ is the weighted average. What can we say about the fraction of the courses in which the student got at least 80? To use Markov we set $\alpha = \frac{8}{7}$ and get:

$$\Pr\left[X \geq 80\right] = \Pr\left[X \geq \frac{8}{7} \cdot \mathbb{E}\left[X\right]\right] \leq \frac{1}{\frac{8}{7}} = \frac{7}{8}$$

Theorem 1.1.19 (Chebyshev's inequality)**.** *For any $c > 0$ we have:*

$$\Pr\left[|X - \mu| \geq c\sigma\right] \leq \frac{1}{c^2}$$

Intuitively, we expect X not to be very different from its expected value. It is convenient to measure its distance from μ in units of σ, and the probability that distance is more than $c\sigma$ is bounded by $\frac{1}{c^2}$.

Proof. Use Markov with $Y = (X - \mu)^2$, $\alpha = c^2$. Notice that $\mathbb{E}\left[Y\right] = \mathrm{Var}\left[X\right]$.

$$\begin{aligned}
\Pr\left[Y \geq \alpha \mathbb{E}\left[Y\right]\right] &\leq \frac{1}{\alpha} \\
\Pr\left[(X-\mu)^2 \geq c^2 \mathbb{E}\left[(X-\mu)^2\right]\right] &\leq \frac{1}{c^2} \\
\Pr\left[(X-\mu)^2 \geq c^2\sigma^2\right] &\leq \frac{1}{c^2} \\
\Pr\left[|X-\mu| \geq c\sigma\right] &\leq \frac{1}{c^2}
\end{aligned}$$

□

Definition 1.1.20. An *indicator* is a random variable $X : \Omega \to \{0, 1\}$. It partitions Ω into two disjoint sets, the elements that get 0 and those that get 1. Notice that for indicators, $\mathbb{E}\left[X\right] = \Pr\left[X = 1\right]$.

Problem 1.1.21. Let Ω_n be a probability space with all permutations on $[n]$, with uniform probability: for every $\pi \in S_n$ we set $\Pr\left[\pi\right] = \frac{1}{n!}$. A fixed point of a permutation is a number which is mapped to itself. For example, $(4, 2, 5, 1, 6, 3, 7)$ has two fixed points (2 and 7). What is the average number of fixed points in Ω_n?

Solution. Define $X : \Omega_n \to \{0, 1, 2, \ldots\}$ such that $X(\pi)$ is the number of fixed points in π. We want to calculate $\mathbb{E}\left[X\right]$.

For every $1 \leq i \leq n$, define X_i to be an indicator of whether i is a fixed point.

$$X_i(\pi) = \begin{cases} 1 & \pi(i) = i \\ 0 & \text{else} \end{cases}$$

Then by definition,

$$X = \sum_{i=1}^{n} X_i$$

By linearity of expectation,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \Pr[X_i = 1]$$

There is nothing unique about any particular i, so all terms are equal to $\Pr[X_1 = 1]$.

$$\mathbb{E}[X] = n \cdot \Pr[X_1 = 1]$$

The number of permutations with $\pi(1) = 1$ is $(n-1)!$, and each has probability $\frac{1}{n!}$, so

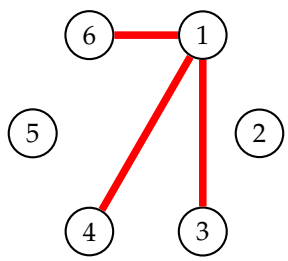
$$\mathbb{E}[X] = n \cdot \frac{(n-1)!}{n!} = n \cdot \frac{1}{n} = 1$$

1.2 Ramsey numbers

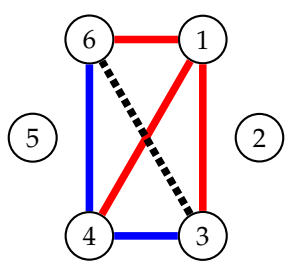
Problem 1.2.1. Show that in a party of 6 people, there are 3 who know each other or 3 who don't know each other (or both). Assume that if a person a knows b then b also knows a.

Solution. Consider the complete graph K_6, where each vertex represents a person. We can color the edges in red and blue, where one color signifies a pair who know each other, and the other color a pair who don't. The problem is then to show there is always a monochromatic triangle.

Consider the vertex 1. It is connected to 5 edges. At least 3 of these belong to one color, say red. Suppose without loss of generality that they look like this:



If $\{3,4\}$ is red or $\{4,6\}$ is red, we have a monochromatic triangle, and we are done. If they are both blue, then $\{3,6\}$ creates such a triangle no matter what color we choose:

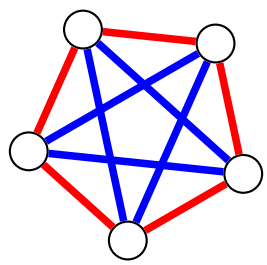


This problem has been studied more generally:

Theorem 1.2.2 (Ramsey)**.** *Let $k, \ell \geq 2$ be natural numbers. For sufficiently large N, every coloring of K_N in red and blue contains a blue k-clique or a red ℓ-clique (or both). The smallest such N is denoted $R(k, \ell)$.*

That is, in a big enough party (with at least $R(k,\ell)$ people), we are guaranteed there are k people who know each other or ℓ who don't (or both).

We have shown above that there is always a monochromatic triangle in K_6, or in other words, $R(3,3) \leq 6$. In fact we have $R(3,3) = 6$: to prove $R(3,3) > 5$, we show a coloring of K_5 with no monochromatic triangles:



Erdős and Szekeres proved the following quantitative version:

Theorem 1.2.3 (Erdős–Szekeres)**.** $R(k,\ell) \leq \binom{k+\ell-2}{k-1}$.

Corollary 1.2.4. $R(k,k) \leq \binom{2k-2}{k-1}$. *With Stirling approximation we can show this is less than* 4^k.

We are also interested in lower bounds on Ramsey numbers:

Proposition 1.2.5. *We have* $R(k,k) \geq \sqrt{2}^k$.

Proof. Let $N = \sqrt{2}^k$, so that $k = 2\log N$. We want to show there exists a coloring of K_N with no monochromatic clique of size k. The idea is to color randomly and prove that the probability there is no monochromatic k-clique is positive (in fact, it is very high).

Consider the probability space Ω of all colorings of K_N with uniform probability. There are 2 options per edge and $\binom{N}{2}$ edges, so there are $2^{\binom{N}{2}}$ elements in Ω.

Fix some natural number t and define $X, Y : \Omega \to \{0,1,2,\ldots\}$ as:

$$X(\omega) = \text{number of blue } t\text{-cliques in } \omega$$
$$Y(\omega) = \text{number of red } t\text{-cliques in } \omega$$

The number of monochromatic t-cliques is $X + Y$.

Key idea: suppose we could show, for some specific t, that the average number of monochromatic t-cliques is < 1 (that is, $\mathbb{E}[X] + \mathbb{E}[Y] < 1$). This would imply the existence of a coloring with no such cliques (similarly, if the average number of cats per household is < 1, there exists a house with no cats).

By symmetry, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are the same, so let's find t such that $\mathbb{E}[X] < \frac{1}{2}$. For every subset $S \subseteq [N]$ of size t, define X_S to be an indicator of whether S forms a blue clique. Then:

$$X = \sum_{|S|=t} X_S$$

where the sum is over all $\binom{N}{t}$ subsets of size t. By linearity of expectation,

$$\mathbb{E}[X] = \sum_{|S|=t} \mathbb{E}[X_S]$$

and by symmetry each term in the sum is equal to the term for a specific $S = \{1, \ldots, t\}$:

$$\mathbb{E}[X] = \binom{N}{t} \Pr[X_S = 1]$$

There are $\binom{t}{2}$ edges in S, and the probability they are all blue is $2^{-\binom{t}{2}}$. So we want to find t such that:

$$\binom{N}{t} 2^{-\binom{t}{2}} < \frac{1}{2}$$
$$\binom{N}{t} < 2^{\binom{t}{2}-1}$$

We will use the approximation $\binom{N}{t} \approx \left(\frac{Ne}{t}\right)^t$ (this strengthens the inequality, see exercise 2):

$$\left(\frac{Ne}{t}\right)^t < 2^{\frac{t(t-1)}{2}-1}$$
$$\frac{Ne}{t} < 2^{\frac{t-1}{2}-\frac{1}{t}}$$

If we take $t = 2\log N$, on the left we have $\frac{N}{\log N}$ (up to a constant) and on the right we have $2^{\frac{2\log N}{2}+O(1)}$ which is about N (up to a constant). Hence the inequality is satisfied for all large enough N, and we conclude there exists a coloring with no monochromatic cliques of size $2\log N$. In other words, $R(k,k) \geq N$. □

1.3 Second moment method

Lecture 2
2020-10-27

Often we work with a probability space that depends on some number n, for example all graphs on n vertices. Then the variables and probabilities we consider are all dependent on n. We will sometimes want to study their behavior when $n \to \infty$. We will say that an event happens "almost surely" or "with high probability" if its probability approaches 1 as $n \to \infty$.

We are interested in the concentration of measure of a random variable: the probability it is very close to its expected value. Chebyshev's inequality gives:

$$\Pr[|X - \mu| \geq c\sigma] \leq \frac{1}{c^2}$$

When choosing c as a function of n, this can potentially give us results of the form

$$\Pr[|X - \mu| > \epsilon] \to 0$$

where ϵ is some small function of n. We can interpret this as X being concentrated around μ.

Proposition 1.3.1. *Let Ω be a probability space defined according to some n. Let $X : \Omega \to \{0, 1, 2, \ldots\}$ be an RV and assume that $\mathbb{E}[X] \to 0$ as $n \to \infty$ (note: $\mathbb{E}[X]$ is a number which depends on n). Then*

$$\Pr[X = 0] \to 1$$

We may also write these as $\mathbb{E}[X] = o(1)$ and $\Pr[X = 0] = 1 - o(1)$.

In the previous lecture we counted the average number $\mathbb{E}[X+Y]$ of monochromatic t-cliques. We showed that it is < 1, so there must be a coloring with no such cliques. In fact the inequality there was strong enough to give $\mathbb{E}[X+Y] = o(1)$, which now we know is an even stronger result: the probability of having a monochromatic t-clique is $o(1)$.

In other words, not only does a desired coloring exist, but a random coloring almost surely has the property we want.

The opposite is not true: if $\mathbb{E}[X] \to \infty$, it does not necessarily mean that $\Pr[X=0] \to 0$. For example, suppose X is 0 with probability $1-\varepsilon$ and $\frac{1}{\varepsilon^2}$ with probability ε. Then:

$$\mathbb{E}[X] = (1-\varepsilon)\cdot 0 + \varepsilon\frac{1}{\varepsilon^2} = \frac{1}{\varepsilon}$$

in this case we see $\mathbb{E}[X] \to \infty$ and $\Pr[X=0] \to 1$ when $\varepsilon \to 0$.

What we can say in such cases is that the variance is large:

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \frac{1}{\varepsilon^4}\cdot\varepsilon - \frac{1}{\varepsilon^2} \\ &= \frac{1}{\varepsilon^3} - \frac{1}{\varepsilon^2} \to \infty\end{aligned}$$

If the variance is not too large, this affects the probability that $X = 0$:

Corollary 1.3.2. *From Chebyshev we know that if X is an RV with $\mathbb{E}[X] \neq 0$,*

$$\Pr[X=0] \leq \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$$

Proof. Chebyshev gives

$$\Pr[|X-\mu| \geq c\sigma] \leq \frac{1}{c^2}$$

Set $c = \frac{\mu}{\sigma}$. The event $X = 0$ is contained in the event $|X-\mu| \geq \mu$:

$$\Pr[X=0] \leq \Pr[|X-\mu| \geq \mu] \leq \frac{\sigma^2}{\mu^2}$$

□

The first moment is expectation and the second moment is variance.

Definition 1.3.3. $G(n,p)$ is a probability space of all undirected graphs on n vertices $1,\ldots n$, where we choose whether an edge $\{u,v\}$ exists by flipping a coin with probability p. Formally, a graph G with $|E|$ edges has probability

$$\Pr[G] = p^{|E|}(1-p)^{\binom{n}{2}-|E|}$$

That is, p per edge and $1-p$ per non-edge.

Remark 1.3.4. Equivalently we can think of K_n where "existing" edges are blue and "non-existing" edges as red. So if we color the edges with uniform probability like in the previous lecture, we are in fact working with $G\left(n,\frac{1}{2}\right)$. Notice that the probability in $G\left(n,\frac{1}{2}\right)$ is uniform, because $\Pr[G] = 2^{-\binom{n}{2}}$ for all G.

We will often consider cases where p is a function of n. Some properties of graphs (such as connectivity) have a *threshold* phenomenon: below a certain value of p they almost never happen; and above a certain of value, they almost always do.

Proposition 1.3.5. *The existence of a copy of K_4 has a threshold of $n^{-2/3}$: for $p = o\left(n^{-2/3}\right)$ we almost never see K_4, and for $p = \omega\left(n^{-2/3}\right)$ we almost always do.*

Proof. We prove both cases.

- Suppose $p \ll n^{-2/3}$. Let X be the number of K_4s in G. As usual, we use indicators:

$$X = \sum_{|T|=4} X_T$$

where the sum is over all subsets T of 4 vertices, and X_T indicates whether T is a clique. By linearity,

$$\mathbb{E}[X] = \sum_{|T|=4} \mathbb{E}[X_T]$$

The probability that 4 particular vertices form a clique is p^6, so:

$$\mathbb{E}[X] = \binom{n}{4} p^6 = \Theta(n^4 p^6)$$

If $p \ll n^{-2/3}$ then $p^6 \ll n^{-4}$ so

$$\mathbb{E}[X] = O(n^4) \cdot o(n^{-4}) = o(1)$$

Therefore $\mathbb{E}[X] \to 0$ and as we have seen above, this means $X = 0$ almost surely.

- Suppose $p \gg n^{-2/3}$. This time $\mathbb{E}[X] \to \infty$, and we want to show $\Pr[X = 0]$ is small. We use the second moment method. We have

$$\Pr[X = 0] \leq \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}$$

To show this approaches 0 it is sufficient to prove

$$\mathrm{Var}[X] = o\left(\mathbb{E}[X]^2\right) = o\left(n^8 p^{12}\right)$$

We calculate the variance as a sum:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{Var}\left[\sum_{|T|=4} X_T\right] \\
&= \sum_{|T|=4} \mathrm{Var}[X_T] + \sum_{\substack{|S|=|T|=4 \\ S \neq T}} \mathrm{Cov}(X_S, X_T) \\
&= \sum_{|T|=4} \left(\mathbb{E}[X_T] - \mathbb{E}[X_T]^2\right) + \sum_{\substack{|S|=|T|=4 \\ S \neq T}} \left(\mathbb{E}[X_S X_T] - \mathbb{E}[X_S]\,\mathbb{E}[X_T]\right)
\end{aligned}$$

The left sum is:

$$\sum_{|T|=4} \left(\mathbb{E}[X_T] - \mathbb{E}[X_T]^2\right) \leq \sum_{|T|=4} \mathbb{E}[X_T] = \mathbb{E}[X]$$

So this part of the variance is indeed $o\left(\mathbb{E}\left[X\right]^2\right)$, because $\mathbb{E}\left[X\right] \to \infty$.

$$\begin{aligned}
\mathbb{E}\left[X\right] &= o\left(\mathbb{E}\left[X\right]^2\right) \\
1 &= o\left(\mathbb{E}\left[X\right]\right) \\
1 &= o\left(n^4 p^6\right) \\
p &= \omega\left(n^{-2/3}\right)
\end{aligned}$$

It is left to show that the right sum (covariances) is also $o\left(\mathbb{E}\left[X\right]^2\right)$. The sum is over all pairs of different subsets S, T of 4 vertices. We split to cases:

- The number of vertices S and T have in common is 0 or 1. In this case, they have no edges in common, so X_S and X_T are independent. Therefore $\mathrm{Cov}(X_S, X_T) = 0$, adding nothing to the sum.
- The number of common vertices is 2. In this case S and T have exactly one edge in common. They form two "conjoined" copies of K_4 with a total of 6 vertices and $12 - 1 = 11$ edges (draw it and count). The probability that these 11 edges exist is p^{11}. The number of ways to choose 6 vertices is $\binom{n}{6} = O(n^6)$, and the number of ways to choose which 2 vertices are common is $\binom{6}{2}$. Therefore:

  $$\sum_{\substack{|S|=|T|=4 \\ |S\cap T|=2}} \mathbb{E}\left[X_S X_T\right] = O\left(n^6 p^{11}\right)$$

  and we have

  $$\begin{aligned}
  \sum_{\substack{|S|=|T|=4 \\ |S\cap T|=2}} \mathrm{Cov}(X_S, X_T) &= \sum_{\substack{|S|=|T|=4 \\ |S\cap T|=2}} \left(\mathbb{E}\left[X_S X_T\right] - \mathbb{E}\left[X_S\right]\mathbb{E}\left[X_T\right]\right) \\
  &\leq \sum_{\substack{|S|=|T|=4 \\ |S\cap T|=2}} \mathbb{E}\left[X_S X_T\right] \\
  &= O\left(n^6 p^{11}\right)
  \end{aligned}$$

  which is $o\left(n^8 p^{12}\right)$ as well.
- The number of common vertices is 3. Then S and T have a triangle in common. Together they have 5 vertices and 9 edges (draw and count), so similarly to the previous case,

  $$\sum_{\substack{|S|=|T|=4 \\ |S\cap T|=3}} \mathrm{Cov}(X_S, X_T) \leq \sum_{\substack{|S|=|T|=4 \\ |S\cap T|=3}} \mathbb{E}\left[X_S X_T\right] = O\left(n^5 p^9\right)$$

  which is $o\left(n^8 p^{12}\right)$ as well.

□

1.4 Balls and bins

We randomly throw balls into n bins, one ball per throw. There are many questions we can ask, such as:

1. When (after how many throws) do we expect the first collision (more than one ball in the same bin)? The answer is about $\sqrt{n}$, and fairly concentrated.
2. Coupon collector: when we do we expect all bins to be non-empty? The answer is about $n \ln n$.
3. When are the bins balanced? For example, when do we have $\frac{\max}{\min} < 1.1$?

We start with 1. A colloquial version of it is called "the birthday paradox": what is the probability that two people in a room share a birthday? Assuming 365 days in a year, if there are 366 people then we are guaranteed a "collision". But surprisingly, with just 23 people the probability of a collision is already $> \frac{1}{2}$.

We will use the following inequality: $e^x \geq 1 + x$ for all x, and it is a good approximation when x is small, since $e^x = 1 + x + O(x^2)$. It is useful to convert $(1 + x)$ factors into exponents.

Problem 1.4.1. What is the probability that there is no collision after r throws?

Solution. After the first ball the probability is 1, then $1 \cdot \left(1 - \frac{1}{n}\right)$ (one bin is not allowed for the second ball), then $1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right)$, and so on. Generally,

$$\begin{aligned}
\Pr\left[\text{no collisions after } r \text{ throws}\right] &= \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{r-1}{n}\right) \\
&\leq e^{-\left(\frac{1}{n} + \frac{2}{n} + \ldots + \frac{r-1}{n}\right)} \\
&= e^{-\frac{r(r-1)}{2n}}
\end{aligned}$$

Therefore if $r \gg \sqrt{n}$, $\frac{r(r-1)}{2n} \to \infty$ and the probability tends to 0. Otherwise, the exponent tends to 0 and the probability tends to 1.

Problem 1.4.2. There are n elements, and we choose random sets A, B with $|A| = a$, $|B| = b$. What is the probability that $A \cap B = \emptyset$?

Solution. When constructing B, initially there are $n - a$ valid elements to choose from out of n total. Then, there are $n - a - 1$ valid elements to choose from, and so on. So the probability is[1]

$$\left(1 - \frac{a}{n}\right)\left(1 - \frac{a+1}{n}\right) \cdots \left(1 - \frac{a+b-1}{n}\right)$$

We can bound this from above similarly:

$$e^{-\left(\frac{a}{n} + \frac{a+1}{n} + \ldots + \frac{a+b-1}{n}\right)} = e^{-\frac{1}{n}(a + (a+1) + \ldots + (a+b-1))}$$

Assume without loss of generality $b \leq a$. There are b terms in the arithmetic series, and each is between a and $2a$. Therefore their sum is $\Theta(ab)$:

$$e^{-\Theta\left(\frac{ab}{n}\right)}$$

If $ab \gg n$ this tends to 0, and if $ab \ll n$ then it tends to 1.

We now turn to question 2: when are all bins non-empty. This is motivated by the problem of buying products with randomly distributed coupons, where the goal is to collect at least one of every coupon type. The types are the bins, and each product purchase is a ball throw.

[1] This seems to be answering a slightly different formulation: it is the probability that a random tuple of b elements has no repetitions and no elements in A.

Definition 1.4.3. For each $k = 0, 1, \ldots$ let T_k be the time from the first moment that k bins are non-empty, until the first moment that $k+1$ bins are non-empty.

When exactly k bins are non-empty, each throw has a probability of $\frac{n-k}{n}$ to enter an empty bin. So T_k is distributed geometrically with parameter $\frac{n-k}{n}$, and we have

$$\mathbb{E}\left[T_k\right] = \frac{n}{n-k}$$

Define T to be the time until all bins are non-empty. Then

$$T = \sum_{k=0}^{n-1} T_k$$

and

$$\mathbb{E}\left[T\right] = \sum_{k=0}^{n-1} \mathbb{E}\left[T_k\right] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{k=0}^{n-1} \frac{1}{n-k} = n \sum_{j=1}^{n} \frac{1}{j} = nH_n = n\left(\ln n + \Theta(1)\right)$$

since the harmonic series H_n grows like $\ln n$ up to an additive constant.

How concentrated is T? Markov gives

$$\Pr\left[T > 10n \ln n\right] \leq \frac{1}{10}$$

Notice that T_k are independent, and the variance of a sum of independent variables is the sum of variances.

$$\mathrm{Var}\left[T\right] = \sum_{k=0}^{n-1} \mathrm{Var}\left[T_k\right]$$

The variance of a geometric RV with parameter p is $\frac{1-p}{p^2} < \frac{1}{p^2}$. Hence

$$\mathrm{Var}\left[T\right] < \sum_{k=0}^{n-1} \frac{n^2}{(n-k)^2} = n^2 \sum_{k=0}^{n-1} \frac{1}{(n-k)^2} = n^2 \sum_{j=1}^{n} \frac{1}{j^2} < n^2 \sum_{j=1}^{\infty} \frac{1}{j^2} = n^2 \cdot \frac{\pi^2}{6}$$

Now with Chebyshev,

$$\Pr\left[T > nH_n + \frac{c\pi n}{\sqrt{6}}\right] \leq \frac{1}{c^2}$$

We can take $c = \sqrt{\ln n}$ so that nH_n is the dominant term, and we find that a deviation of $n\sqrt{\ln n}$ from $\mathbb{E}\left[T\right]$ has probability $\leq \frac{1}{\ln n}$.

1.5 Random walks

Lecture 3
2020-11-03

One dimensional random walk over $\mathbb{Z}$: a person starts at 0. In each step, they flip a coin and go 1 left or 1 right accordingly. This has been studied extensively, and there are many questions we can ask:

1. Where is the person likely to be at time n?
2. What is the farthest from 0 the person has gotten to by time n?

3. How many times has the person visited 0 by time n?

Formally, for every $i \in \mathbb{N}$ define X_i to be the step at time i, where the value of each X_i is 1 or -1, and they are i.i.d (independent and identically distributed). Question 1 is concerned with the following quantity:

$$X = \sum_{i=1}^{n} X_i$$

We will have something to say about X, since sums of i.i.d variables have nice properties. Question 2 is about:

$$Z = \max_{1 \le k \le n} \left| \sum_{i=1}^{k} X_k \right|$$

Question 3 is harder.

Random walks: preliminaries

Before embarking on the journey of random walks, we need to pack a few sandwiches.

Proposition 1.5.1. *For all real t,*

$$\frac{e^t + e^{-t}}{2} \le e^{t^2/2}$$

Proof. Taylor expansion gives:

$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots$$
$$e^{-t} = 1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \dots$$

Then their average is:

$$\frac{e^t + e^{-t}}{2} = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \dots = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!}$$

On the other hand,

$$e^{t^2/2} = 1 + \frac{t^2}{2} + \frac{t^4}{8} + \frac{t^6}{48} + \dots = \sum_{k=0}^{\infty} \frac{t^{2k}}{k! \cdot 2^k}$$

Each term in the latter is $\ge$ each corresponding term in the former:

$$\frac{t^{2k}}{k! \cdot 2^k} \ge \frac{t^{2k}}{(2k)!}$$
$$\frac{(2k)!}{k!} \ge 2^k$$

On the left we have a product of k numbers, $(k+1) \cdots (2k)$, and each one is at least 2. $\square$

Next, we estimate $\binom{n}{k}$ for various values of k.

Proposition 1.5.2. *For $k \ll n$,*

$$\binom{n}{k} \approx \left(\frac{ne}{k}\right)^k$$

For $k = \alpha n$ we need:

Theorem 1.5.3 (Stirling's approximation)**.** *We have*

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

Equivalently,

$$n! = (1 + o(1)) \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$
$$\ln n! = n \ln n - n + \frac{1}{2} \ln n + \frac{1}{2} \ln(2\pi) + o(1)$$

The easier, less precise version:

$$\ln n! = n \ln n - n + O(\ln n)$$

The easier version can be shown by approximating $\ln n! = \ln 1 + \ldots + \ln n$ with an integral (see exercise 2). With a bit more work we can get the $\frac{1}{2} \ln n$ part, but the $\frac{1}{2} \ln(2\pi)$ requires much more. We will use it without proof.

Proposition 1.5.4. *For $k = \frac{n}{2}$,*

$$\binom{n}{\frac{n}{2}} \sim \frac{2^n}{\sqrt{n \frac{\pi}{2}}}$$

Proof. We apply Stirling:

$$\begin{aligned}
\binom{n}{\frac{n}{2}} &= \frac{n!}{\left(\frac{n}{2}\right)!^2} \\
&\sim \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\left(\frac{n}{2e}\right)^{\frac{n}{2}} \sqrt{2\pi \frac{n}{2}}\right)^2} \\
&= \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \cdot \frac{1}{\left(\frac{n}{2e}\right)^n \pi n} \\
&= \frac{2^n}{\sqrt{n \frac{\pi}{2}}}
\end{aligned}$$

□

Remark 1.5.5. Notice that we have

$$2^n = (1+1)^n = \sum_{k=0}^{n} \binom{n}{k}$$

So a naive upper bound on the middle coefficient is $\binom{n}{\frac{n}{2}} < 2^n$. Notice that the middle is also the largest term, so a naive lower bound is

$$2^n = \sum_{k=0}^{n} \binom{n}{k} \leq \sum_{k=0}^{n} \binom{n}{\frac{n}{2}} = (n+1)\binom{n}{\frac{n}{2}}$$
$$\frac{2^n}{n+1} \leq \binom{n}{\frac{n}{2}}$$

So even without Stirling we know the order of magnitude of $\binom{n}{\frac{n}{2}}$ is somewhere between $\frac{2^n}{n+1}$ and 2^n. With Stirling, we concluded it is $\Theta\left(\frac{2^n}{\sqrt{n}}\right)$.

Back to $k = \alpha n$, notice that $\binom{n}{k} = \binom{n}{n-k}$, so it is sufficient to consider $0 < \alpha < \frac{1}{2}$.

Remark 1.5.6. αn might not be an integer. This won't matter to the asymptotic behavior.

Proposition 1.5.7. *Let* $0 < \alpha < \frac{1}{2}$. *We have:*

$$\binom{n}{\alpha n} \sim 2^{nH(\alpha)+O(\log n)}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

Proof. By Stirling,

$$\begin{aligned}\binom{n}{\alpha n} &\approx \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{\alpha n}{e}\right)^{\alpha n} \sqrt{2\pi\alpha n} \left(\frac{(1-\alpha)n}{e}\right)^{(1-\alpha)n} \sqrt{2\pi(1-\alpha)n}} \\ &= \frac{1}{\sqrt{2\pi n}\sqrt{\alpha(1-\alpha)}\left(\alpha^\alpha (1-\alpha)^{(1-\alpha)}\right)^n}\end{aligned}$$

Asymptotically the main player here is the exponential factor:

$$\begin{aligned}\frac{1}{\left(\alpha^\alpha (1-\alpha)^{(1-\alpha)}\right)^n} &= 2^{-n\log\left(\alpha^\alpha (1-\alpha)^{(1-\alpha)}\right)} \\ &= 2^{n(-\alpha\log\alpha-(1-\alpha)\log(1-\alpha))} \\ &= 2^{nH(\alpha)}\end{aligned}$$

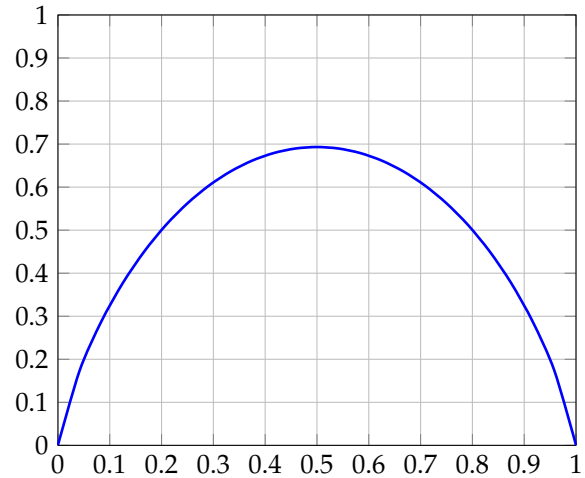
The other factor, $1/\sqrt{2\pi n\alpha(1-\alpha)}$, is $2^{O(\log n)}$, so overall we have $2^{nH(\alpha)+O(\log n)}$. □

Remark 1.5.8. H is called the entropy function. It can be defined for an RV X that can take values $x_1, \ldots, x_k$ with probabilities $p_1, \ldots, p_k$:

$$H(X) = -\sum_{i=1}^{k} p_i \ln p_i$$

Here we used logarithms with base e, but other bases are sometimes of interest, like base 2 above. $H(X)$ captures the notion of amount of information gathered from observing X. For example, if X is an indicator with $\Pr[X = 1] = p$, then

$$H(p) = -p\ln p - (1-p)\ln(1-p)$$



Observing X when it's very likely (or unlikely) gives little information, since the result is more certain. The amount of information gained is maximized when X is a fair coin flip, because the result is very uncertain.

Back to random walks

Suppose the step X_i is ± 1 with probability $p = \frac{1}{2}$, for every i. We will answer question 1: where is the person likely to be after N steps? The answer turns out to be within the range $\left[-\sqrt{N}, \sqrt{N}\right]$ with very high probability.

Proposition 1.5.9. *For $X = \sum_{i=1}^{N} X_i$ we have*

$$\Pr\left[X \geq a\right] \leq e^{-\frac{a^2}{2N}}$$

Proof. Each step has two options, left or right, so there are 2^N possible sequences of steps $\epsilon_1, \ldots, \epsilon_N$ (where $\epsilon_i \in \{1, -1\}$). To arrive at $t \in \mathbb{Z}$, their sum must be t:

$$\Pr\left[X = t\right] = \frac{\left|\{\text{sequences } \epsilon_1, \ldots, \epsilon_N \text{ that sum to } t\}\right|}{2^N}$$

Such a sequence must have exactly $\frac{N+t}{2}$ times 1 and $\frac{N-t}{2}$ times -1. Notice that t and N must be of the same parity: we cannot get to $t = 3$ after $N = 100$ steps, since we alternate between evens and odds. A precise expression for $\Pr\left[X = t\right]$ should mention probability 0 in that case.

The number of such sequences is $\binom{N}{\frac{N-t}{2}}$, because that is the number of ways to choose which subset of $\epsilon_1, \ldots, \epsilon_N$ are -1s.

A technique which is sometimes useful to reason about RV bounds: fix some real number[2] $d > 0$, and notice that the event $X \geq a$ is the same as $e^{dX} \geq e^{da}$. By Markov,

$$\Pr\left[X \geq a\right] = \Pr\left[e^{dX} \geq e^{da}\right] \leq \frac{1}{e^{da}} \mathbb{E}\left[e^{dX}\right]$$

[2]Originally denoted by t in the lecture. I tried to avoid overloading the letter.

And since X is the sum of X_i which are i.i.d,

$$\mathbb{E}\left[e^{dX}\right] = \mathbb{E}\left[e^{d\sum_{i=1}^{N} X_i}\right] = \prod_{i=1}^{N} \mathbb{E}\left[e^{dX_i}\right]$$

With probability $\frac{1}{2}$, X_i is 1 and $e^{dX_i} = e^d$. Otherwise, it is -1 and $e^{dX_i} = e^{-d}$. Therefore:

$$\mathbb{E}\left[e^{dX_i}\right] = \frac{e^d + e^{-d}}{2}$$

And we have seen that this is $\leq e^{d^2/2}$, so

$$\mathbb{E}\left[e^{dX}\right] \leq \prod_{i=1}^{N} e^{\frac{d^2}{2}} = e^{\frac{d^2 N}{2}}$$

Plugging this back in we get:

$$\Pr\left[X \geq a\right] \leq \frac{e^{\frac{d^2 N}{2}}}{e^{da}} = e^{\frac{d^2 N}{2} - da}$$

This is true for any $d > 0$. To make the most of this inequality, we choose $d = \frac{a}{N}$ which minimizes $\frac{d^2 N}{2} - da$. We get:

$$\frac{d^2 N}{2} - da = \frac{a^2 N}{2N^2} - \frac{a^2}{N} = -\frac{a^2}{2N}$$

Hence $\Pr\left[X \geq a\right] \leq e^{-\frac{a^2}{2N}}$ as claimed. □

Corollary 1.5.10. *We are almost never farther than $O(\sqrt{N})$ from the origin: if $a \gg \sqrt{N}$, then* $\Pr\left[X \geq a\right]$ *is $e^{-\frac{a^2}{2N}} \to 0$.*

Notice also that the probability to arrive at exactly t is $\frac{1}{2^N}\binom{N}{\frac{N-t}{2}}$, which is a binomial distribution with a peak around $\left[-\sqrt{N}, \sqrt{N}\right]$. In particular for $t = 0$, we have seen $\frac{1}{2^N}\binom{N}{\frac{N}{2}} = \Theta\left(\frac{1}{\sqrt{N}}\right)$. The sum of $\frac{1}{\sqrt{N}}$ diverges, so we expect to return to the origin infinitely many times. This is called a recurrent walk, or a drunk walk.

Such calculations for $\mathbb{Z}^2$ yield $\frac{1}{N}$, where the sum also diverges so it is recurrent. However, for $\mathbb{Z}^3$ the probability is $\frac{1}{\sqrt{N}^3}$, and the sum does not diverge (a drunk bird does not return home).

The peak around $\pm\sqrt{N}$ is a special case of the central limit theorem:

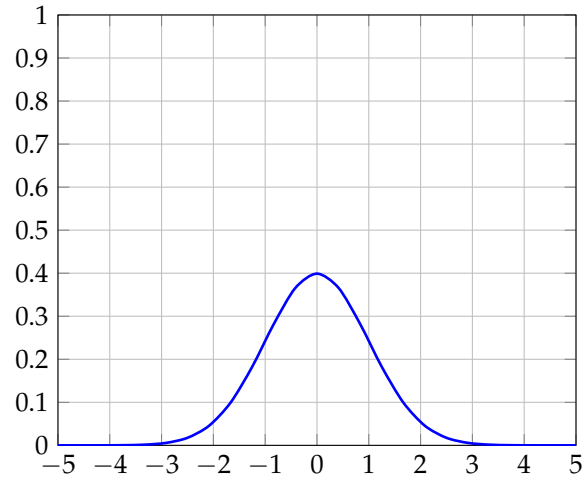
Theorem 1.5.11 (The central limit theorem)**.** *For $X_1, \ldots, X_n$ i.i.d with expectation μ and variance σ^2, define*

$$Z = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}$$

Then Z converges in distribution to $\mathcal{N}(0,1)$. That is,

$$\lim_{n\to\infty} \Pr\left[Z \leq a\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2} dt$$

Generally a density function is a non-negative function with integral 1. The standard normal distribution is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.



Theorem 1.5.12 (Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be independent indicators, and let $X = \sum_{i=1}^n X_i$. Then for all $\delta > 0$,*

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

This is exponentially small because $\frac{e^\delta}{(1+\delta)^{1+\delta}} < 1$. A similar calculation will give

$$\Pr\left[X \leq (1-\delta)\mu\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^\mu$$

Proof. Let $a = (1+\delta)\mu$. As before with Markov and any $t > 0$,

$$\Pr\left[X \geq a\right] = \Pr\left[e^{tX} \geq e^{ta}\right] \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}}$$

and by independence,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t\sum X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{tX_i}\right]$$

Denote $\Pr\left[X_i = 1\right] = p_i$. Then with probability p_i we have $e^{tX_i} = e^t$, and with probability $1 - p_i$ we have $e^{tX_i} = e^0 = 1$. Therefore:

$$\begin{aligned}
\prod_{i=1}^n \mathbb{E}\left[e^{tX_i}\right] &= \prod_{i=1}^n \left(p_i e^t + (1-p_i)\right) \\
&= \prod_{i=1}^n \left(1 + p_i\left(e^t - 1\right)\right) \\
&\leq \prod_{i=1}^n e^{p_i\left(e^t-1\right)} \\
&= e^{\sum_{i=1}^n p_i\left(e^t-1\right)}
\end{aligned}$$

Notice that $\mu = \sum_{i=1}^n \Pr\left[X_i = 1\right] = \sum_{i=1}^n p_i$, so this is equal to:

$$e^{\mu\left(e^t-1\right)}$$

Overall for all $t > 0$ we have:

$$\Pr[X \geq a] \leq e^{\mu(e^t - 1) - ta}$$

Choose t to minimize this expression.

$$\left(\mu(e^t - 1) - ta\right)' = \mu e^t - a$$

and we get $t = \ln \frac{a}{\mu} = \ln(1+\delta)$. This gives:

$$\Pr[X \geq a] \leq e^{\mu(e^{\ln(1+\delta)} - 1) - \mu(1+\delta)\ln(1+\delta)} = \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

□

2 Linear algebra

Lecture 4/5
2020-11-10/17

2.1 Basics

Definition 2.1.1. Let U, V be vector spaces (also called linear spaces) over $\mathbb{R}$. A *linear transformation* is a function $T : U \to V$ such that:

- T preserves scalar multiplication: for every $x \in U$ and $\alpha \in \mathbb{R}$, we have $T(\alpha x) = \alpha T(x)$.
- T preserves addition: for every $x, y \in U$ we have $T(x + y) = T(x) + T(y)$.

Every linear transformation corresponds to a matrix A such that $T(x) = Ax$ for all $x \in U$, where Ax is the product of a matrix and a vector.

Remark 2.1.2. The product Ax of a matrix and a column vector gives a column vector. We may also be interested in the row vector xA, where x is itself a row vector. We may omit the transpose notation (x^T) when it is obvious from context whether x should be a row or a column.

One of our goals is to understand linear transformations geometrically. For example *scaling* is a linear transformation (say, $T : \mathbb{R}^2 \to \mathbb{R}^2$ which multiplies the x coordinate of every vector by 2). Another is *rotation* of the space around the origin. We will eventually see the SVD theorem, which says every linear map can be viewed as a sequence of such actions.

Definition 2.1.3. Let V be a vector space. The vectors $v_1, \ldots, v_n$ are called a *basis* of V if they are linearly independent, and also every $v \in V$ can be written as $v = \sum_{i=1}^n \alpha_i v_i$ where α_i are scalars.

Example 2.1.4. Suppose T is a linear map which performs scaling on each vector in the basis, that is, $T(v_i) = \gamma_i v_i$ for every $1 \leq i \leq n$ where γ_i are scalars. Then

$$T(v) = T\left(\sum_{i=1}^n \alpha_i v_i\right) = \sum_{i=1}^n T(\alpha_i v_i) = \sum_{i=1}^n \alpha_i T(v_i) = \sum_{i=1}^n \alpha_i \gamma_i v_i$$

Metrics and norms

We want to measure sizes of vectors and distances between them. This is motivated by applications, for example, when we want to approximate a solution to given equations which are too hard (or even impossible) to solve completely.

In order to talk about approximations mathematically, we first define a metric, which captures our intuitive notion of distance.

Definition 2.1.5. Let X be a set and $d : X \times X \to \mathbb{R}$ be a function. (X, d) is called a *metric space*, and d is called a *metric*, if the following hold:

- For all $x, y \in X$ we have $d(x, y) = d(y, x)$ (distance is symmetric).
- For all $x, y \in X$ we have $d(x, y) \geq 0$ (distance is never negative).
- For all $x, y \in X$ we have $d(x, x) = 0$, and if $x \neq y$ then $d(x, y) > 0$ (distance is positive if and only if the points are different).
- For all $x, y, z \in X$ we have $d(x, y) + d(y, z) \geq d(x, z)$ (distance obeys triangle inequality).

Example 2.1.6. Physical (Euclidean) distance is the obvious example, but many other metrics are useful: time, edit distance (amount of modifications needed to turn one string to another), etc. Interestingly, air travel cost is not always a metric, because traveling from x to y through z is sometimes cheaper than a direct flight.

Norms captures our notion of vector length (or size):

Definition 2.1.7. Let U be a vector space. A *norm* is a function $\|\| : U \to \mathbb{R}$ such that:

- For every $u \in U$ we have $\|u\| \geq 0$ (size is never negative).
- We have $\|u\| = 0$ if and only if $u = \mathbf{0}$ (size is zero only at the origin).
- For every $u \in U$ and scalar α, we have $\|\alpha u\| = |\alpha| \, \|u\|$ (scaling a vector scales its size).
- For every $u, v \in U$, we have $\|u + v\| \leq \|u\| + \|v\|$ (triangle inequality).

This definition is similar to the metric definition if we consider distances from the origin (compare by putting $y = \mathbf{0}$ above). One different requirement here is $\|\alpha u\| = |\alpha| \, \|u\|$, which is justified by our intuition for size (for example, we want $\|u\| = \|-u\|$).

Definition 2.1.8. For a normed vector space $(U, \|\|)$ we define the *unit ball*:

$$B_{\|\|} = \{u \in U \mid \|u\| \leq 1\}$$

Observe some important properties of the unit ball:

Definition 2.1.9. A set S of points is called *convex* if for all $x, y \in S$, the segment between x, y is contained in S.

The line[3] passing through $x, y \in S$ (assuming $x \neq y$) is the set of points

$$L_{x,y} = \{tx + (1 - t)y \mid t \in \mathbb{R}\}$$

The segment between x, y is given by restricting to $0 \leq t \leq 1$. We show that the unit ball is convex: for any such $0 \leq t \leq 1$,

$$\|tx + (1 - t)y\| \leq \|tx\| + \|(1 - t)y\| = t \, \|x\| + (1 - t) \, \|y\| \leq t + (1 - t) = 1$$

Definition 2.1.10. A set S of points is called *centrally symmetric* if for all $x \in S$ we have $-x \in S$.

We define three commonly used norms for $\mathbb{R}^n$. The unit ball will look different for each.

[3] This was shown in lecture 5.

Definition 2.1.11. The ℓ_2 norm $\|\|_2$, also called the Euclidean norm, is defined for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ as:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

We show this is indeed a norm by our definition. Clearly it is non-negative, and $\|x\|_2 = 0$ if and only if $x = \mathbf{0}$. We also have $\|\alpha x\|_2 = |\alpha| \, \|x\|_2$ following directly from the definition. For the triangle inequality, consider $x, y \in \mathbb{R}^n$:

$$\begin{aligned}
\|x+y\|_2 &\leq \|x\| + \|y\| \\
\sqrt{\sum_{i=1}^{n} (x_i + y_i)^2} &\leq \sqrt{\sum_{i=1}^{n} x_i^2} + \sqrt{\sum_{i=1}^{n} y_i^2} \\
\sum_{i=1}^{n} (x_i + y_i)^2 &\leq \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 + 2\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \\
2\sum_{i=1}^{n} x_i y_i &\leq 2\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \\
\langle x, y \rangle &\leq \|x\| \, \|y\|
\end{aligned}$$

This follows from Cauchy-Schwarz.

With the Euclidean norm, a unit ball looks like an actual ball in $\mathbb{R}^3$, or a circle in $\mathbb{R}^2$, and generally an n-dimensional ball in $\mathbb{R}^n$.

Definition 2.1.12. The ℓ_1 norm is defined as:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Here too it is easy to verify the norm properties. The triangle inequality for vectors follows from real numbers:

$$\|x+y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} (|x_i| + |y_i|) = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1$$

With ℓ_1 norm, the unit "ball" looks like a diamond in $\mathbb{R}^2$, or an octahedron in $\mathbb{R}^3$.

Definition 2.1.13. The ℓ_∞ norm is defined as:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

In this case the triangle inequality is:

$$\max_{1 \leq \alpha \leq n} |x_\alpha + y_\alpha| \leq \max_{1 \leq \beta \leq n} |x_\beta| + \max_{1 \leq \gamma \leq n} |y_\gamma|$$

Given any norm $\|\|$, we can define the distance between vectors x, y as $\|x - y\|$.

Remark 2.1.14. These norms are all special cases of ℓ_p , defined as $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$ (this works for $p = \infty$ as well, when taking the limit).

Example 2.1.15. Suppose we have a function $f : [-1, 1] \to \mathbb{R}$ which is difficult to work with, so we want to approximate it by some polynomial P. We need to define what a good approximation means, and this depends on the specific application. For example, we may want to define a norm using an integral (analogous to Euclidean):

$$\|f\|_2 = \sqrt{\int_{-1}^{1} f(x)^2 dx}$$

and then we are looking for a polynomial with minimal distance:

$$\min_P \|f - P\|_2^2 = \min_P \int_{-1}^{1} (f(x) - P(x))^2 \, dx$$

Depending on what kind of errors we are willing to tolerate in the application, we may prefer other norms, such as $\max_{-1 \le x \le 1} |f(x)|$ (analogous to ℓ_∞) or $\int_{-1}^{1} |f(x)| \, dx$ (analogous to ℓ_1).

In this example the vector space is the set of all continuous functions on $[-1, 1]$. Note that its dimension is infinite. The set of polynomials is another such space.

Example 2.1.16. Every norm has its unit ball, which is a convex centrally symmetric shape. We can go the other way around: start with such a shape, say a hexagon B centered at the origin. Take any point x outside the hexagon and connect it to the origin. The segment meets some unique z on the hexagon's boundary, with some unique scalar λ such that $\lambda z = x$. We can define $\|x\| = \lambda$. Formally[4],

$$\|x\|_B = \frac{1}{\max\{\lambda > 0 \mid \lambda x \in B\}}$$

Note this works as expected when x is inside B as well.

We are also interested in measuring sizes of matrices.

Definition 2.1.17. Let $(U, \|\|)$ and $(V, |||\|||)$ be normed spaces, and let $T : U \to V$ be a linear transformation. We define the *operator norm* of T:

$$\|T\|_{op} = \max_{\mathbf{0} \neq u \in U} \frac{|||T(u)|||}{\|u\|}$$

That is, the operator norm of T is the largest multiplicative effect it has on a vector's size.

Note: in certain cases the maximal u is not guaranteed to exist, so we can define with sup instead of max.

Example 2.1.18. Can T "stretch" different vectors differently, as the definition suggests? Yes: let $T(x) = x$, define $U = \mathbb{R}^2$ with ℓ_1 and $V = \mathbb{R}^2$ with ℓ_∞.

$$\|T\|_{op} = \max_{\mathbf{0} \neq x \in \mathbb{R}^2} \frac{\|x\|_\infty}{\|x\|_1} = \max_{\mathbf{0} \neq x \in \mathbb{R}^2} \frac{\max(|x_1|, |x_2|)}{|x_1| + |x_2|}$$

Clearly the denominator is $\ge$ the numerator, so $\|T\|_{op} \le 1$. For $x = (1, 0)$ the ratio is exactly 1, so $\|T\|_{op} = 1$. However, it can be strictly less than 1, for example with $x = (1, 1)$.

[4] This was shown in lecture 5.

Eigenvalues

Definition 2.1.19. The *right kernel* of a matrix M is the set of all vectors y such that $My = \mathbf{0}$. The *left kernel* is similarly defined with $yM = \mathbf{0}$. Notice that $y = \mathbf{0}$ always fits, so we will usually be interested in non-trivial solutions.

Definition 2.1.20. If a linear transformation T satisfies $T(x) = \lambda x$ for some scalar λ and vector x, we say that x is an *eigenvector* of T with *eigenvalue* λ. We require that $x \neq \mathbf{0}$ to avoid the trivial case.

Since there is a corresponding matrix A, finding eigenvectors and eigenvalues is equivalent to solving $Ax = \lambda x$, which is the same as $(\lambda I - A)x = \mathbf{0}$. So we are looking for non-trivial vectors in the right kernel of the matrix $\lambda I - A$.

A matrix has a non-trivial kernel iff it is singular (that is, determinant 0). So the eigenvalues are characterized by the roots of the n-degree polynomial

$$\det(\lambda I - A) = 0$$

(with λ as a variable). This is called the *characteristic polynomial* of A.

Definition 2.1.21. Matrices A, B are called *similar* if there exists an invertible matrix P such that $B = PAP^{-1}$ (equivalently, $A = P^{-1}BP$).

Definition 2.1.22. Diagonalization of a matrix A is the process of finding a diagonal matrix B which is similar to A.

Lecture 5
2020-11-17

The fundamental theorem of algebra states that every polynomial with degree n has n roots, counted with multiplicities. Generally these roots are complex numbers. So a characteristic polynomial $p_A(\lambda) = \det(\lambda I - A)$ can be decomposed as $\prod_{i=1}^{n}(\lambda - \alpha_i)$ with possibly some repeated α_i.

Definition 2.1.23. The algebraic multiplicity of an eigenvalue is its multiplicity in the characteristic polynomial.

Definition 2.1.24. The geometric multiplicity of an eigenvalue is the dimension of the subspace spanned by the corresponding eigenvectors.

The geometric multiplicity is at most the algebraic multiplicity, and we don't always have equality.

Example 2.1.25. Let $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Then $\lambda I - A = \begin{pmatrix} \lambda & -1 \\ 0 & \lambda \end{pmatrix}$ and $p_A(\lambda) = \lambda^2$. Therefore there is a single eigenvalue $\lambda = 0$ with algebraic multiplicity 2. The vectors in the eigenspace are those satisfying

$$\begin{aligned} Av &= \mathbf{0} \\ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \mathbf{0} \\ \begin{pmatrix} x_2 \\ 0 \end{pmatrix} &= \mathbf{0} \\ x_2 &= 0 \end{aligned}$$

Therefore the vectors are exactly those of the form $v = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$, hence the geometric multiplicity is 1.

A fact which is not difficult: the matrices with no multiplicities (that is, where every eigenvalue is simple) are dense in the set of all matrices. In other words, for every matrix A and $\epsilon > 0$ we can find a matrix B with no multiplicities such that $|a_{i,j} - b_{i,j}| < \epsilon$ for all i, j. This is related to the fact that we can change a polynomial "slightly" to remove repeated roots.

Definition 2.1.26. We distinguish between left eigenvectors and right eigenvectors. A right eigenvector x satisfies $Ax = \lambda x$, and a left eigenvector y satisfies $yA = \lambda y$. There is no such distinction for eigenvalues.

Notice that for every eigenvalue λ there is at least one left eigenvector and one right eigenvector, because $\lambda I - A$ is singular, so its kernel is non-trivial.

Proposition 2.1.27. *Similar matrices have the same eigenvalues, and in fact the same characteristic polynomial.*

Proof. Suppose $B = PAP^{-1}$. We have

$$\begin{aligned} |\lambda I - B| &= \left|\lambda I - PAP^{-1}\right| \\ &= \left|P(\lambda I - A)P^{-1}\right| \\ &= |P|\,|\lambda I - A|\left|P^{-1}\right| \\ &= |\lambda I - A| \end{aligned}$$

□

Generally the easiest way to find the eigenvalues of a matrix A is to diagonalize it: find a similar diagonal matrix D (if it exists) and Q such that $A = QDQ^{-1}$. The eigenvalues of a diagonal matrix are the elements on its main diagonal.

Observe that this is the same as $AQ = QD$. Let $q_1, \ldots, q_n$ be the columns of Q. Then

$$\begin{aligned} A(q_1 \mid \ldots \mid q_n) &= (q_1 \mid \ldots \mid q_n) \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \\ (Aq_1 \mid \ldots \mid Aq_n) &= (\lambda_1 q_1 \mid \ldots \mid \lambda_n q_n) \end{aligned}$$

Therefore the columns of Q are eigenvectors of A that correspond to its eigenvalues $\lambda_1, \ldots, \lambda_n$.

Lemma 2.1.28. *Let $v_1, \ldots, v_k$ be eigenvectors of A corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Then $v_1, \ldots, v_k$ are independent.*

Proof. Assume for contradiction that they are dependent, and consider a minimal linearly dependent subset $v_1, \ldots, v_\ell$ where $\ell \leq k$. There are $\alpha_i \neq 0$ such that

$$\sum_{i=1}^{\ell} \alpha_i v_i = \mathbf{0}$$

We have

$$\mathbf{0} = A\mathbf{0} = A\left(\sum_{i=1}^{\ell} \alpha_i v_i\right) = \sum_{i=1}^{\ell} \alpha_i \lambda_i v_i$$

and also $\lambda_\ell \sum_{i=1}^{\ell} \alpha_i v_i = \mathbf{0}$, and subtracting we get

$$\sum_{i=1}^{\ell-1} \alpha_i (\lambda_\ell - \lambda_i) v_i = \mathbf{0}$$

which contradicts the minimality of ℓ. □

Isometries

We want to understand rotation as a linear transformation. We define it to be a linear transformation that preserves the Euclidean norm. The fact that all distances and sizes are preserved captures our intuition of rotations (and reflections, which we include here).

Definition 2.1.29. Multiplying by a matrix A is a *rotation*, or an *isometry of* ℓ_2, if for all x we have $\|Ax\|_2 = \|x\|_2$.

Since $\|u\|_2^2 = \langle u, u\rangle$, it is a rotation iff $\langle Ax, Ax\rangle = \langle x, x\rangle$. Since generally $\langle Ax, By\rangle = \left\langle B^T Ax, y\right\rangle$, we can write this as $\langle x, x\rangle = \left\langle A^T Ax, x\right\rangle$.

Definition 2.1.30. A matrix A is called *orthogonal* if $A^T A = I$ (equivalently, $AA^T = I$), that is, $A^T = A^{-1}$.

Remark 2.1.31. If M is a matrix, not necessarily square, then MM^T is a square matrix where in i, j we have the inner product of M's i^{th} and j^{th} rows. Symmetrically the same can be said for $M^T M$ and columns.

Therefore, if $AA^T = I$ this means every two different rows (or columns) of A are orthogonal to each other, and every row and column is a unit vector.

Proposition 2.1.32. *If multiplying by A is a rotation then A is orthogonal.*

Proof. Suppose $\|Ax\|_2 = \|x\|_2$ for all x. Consider two vectors x, y. We have

$$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 + 2\langle x, y\rangle$$

On the other hand,

$$\begin{aligned}
\|x+y\|_2^2 &= \|A(x+y)\|_2^2 \\
&= \|Ax\|_2^2 + \|Ay\|_2^2 + 2\langle Ax, Ay\rangle \\
&= \|Ax\|_2^2 + \|Ay\|_2^2 + 2\left\langle A^T Ax, y\right\rangle \\
&= \|x\|_2^2 + \|y\|_2^2 + 2\left\langle A^T Ax, y\right\rangle
\end{aligned}$$

So for any x, y we have

$$\begin{aligned}
\left\langle A^T Ax, y\right\rangle &= \langle x, y\rangle \\
\left\langle (A^T A - I)x, y\right\rangle &= 0
\end{aligned}$$

For this to be true for any x, y we must have $A^T A - I = \mathbf{0}$. □

Theorem 2.1.33 (Spectral theorem for real symmetric matrices)**.** *Every real, symmetric matrix can be diagonalized with an orthogonal matrix. That is, there is an orthogonal matrix V and a diagonal matrix Λ such that*

$$A = V\Lambda V^T$$

The diagonal of Λ has the eigenvalues of A, and the columns of V are the corresponding eigenvectors of A.

More norms

Definition 2.1.34. For any vector u consider the map $x \mapsto \langle u, x \rangle$. We define a norm:

$$\|u\|^* = \max_{x \neq \mathbf{0}} \frac{|\langle u, x \rangle|}{\|x\|} = \max_{\|x\|=1} |\langle u, x \rangle|$$

It is called the *dual norm* of $\|\|$.

It is a norm: clearly it is non-negative, and 0 iff $u = \mathbf{0}$. We have $\|\alpha u\|^* = |\alpha| \, \|u\|^*$ since we can take α out of the expression $\langle \alpha u, x \rangle$. The triangle inequality also holds.

Proposition 2.1.35. *We have the following relations between $\ell_1, \ell_2, \ell_\infty$:*

$$\|\|_1^* = \|\|_\infty$$
$$\|\|_\infty^* = \|\|_1$$
$$\|\|_2^* = \|\|_2$$

Partial proof. For ℓ_1 we have:

$$\|u\|_1^* = \max_{\|x\|_1=1} \left| \sum_{i=1}^n u_i x_i \right|$$

To maximize the sum of $u_i x_i$, we would choose the signs of x_i to match u_i. Then, since $\|x\|_1 = 1$, the maximum is obtained when the largest u_i is given all the weight, resulting in $\max_{1 \leq i \leq n} |u_i|$ which is exactly $\|u\|_\infty$.

For ℓ_2 we have:

$$\|u\|_2^* = \max_{\sum_{i=1}^n x_i^2 = 1} |\langle x, u \rangle|$$

Again we choose the signs of x_i to match u_i. The best x is $\frac{u}{\|u\|_2}$ and we get $\langle x, u \rangle = \frac{\|u\|_2^2}{\|u\|_2} = \|u\|_2$. $\square$

Proposition 2.1.36. *For finite dimensional spaces we have $\|u\|^{**} = \|u\|$.*

Theorem 2.1.37. *Let $1 \leq p \leq \infty$, and q such that $\frac{1}{p} + \frac{1}{q} = 1$ (define $\frac{1}{p} = 0$ for $p = \infty$). Then L_p is the dual norm of L_q.*

Remark 2.1.38. Let $\|\|$ and $\|\|^*$ be dual norms. Then for all x, y,

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|^*$$

This follows directly from definition of $\|y\|^*$.

Theorem 2.1.39 (Hölder's inequality)**.** *Let p, q be as above, then*

$$\|x\|_p \, \|y\|_q \geq |\langle x, y \rangle|$$

Lecture 6
2020-11-24

We will prove this using the following auxiliary inequality:

Theorem 2.1.40 (Young's inequality)**.** *Let $1 \leq p, q \leq \infty$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then for any $a, b > 0$ we have:*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Proof. The logarithm function is concave, that is, every segment connecting two points on its plot is below the function. Formally, for all x_1, x_2 and $0 < \alpha < 1$,

$$\ln(\alpha x_1 + (1-\alpha)x_2) \geq \alpha \ln x_1 + (1-\alpha) \ln x_2$$

Read: logarithm of a point in (x_1, x_2) on the x axis is $\geq$ the corresponding point in the segment connecting $(x_1, \ln x_1)$ and $(x_2, \ln x_2)$. So we have:

$$\ln(ab) = \ln a + \ln b = \frac{1}{p} \ln a^p + \frac{1}{q} \ln b^q \leq \ln\left(\frac{a^p}{p} + \frac{b^q}{q}\right)$$

where the last inequality follows from concavity, with $\alpha = \frac{1}{p}$, $1-\alpha = \frac{1}{q}$, $x_1 = a^p$ and $x_2 = b^q$. $\square$

Proof of Hölder's inequality. We have:

$$\frac{|\langle x, y\rangle|}{\|x\|_p \|y\|_q} = \frac{|\sum_{i=1}^n x_i y_i|}{\|x\|_p \|y\|_q} \leq \sum_{i=1}^n \frac{|x_i| |y_i|}{\|x\|_p \|y\|_q}$$

Let $a = \frac{|x_i|}{\|x\|_p}$ and $b = \frac{|y_i|}{\|y\|_p}$ in Young's inequality:

$$\frac{|x_i| |y_i|}{\|x\|_p \|y\|_q} \leq \frac{|x_i|^p}{p \|x\|_p^p} + \frac{|y_i|^q}{q \|y\|_q^q}$$

So

$$\begin{aligned}
\frac{|\langle x, y\rangle|}{\|x\|_p \|y\|_q} &\leq \sum_{i=1}^n \left(\frac{|x_i|^p}{p \|x\|_p^p} + \frac{|y_i|^q}{q \|y\|_q^q}\right) \\
&= \frac{1}{p \|x\|_p^p} \sum_{i=1}^n |x_i|^p + \frac{1}{q \|y\|_q^q} \sum_{i=1}^n |y_i|^q \\
&= \frac{1}{p \|x\|_p^p} \|x\|_p^p + \frac{1}{q \|y\|_q^q} \|y\|_q^q \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1
\end{aligned}$$

$\square$

Definition 2.1.41. A matrix is called *positive definite* if all of its eigenvalues are positive, and *positive semidefinite* if all of its eigenvalues are non-negative. We sometimes write PSD_n for the set of positive semidefinite $n \times n$ matrices.

Theorem 2.1.42. *Let A be real and symmetric. The following are equivalent:*

1. *A is positive semidefinite.*
2. *$A = MM^T$ for some real matrix M.*
3. *The quadratic form xAx^T is non-negative for all x (note: this notation implies x is a row vector).*

Proof. We show each condition implies the next:

- $1 \Rightarrow 2$: By the spectral theorem, A can be written as $A = U\Lambda U^T$ where U is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Since $A \in PSD_n$ then the diagonal of Λ is non-negative. Let $\Lambda^{1/2} = \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ and define $M = U\Lambda^{1/2}$. Then we have

$$\begin{aligned} MM^T &= \left(U\Lambda^{1/2}\right)\left(U\Lambda^{1/2}\right)^T \\ &= U\Lambda^{1/2}\left(\Lambda^{1/2}\right)^T U^T \\ &= U\Lambda^{1/2}\Lambda^{1/2}U^T \\ &= U\Lambda U^T \\ &= A \end{aligned}$$

- $2 \Rightarrow 3$: Suppose $A = MM^T$ for some real matrix M. We have:

$$xAx^T = xMM^Tx^T = \langle xM, xM \rangle = \|xM\|^2 \geq 0$$

- $3 \Rightarrow 1$: If $xA = \lambda x$ then $xAx^T = \lambda xx^T = \lambda \|x\|^2$. Therefore $\lambda = \frac{\|x\|^2}{xAx^T} \geq 0$, because the norm is non-negative and the denominator is non-negative by assumption.

□

Remark 2.1.43. From the condition $xAx^T \geq 0$ we can verify that the set PSD_n is a *cone*: a set closed under addition and under multiplication by positive scalars.

2.2 Singular value decomposition (SVD)

Theorem 2.2.1 (SVD)**.** *Every $A \in M_{m\times n}(\mathbb{R})$ can be written as $A = UDV^T$ such that $U \in M_{m\times m}(\mathbb{R})$, $D \in M_{m\times n}(\mathbb{R})$, $V = M_{n\times n}(\mathbb{R})$, where U and V are orthogonal and D is "diagonal" except possibly some extra rows of 0s (if $m > n$) or extra columns of 0s (if $n < m$).*

We think of U as rotation, D as stretching, and V as rotation again. The fact that we use V^T instead of V is just a convention.

The values on the diagonal of D are called the *singular values* of A, and they are denoted $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$. There are m or n such values, whichever is smaller.

$$D = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{or} \quad D = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_m & 0 & 0 \end{pmatrix}$$

When ordered in decreasing order, the singular values of A are unique. The proof of the theorem will also give us $\sigma_1 = \|A\|_{op}$.

Proof. Write $D = U^TAV$. The proof is by induction on $n + m$. The plan is:

1. We find orthogonal U', V' such that U'^TAV' is not necessarily diagonal, but its first column is what we want, $(\sigma_1, 0, \ldots, 0)$.

2. We show that the first row is also what we want. The bottom right part of the matrix can still be anything.

3. By induction, we represent that part in SVD form, and use some block calculations to obtain an SVD form of A.

In full detail:

1. Let u_1, v_1 be unit vectors, such that $Av_1 = \sigma_1 u_1$ and σ_1 is the maximum possible. In other words, σ_1 is the maximal amount A stretches a unit vector, also known as $\|A\|_{op}$, and v_1, u_1 are an example of such stretching. Extend each of them to an orthonormal basis U', V' respectively:

$$U' = \left(\begin{array}{c|c} u_1 & U_2 \end{array}\right) \quad V' = \left(\begin{array}{c|c} v_1 & V_2 \end{array}\right)$$

where the first columns are u_1 and v_1 respectively. Then we have:

$$U'^T AV' = \begin{pmatrix} u_1^T \\ U_2^T \end{pmatrix} A \begin{pmatrix} v_1 & V_2 \end{pmatrix} = \begin{pmatrix} u_1^T \\ U_2^T \end{pmatrix} \begin{pmatrix} Av_1 & AV_2 \end{pmatrix} = \begin{pmatrix} u_1^T Av_1 & u_1^T AV_2 \\ U_2^T Av_1 & U_2^T AV_2 \end{pmatrix}$$

Notice that $u_1^T Av_1 = \sigma_1$ directly from definition, and $U_2^T Av_1 = U_2^T \sigma_1 u_1 = \mathbf{0}$ because u_1 is orthogonal to all the other vectors in that basis. Therefore:

$$U'^T AV' = \left(\begin{array}{c|c} \sigma_1 & w \\ \hline \mathbf{0} & B \end{array}\right)$$

for a row vector $w = u_1^T AV_2$ and matrix $B = U_2^T AV_2$. We have the correct first column.

2. Suppose $w \neq \mathbf{0}$ and consider the column vector (σ_1, w^T). We have:

$$U'^T AV' \begin{pmatrix} \sigma_1 \\ w^T \end{pmatrix} = \left(\begin{array}{c|c} \sigma_1 & w \\ \hline \mathbf{0} & B \end{array}\right) \begin{pmatrix} \sigma_1 \\ w^T \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \|w\|^2 \\ \vdots \end{pmatrix}$$

The length of the original vector is $\sqrt{\sigma_1^2 + \|w\|^2}$, and the length of the resulting vector is at least $\sigma_1^2 + \|w\|^2$. This means that $U'^T AV'$ stretches it by at least $\frac{\sigma_1^2 + \|w\|^2}{\sqrt{\sigma_1^2 + \|w\|^2}} = \sqrt{\sigma_1^2 + \|w\|^2} > \sigma_1$. The matrices U'^T and V' are orthogonal, so they only rotate without stretching. The stretching is done by A, which means there exists a vector it stretches by more than σ_1. This contradicts the maximality of σ_1.

3. By induction, there are orthogonal matrices U'', V'' such that $U''^T BV''$ is diagonal (up to extra rows or columns of 0s). So we have:

$$\left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & U''^T \end{array}\right) \left(\begin{array}{c|c} \sigma_1 & \mathbf{0} \\ \hline \mathbf{0} & B \end{array}\right) \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & V'' \end{array}\right) = \left(\begin{array}{c|c} \sigma_1 & \mathbf{0} \\ \hline \mathbf{0} & U''^T BV'' \end{array}\right)$$

which has the desired form. The product of orthogonal matrices is again orthogonal (it's a group), so we can set

$$U^T = \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & U''^T \end{array}\right) U'^T$$

$$V = V' \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & V'' \end{array}\right)$$

to get $A = UDV^T$.

□

Remark 2.2.2. As often happens, multiplying a matrix by its transpose gives interesting results. If $A = UDV^T$ then

$$AA^T = UDV^T(UDV^T)^T = UDV^TVD^TU^T = UDD^TU^T$$

Notice that DD^T is diagonal with values $\sigma_1^2, \sigma_2^2, \ldots$ and possibly some 0s. Multiplying by U from the left and by $U^T = U^{-1}$ from the right is a change of basis, which does not affect eigenvalues. So the eigenvalues of AA^T are the squares of the singular values of A.

Motivation and usage

The fact that any real matrix can be viewed as rotate-stretch-rotate is very useful in applied mathematics, especially in cleaning up noisy data. Often the larger singular values come from the "signal", and there is a tail of smaller ones which are probably noise. That is, we assume there is a clean ideal matrix we want to find, which has relatively low rank because it's nice and tidy (note that lower rank means fewer positive singular values). When we measure a high-rank approximation of it, some of the singular values that "should" be 0 appear as small positive numbers instead.

Useful definition for distance between matrices:

Definition 2.2.3. The Frobenius norm of a matrix $A \in M_{n\times m}(\mathbb{R})$ is defined like ℓ_2 for vectors:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2}$$

Notice that we have $\|A\|_F^2 = \mathrm{tr}(AA^T)$.

Proposition 2.2.4. *The Frobenius form is invariant under multiplication by orthogonal matrices. That is, if U, V are orthogonal, then $\|A\|_F = \|AU\|_F = \|VA\|_F$.*

Proof. We have:

$$\|AU\|_F^2 = \mathrm{tr}((AU)(AU)^T) = \mathrm{tr}(AUU^TA^T) = \mathrm{tr}(AA^T) = \|A\|_F^2$$
$$\|VA\|_F^2 = \mathrm{tr}((VA)(VA)^T) = \mathrm{tr}(VAA^TV^T) = \mathrm{tr}(V^TVAA^T) = \mathrm{tr}(AA^T) = \|A\|_F^2$$

In the second row we used the fact that $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ (careful: we can't reorder the product however we like. Generally we are only allowed cyclic shifts). □

Problem. Given some matrix A and a natural number k, what is the closest matrix B with $\operatorname{rank}(B) \leq k$? We can ask this with different notions of distance between A and B, such as operator norm or Frobenius norm.

Some years ago, Netflix published a challenge to help build its recommendation system. There is a database that can be represented as a large matrix, with one row per person and one column per movie, where the numbers are ratings that people gave movies they watched. Most of the elements in this matrix are missing, because most people haven't watched most movies. What movies should be recommended to a given user? This is a hard problem. Our goal is not to present a solution here, but to show the relevance of SVD.

Movies can be grouped into a relatively small number of categories like action, drama, etc. Consider two smaller matrices:

1. A matrix with a row per person and a column per category, measuring how much a person likes a category.
2. A matrix with a row per category and a column per movie, measuring how fitting a movie is to a category.

This is a rough simplification, but we can guess that the original matrix is the product of these two matrices (we are assuming here that a person who loves action is likely to enjoy all action movies, which is not ideal, but it's a start).

Note that this guess means the rank of the original matrix is small: it is bounded by the number of categories, because the rank of a product is bounded by the rank of each factor, thus motivating the above problem.

Theorem 2.2.5. *Let $A = M_{m\times n}(\mathbb{R})$ and $k \in \mathbb{N}$. Let $A = UDV^T$ be its SVD form, and let*

$$B = UD^{(k)}V^T$$

where $D^{(k)}$ is D with only the first k elements on the diagonal, that is, all of $\sigma_{k+1}, \sigma_{k+2}, \ldots$ replaced with 0s. Then B is a solution to both of these minimization problems:

$$\min_{\operatorname{rank}(B)\leq k} \|A - B\|_{op}$$
$$\min_{\operatorname{rank}(B)\leq k} \|A - B\|_{F}$$

Lecture 7
2020-12-01

Proof. For the chosen B, notice that $A - B = U(D - D^{(k)})V^T$, so the maximal stretching that $A - B$ can do is $\|A - B\|_{op} = \sigma_{k+1}$. We will show that this is the best possible, that is, for any matrix C with $\operatorname{rank}(C) \leq k$ we have $\|A - C\|_{op} \geq \sigma_{k+1}$.

Our goal then is to find a unit vector z such that $\|(A - C)z\| \geq \sigma_{k+1}$. Because the only thing we know about C is $\operatorname{rank}(C) \leq k$, it will be convenient to choose $z \in \ker(C)$, so that $(A - C)z = Az$. The kernel is big: we have $\dim\ker(C) \geq n - k$, so it has a non-trivial intersection with any $(k+1)$-dimensional subspace. In particular, there is a non-trivial vector in $\ker(C)$ which is also in

$$\operatorname{span}(v_1, \ldots, v_{k+1})$$

where v_i are the rows of V. Let z be such a vector after normalization. Write it as:

$$z = \sum_{i=1}^{k+1} \alpha_i v_i$$

We have:

$$A = \sum_{i=1}^{\min(n,m)} \sigma_i u_i \otimes v_i$$

where u_i are the columns of U. Hence:

$$Az = \sum_{i=1}^{\min(n,m)} \sigma_i u_i \otimes v_i z = \sum_{i=1}^{k+1} \sigma_i u_i \langle v_i, z \rangle$$

Notice that u_i are an orthonormal basis, so the norm is:

$$\|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 \langle v_i, z \rangle^2 \geq \sum_{i=1}^{k+1} \sigma_{k+1}^2 \langle v_i, z \rangle^2 = \sigma_{k+1}^2 \sum_{i=1}^{k+1} \langle v_i, z \rangle^2 = \sigma_{k+1}^2$$

□

2.3 Variational characterization of eigenvalues

Theorem 2.3.1 (Rayleigh-Ritz)**.** *Let A be a real symmetric matrix with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. Then:*

$$\lambda_1 = \max_{x \neq \mathbf{0}} \frac{xAx^T}{\|x\|^2} = \max_{\|x\|_2 = 1} xAx^T$$

The expression $\frac{xAx^T}{\|x\|^2}$ is called the Rayleigh quotient of A and x. Choose some x_1 which maximizes it. Then:

$$\lambda_2 = \max_{x \perp x_1} \frac{xAx^T}{\|x\|^2}$$

Again choose some $x_2 \perp x_1$ which maximizes this expression. Generally for $0 \leq k < n$,

$$\lambda_{k+1} = \max_{x \perp x_1, \ldots, x_k} \frac{xAx^T}{\|x\|^2}$$

where $x_1, \ldots, x_k$ are the choices we made along the way.

On every step, we maximize with some vector, obtain an eigenvalue, then limit ourselves to the subspace which is orthogonal to everything so far. Similarly, we can start with $\lambda_n = \min \frac{xAx^T}{\|x\|^2}$ and do it all in the other direction.

Proof. We show that the maximal Rayleigh quotient of step k is bounded from above and below by λ_{k+1}. The lower bound is immediate: take x^T to be an eigenvector corresponding to λ_{k+1}.

$$\begin{aligned} Ax^T &= \lambda_{k+1} x^T \\ xAx^T &= \lambda_{k+1} xx^T \\ xAx^T &= \lambda_{k+1} \|x\|^2 \end{aligned}$$

Upper bound: since A is real and symmetric, we can write $A = V \Lambda V^T$ where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and V's columns are an orthonormal basis of eigenvectors. Then we have:

$$\begin{aligned} xAx^T = xV\Lambda V^T x^T = (xV)\Lambda(xV)^T &= \sum_{i=1}^{n} \lambda_i (xV)_i^2 \\ &\leq \sum_{i=1}^{n} \lambda_1 (xV)_i^2 = \lambda_1 \sum_{i=1}^{n} (xV)_i^2 = \lambda_1 \|x\|^2 \end{aligned}$$

because V is orthogonal. Generally, if $x \perp v_1, \ldots, v_k$ then

$$xAx^T = \sum_{i=1}^{n} \lambda_i (xV)_i^2 = \sum_{i=k+1}^{n} \lambda_i (xV)_i^2 \leq \lambda_{k+1} \sum_{i=k+1}^{n} (xV)_i^2 = \lambda_{k+1} \|x\|^2$$

□

Alternative proof for λ_1. We show $\lambda_1 = \max \frac{yAy^T}{\|y\|^2}$ by analyzing the Rayleigh quotient as a differentiable function. Consider the partial derivative for the ith coordinate in the denominator:

$$\frac{\partial}{\partial y_i} \|y\|^2 = \frac{\partial}{\partial y_i} \sum_{j=1}^{n} y_j^2 = 2y_i$$

and in the numerator:

$$\frac{\partial}{\partial y_i} yAy^T = \frac{\partial}{\partial y_i} \sum_{1 \leq j,k \leq n} A_{j,k} y_j y_k = 2 \sum_{k=1}^{n} A_{i,k} y_k = 2(Ay)_i$$

Generally the derivative of a quotient is $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$ which is 0 when $f'g = fg'$. So we want for all i:

$$2(Ay)_i \|y\|^2 = 2y_i y A^T y^T$$
$$(Ay)_i = \frac{yA^T}{\|y\|^2} y_i$$

Requiring this for all i is equivalent to $Ay = \frac{yAy^T}{\|y\|^2} y$, so y is an eigenvector with eigenvalue $\frac{yAy^T}{\|y\|^2}$. □

Theorem 2.3.2 (Courant–Fischer)**.** *Let A be real and symmetric with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. Then for all $0 \leq i < n$:*

$$\lambda_{i+1} = \min_{\dim F = i} \max_{x \perp F} \frac{xAx^T}{\|x\|^2}$$

that is, we minimize over subspaces F with dimension i, and maximize over vectors orthogonal to F. Similarly,

$$\lambda_{i+1} = \max_{\dim G = n-i-1} \min_{x \perp G} \frac{xAx^T}{\|x\|^2}$$

Corollary 2.3.3 (Interlacing theorem)**.** *Let A be real and symmetric with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$, and let B be a matrix obtained from A by removing the ith row and column, for some i. Let $\mu_1 \geq \ldots \geq \mu_{n-1}$ be the eigenvalues of B. Then we have:*

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \ldots \geq \mu_{n-1} \geq \lambda_n$$

2.4 Perron-Frobenius

Lecture 7/8
2020-12-01/08

In this section[5] we study matrices in which all entries are positive, matrices in which all entries are non-negative, and the relation between them. This is motivated by the next section about Markov chains, where the matrices will contain probabilities.

[5] I took more liberty than usual with reordering this material, hopefully in a useful way.

Positive matrices

Definition 2.4.1. The *spectral radius* of a square matrix is the largest absolute value of its eigenvalues. Note that the eigenvalues may be complex, and we define absolute value in the usual way (distance from the origin).

Theorem 2.4.2 (Perron)**.** *Let $A \in M_{n\times n}(\mathbb{R})$ be a positive matrix ($A_{i,j} > 0$ for all i, j). Then A has a real, positive eigenvalue ρ such that:*

1. *The eigenvector v corresponding to ρ is positive.*
2. *ρ is the spectral radius: for every other eigenvalue $\mu \neq \rho$, we have $|\mu| < \rho$.*
3. *ρ is simple: its algebraic and geometric multiplicity is 1.*
4. *v is the only positive eigenvector (up to multiplication by a constant).*

Proof. We follow a proof by Wielandt[6]. First notice that for any vector $x \neq \mathbf{0}$, if $x \geq \mathbf{0}$ (that is, each coordinate in x is non-negative) then $Ax > \mathbf{0}$. This holds because for every row i we have

$$(Ax)_i = \sum_{j=1}^{n} A_{i,j}x_j$$

All $A_{i,j}$ are positive, the x_j are non-negative, and at least one x_j is positive.

We want to find a maximal positive eigenvalue and corresponding eigenvector. Consider a set S of all vectors $x \geq \mathbf{0}$ with $\|x\|_2 = 1$ (imagine the top-right quarter circle, or octant of a sphere). Define:

$$L(x) = \min_{\substack{1\leq i\leq n \\ x_i\neq 0}} \frac{(Ax)_i}{x_i}$$

Intuition: if x is an eigenvector then all ratios $\frac{(Ax)_i}{x_i}$ are the same, and equal to its eigenvalue. In L we don't require x to be an eigenvector, but we consider that ratio anyway, and $L(x)$ tells us the "worst" ratio in x. The idea is that if the ratios are not all equal, we can make $L(x)$ larger by changing x slightly, raising the minimal ratio in expense of other coordinates. If $L(x)$ can't be made larger, then all ratios are the same, and we have an eigenvector.

Formally, the set S is compact and L is continuous, so there exists a vector $v \in S$ such that $L(v)$ is maximal. We show that v is an eigenvector, and that $L(v) = \rho$ is an eigenvalue.

ρ is the minimal ratio of $\frac{(Av)_i}{v_i}$, so $Av \geq \rho v$ (per coordinate), equivalently $Av - \rho v \geq \mathbf{0}$. Suppose for contradiction that $Av \neq \rho v$. Then $Av - \rho v$ is a non-zero, non-negative vector, hence A sends it to a positive vector:

$$\begin{aligned} A(Av - \rho v) &> \mathbf{0} \\ A(Av) &> \rho(Av) \end{aligned}$$

Therefore[7] we have $\frac{(A(Av))_i}{(Av)_i} > \rho$ for all i, so Av has a higher minimal ratio. We normalize it and get

$$L\left(\frac{Av}{\|Av\|_2}\right) > \rho$$

which contradicts the maximality of ρ.

Now that we have established ρ and v, we show they satisfy the claims.

[6] http://www.maths.nuigalway.ie/~rquinlan/linearalgebra/lecture2526.pdf
[7] The paper introduces $\epsilon > 0$ here, but the proof seems to work without it.

1. A sends non-zero non-negative vectors to positive vectors, so we have $\rho v = Av > \mathbf{0}$. Thus $\rho v_i > 0$ for all i, so $v_i > 0$ as well. Therefore v is positive.

2. Let μ be any eigenvalue of A, with corresponding unit eigenvector y. From $Ay = \mu y$ we have

$$(Ay)_i = \mu y_i$$
$$\sum_{j=1}^{n} A_{i,j} y_j = \mu y_i$$

So by the triangle inequality,

$$|\mu| \, |y_i| \leq \sum_{j=1}^{n} A_{i,j} \, |y_j|$$

Let $|y| = (|y_1|, \ldots, |y_n|)$, and notice that $|y| \in S$. The inequalities for all i can be summarized as $|\mu| \, |y| \leq A \, |y|$, which means $|\mu| \leq L(|y|) \leq \rho$ because ρ is the maximal L.

3. We will show the geometric multiplicity is 1 (the proof for algebraic multiplicity is in the paper). Suppose for contradiction that apart from v there is another eigenvector u with eigenvalue ρ, and that they are linearly independent. Assume without loss of generality that u is real (if it is complex, we need to separate to real and imaginary components).

We can choose $\epsilon \neq 0$ such that $v + \epsilon u$ is non-negative, and has at least one 0 coordinate. We assume u, v are independent, so $v + \epsilon u \neq \mathbf{0}$. This vector is in the eigenspace of ρ, so

$$A(v + \epsilon u) = \rho(v + \epsilon u)$$

But on the left we have a strictly positive vector, and on the right there is at least one 0, contradiction.

4. We omit this part (it is shown in the paper).

□

Non-negative matrices

The main application of this section will be for matrices of probabilities, so we are interested in non-negative matrices as well. The idea is to try and "reduce" the non-negative case to the positive case.

Raising a matrix A to a power A^k preserves all eigenvectors of A (if $Au = \lambda u$ then $A^k u = \lambda^k u$). So if a non-negative matrix A happens to have $A^k > \mathbf{0}$ then we can hope to learn something about it with Perron's theorem.

We only care whether an entry is 0 or not, so we can replace each positive entry in A with 1. Define a matrix M:

$$M_{i,j} = \begin{cases} 1 & A_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

We have $M^k > \mathbf{0}$ iff $A^k > \mathbf{0}$. The advantage is that M is the adjacency matrix of a directed graph (the edge $i \to j$ exists iff $M_{i,j} = 1$). That is, we converted the question of whether $A^k > \mathbf{0}$ into a combinatorial problem.

Recall that $M^k_{i,j}$ is the number of paths from i to j of length k. If some j is not reachable from some i, then $M^k_{i,j}$ will be 0 for all k. This leads us to define:

Definition 2.4.3. A directed graph is *strongly connected* if there exists a path from every vertex to every vertex.

So for $M^k > \mathbf{0}$ to be possible, the graph must be strongly connected. This is necessary but not sufficient: another obstacle is that the graph may have a cyclical structure that creates a problematic pair of vertices for every k. As a simple example, take a cycle of length 2:

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

We cannot get rid of the 0s by raising to a power. More generally, we say a graph has a cyclical structure if the vertices can be partitioned into $r > 1$ sets $S_0, \ldots, S_{r-1}$ such that every outgoing edge from S_i goes to $S_{(i+1) \bmod r}$.

Here we need the following concept, though we will not prove its relevance:

Definition 2.4.4. Let x be a vertex in a directed graph, and suppose x is on a cycle (there is a non-trivial path from x to x). We say that x is *aperiodic* if the greatest common divisor of all path lengths from x to x is 1. We say that a graph is aperiodic if all its vertices are aperiodic.

Theorem 2.4.5 (Frobenius)**.** *Let $A \in M_{n\times n}(\mathbb{R})$ be non-negative. If the corresponding directed graph is strongly connected and aperiodic, then Perron's theorem applies to A.*

Theorem 2.4.6. *Let A be a matrix for which Perron's theorem is applicable (positive, or alternatively non-negative and strongly connected and aperiodic). Let x_ℓ, x_r be the left and right positive eigenvectors respectively, with eigenvalue ρ. Then:*

$$\lim_{t\to\infty} \left(\frac{A}{\rho}\right)^t = \frac{x_r \otimes x_\ell}{\langle x_r, x_\ell \rangle}$$

2.5 Markov chains

A Markov chain is a system of n states $1, \ldots, n$. At any given time the system is in one of the states, and every moment it transitions from the current state to a possibly different state. The transition is probabilistic: for all $1 \le i, j \le n$, the system has some fixed probability $P_{i,j}$ to transition from i to j. The matrix P is called the *transition matrix* of the Markov chain. Since the values $P_{i,1}, \ldots, P_{i,n}$ form a distribution, each row in the matrix P sums to 1 and all entries are non-negative.

Definition 2.5.1. A matrix P is called *stochastic* if its elements are non-negative and the sum of every row is 1.

Let X_t be the state of the system at time t. The transition rule can be defined as:

$$\Pr\left[X_{t+1} = j \mid X_t = i\right] = P_{i,j}$$

Note that a transition only depends on the previous state and not on the history. Therefore a Markov chain is *memoryless* stochastic process.

Suppose the probability that the system is in state i is x_i, for $1 \le i \le n$. The vector $x = (x_1, \ldots, x_n)$ is a distribution vector, that is, $x_i \ge 0$ and $\sum_{i=1}^n x_i = 1$. We want to know the distribution vector $y = (y_1, \ldots, y_n)$ at the next moment, after the system completes one transition. The probability to get to state j is the probability that the previous state was i and that the transition from i to j took place, for any i. Hence:

$$y_j = \sum_{i=1}^n P_{i,j} x_i$$

By definition of matrix multiplication, we have $y = xP$ (we think of the distributions as row vectors), so advancing by one step is done by multiplying by P. Advancing t steps is done by successive multiplications, xP^t. There are some natural questions we can study here, like the long term behavior of the chain (what happens when $t \to \infty$) or the existence of distributions that are not affected by transition.

Definition 2.5.2. Given a stochastic matrix P, a distribution is called *stationary* and denoted π if we have $\pi P = \pi$. It is a left eigenvector of P with eigenvalue 1.

It is not obvious when to expect such a distribution to exist. The theorem below shows that for some Markov chains, there is a unique stationary distribution and the process approaches it regardless of the initial distribution. We use Perron-Frobenius, which requires further assumptions.

Definition 2.5.3. A Markov chain is called *irreducible* if the directed graph corresponding to its transition matrix (in the sense of Perron-Frobenius) is strongly connected.

Definition 2.5.4. A Markov chain is called *aperiodic* if the directed graph corresponding to its transition matrix is aperiodic.

Definition 2.5.5. A Markov chain with matrix P is called *ergodic* if it is both irreducible and aperiodic.

Lecture 9
2020-12-15

Theorem 2.5.6. *Let P be the transition matrix of an ergodic Markov chain. Then:*

$$\lim_{t\to\infty} P^t = \begin{pmatrix} \pi \\ \hline \vdots \\ \hline \pi \end{pmatrix}$$

where π is a distribution row vector such that $\pi P = \pi$.

In particular, $xP^t \to \pi$ for any initial distribution x.

Proof. We have $P \cdot \mathbf{1} = \mathbf{1}$, so 1 is an eigenvalue and $\mathbf{1}$ is a right eigenvector. There is a corresponding left eigenvector π with eigenvalue 1, such that π is a distribution[8]. Let Z be the target matrix where every row is π, and let k be such that $P^k > \mathbf{0}$. Such k exists by Perron-Frobenius. We want to show that P^ℓ and Z get arbitrarily close as ℓ grows. The plan is:

1. We show that the sequence $\left\|P^\ell - Z\right\|_\infty$ is non-increasing.
2. We show that its subsequence $\left\|P^{\ell k} - Z\right\|_\infty$ (where we only consider powers that are multiples of k) approaches 0.

From these we can conclude that $\left\|P^\ell - Z\right\|_\infty \to 0$. In more detail:

1. First, observe that $ZP = PZ = Z$: we have $ZP = Z$ because for each row $\pi P = \pi$, and $PZ = Z$ because every column in Z is constant and $P \cdot \mathbf{1} = \mathbf{1}$. Now we write the matrix $P^{\ell+1} - Z$ as follows:
$$P^{\ell+1} - Z = P(P^\ell - Z)$$
That is, every element of $P^{\ell+1} - Z$ is a convex combination of elements in $P^\ell - Z$, with non-negative coefficients that sum to 1. This cannot increase the largest absolute value.

[8] We did not prove it in class.

2. Let $c > 0$ be maximal such that $P^k \geq cZ$ (such c exists because $P^k > \mathbf{0}$). Both P^k and Z are stochastic, so we have $c \leq 1$, otherwise the row sums in P^k will be too large. If $c = 1$, then they must be equal, so P^k is already Z without taking limits. Multiplying by P more times will not change the result, since $ZP = Z$.

   Suppose then that $0 < c < 1$, and define the matrix:

$$N = \frac{1}{1-c}\left(P^k - cZ\right)$$

   N is stochastic: from $P^k \geq cZ$ we get $N \geq \mathbf{0}$, and summing rows we get:

$$N \cdot \mathbf{1} = \frac{1}{1-c}\left(P^k - cZ\right) \cdot \mathbf{1} = \frac{1}{1-c}\left(\mathbf{1} - c\mathbf{1}\right) = \mathbf{1}$$

   because $P \cdot \mathbf{1} = \mathbf{1}$ and $Z \cdot \mathbf{1} = \mathbf{1}$. Also, we have $NZ = ZN = Z$:

$$NZ = \frac{1}{1-c}\left(P^k - cZ\right)Z = \frac{P^kZ - cZ^2}{1-c} = \frac{Z - cZ}{1-c} = Z$$
$$ZN = \frac{1}{1-c}Z\left(P^k - cZ\right) = \frac{ZP^k - cZ^2}{1-c} = \frac{Z - cZ}{1-c} = Z$$

   where we used the fact that $Z^2 = Z$ and $P^kZ = ZP^k = Z$.

   Using this identity we can easily prove by induction each of the following identities:

$$(N - Z)^\ell = N^\ell - Z$$
$$P^{k\ell} - Z = (P^k - Z)^\ell$$

   Now rearrange the definition of N as follows:

$$P^k - Z = (1-c)(N - Z)$$

   and we have:

$$P^{k\ell} - Z = \left(P^k - Z\right)^\ell = (1-c)^\ell (N - Z)^\ell = (1-c)^\ell (N^\ell - Z)$$

   The elements of $N^\ell - Z$ are between 1 and -1, since N^ℓ and Z are both stochastic. Hence:

$$\left\|P^{k\ell} - Z\right\|_\infty \leq (1-c)^\ell \to 0$$

□

Random walks

An important example of a Markov chain is a simple random walk (SRW) on a graph: every transition is from a vertex to one of its neighbors, chosen uniformly.

Formally, let G be connected and non-bipartite. We define a Markov chain with V as the set of states, and the transition from vertex i to j has probability:

$$P_{i,j} = \begin{cases} \frac{1}{\deg i} & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

The stationary distribution π is proportional to the degrees:

$$\pi_i = \frac{\deg i}{\sum_{v \in V} \deg v} = \frac{\deg i}{2\,|E|}$$

2.6 Expander graphs

Informally, expanders are graphs without "bottlenecks". For example, if a connected graph is made of two large sets of vertices that are only connected by a few edges, then those edges form a bottleneck, and the graph is not a good expander. The more significant the bottleneck, the worse the graph is as an expander. A good expander is difficult to break apart into large pieces. This has many applications, like robustness of computer networks.

To quantify bottlenecks, we want to take the sizes of the sets into account. If one of the sets is very small, the bottleneck is not as significant (this is different from the min-cut problem, where we only care about the minimal number of edges needed to disconnect the graph).

Formally, let $S \subseteq V$ be a non-trivial subset of vertices ($S \neq \emptyset, V$). Denote its complement by $\overline{S} = V \backslash S$. Let $e(S, \overline{S})$ be the number of edges that connect them. If $e(S, \overline{S})$ is small compared to $|S|$ or $|\overline{S}|$, the bottleneck is significant. We usually assume without loss of generality that $|S| \leq \frac{|V|}{2}$ and $|\overline{S}| \geq \frac{|V|}{2}$.

Definition 2.6.1. We say that a graph has *edge expansion* δ if for every $|S| \leq \frac{|V|}{2}$ we have:

$$e(S, \overline{S}) \geq \delta \, |S|$$

Definition 2.6.2. The *Cheeger constant* of a graph G is defined as:

$$h(G) = \min_{0 < |S| \leq \frac{|V|}{2}} \frac{e(S, \overline{S})}{|S|}$$

Remark 2.6.3. Isoperimetric problems are questions of the form "what shape has minimal perimeter-to-area ratio?". In higher dimensions we can think of surface-area-to-volume. In Euclidean geometry the answer is an n-dimensional ball. Here, we can think of $e(S, \overline{S})$ and $|S|$ as the graph theory variants of surface area and volume of S, respectively.

There is a relationship between Cheeger's constant and eigenvalues:

Theorem 2.6.4 (Cheeger's inequalities)**.** *Let G be d-regular and let $d = \lambda_1 \geq \ldots \geq \lambda_n$ be the spectrum of the adjacency matrix A. Then:*

$$\frac{d - \lambda_2}{2} \leq h(G) \leq \sqrt{d^2 - \lambda_2^2}$$

Partial proof. We will show the lower bound. Rearrange it as:

$$d - 2h(G) \leq \lambda_2$$

By the variational characterization we have:

$$\lambda_2 = \max_{x \perp \mathbf{1}} \frac{x A x^T}{\|x\|^2}$$

where we maximize over all x such that $\langle x, \mathbf{1} \rangle = 0$, that is, $\sum_{i=1}^{n} x_i = 0$. Choosing any specific x will give us a lower bound on λ_2. Let $S \subseteq V$ and define x as:

$$x_i = \begin{cases} |\overline{S}| & i \in S \\ -|S| & i \notin S \end{cases}$$

There are $|S|$ elements equal to $\left|\overline{S}\right|$, and $\left|\overline{S}\right|$ elements equal to $-|S|$, so the overall sum of x is $|S|\left|\overline{S}\right| - \left|\overline{S}\right||S| = 0$. We calculate the Rayleigh quotient explicitly. The denominator is:

$$\begin{aligned}
\|x\|_2^2 &= \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i\in S} x_i^2 + \sum_{i\notin S} x_i^2 \\
&= \sum_{i\in S} \left|\overline{S}\right|^2 + \sum_{i\notin S} |S|^2 \\
&= |S|\left|\overline{S}\right|^2 + \left|\overline{S}\right||S|^2 \\
&= |S|\left|\overline{S}\right|\left(\left|\overline{S}\right| + |S|\right) \\
&= n\,|S|\left|\overline{S}\right|
\end{aligned}$$

The numerator is:

$$xAx^T = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j A_{i,j} = \sum_{i\sim j} x_i x_j$$

where the last sum is over all $1 \le i, j \le n$ such that i and j are neighbors (there $A_{i,j} = 1$, and all other terms vanish). We split this sum into 3 parts:

- $i, j \in S$: We have $x_i = x_j = \left|\overline{S}\right|$ and there are $2e(S)$ such pairs, contributing $2e(S)\left|\overline{S}\right|^2$ to the sum.
- $i, j \notin S$: We have $x_i = x_j = -|S|$ and there are $2e(\overline{S})$ such pairs, contributing $2e(\overline{S})|S|^2$ to the sum.
- $i \in S, j \notin S$ or vice versa: We have $x_i x_j = -|S|\left|\overline{S}\right|$ and there are $2e(S,\overline{S})$ such pairs, contributing $-2e(S,\overline{S})|S|\left|\overline{S}\right|$ to the sum.

Overall,

$$\lambda_2 \ge \frac{2e(S)\left|\overline{S}\right|^2 + 2e(\overline{S})|S|^2 - 2e(S,\overline{S})|S|\left|\overline{S}\right|}{n\,|S|\left|\overline{S}\right|}$$

The term $e(S,\overline{S})$ is desirable, but the others are not, so we express them in a different way. If we sum the degrees in S, we get $d\,|S|$. This counts every edge in S twice, and every edge from S to $\overline{S}$ once. A similar argument can be made for $\overline{S}$, and we get:

$$\begin{aligned}
d\,|S| &= 2e(S) + e(S,\overline{S}) \\
d\left|\overline{S}\right| &= 2e(\overline{S}) + e(S,\overline{S})
\end{aligned}$$

Substituting,

$$\lambda_2 \ge \frac{\left(d\,|S| - e(S,\overline{S})\right)\left|\overline{S}\right|^2 + \left(d\left|\overline{S}\right| - e(S,\overline{S})\right)|S|^2 - 2e(S,\overline{S})|S|\left|\overline{S}\right|}{n\,|S|\left|\overline{S}\right|}$$

In this expression, the coefficient of d is

$$\frac{|S|\left|\overline{S}\right|^2 + \left|\overline{S}\right||S|^2}{n\,|S|\left|\overline{S}\right|} = \frac{|S|\left|\overline{S}\right|\left(\left|\overline{S}\right| + |S|\right)}{n\,|S|\left|\overline{S}\right|} = \frac{|S|\left|\overline{S}\right|n}{n\,|S|\left|\overline{S}\right|} = 1$$

and the coefficient of $e(S,\overline{S})$ is

$$\frac{-\left|\overline{S}\right|^2-|S|^2-2\,|S|\,\left|\overline{S}\right|}{n\,|S|\,\left|\overline{S}\right|}=\frac{-\left(\left|\overline{S}\right|+|S|\right)^2}{n\,|S|\,\left|\overline{S}\right|}=\frac{-n^2}{n\,|S|\,\left|\overline{S}\right|}=-\frac{n}{|S|\,\left|\overline{S}\right|}$$

So we can simplify:

$$\lambda_2\geq d-\frac{n\cdot e(S,\overline{S})}{|S|\,\left|\overline{S}\right|}$$

This is true for any $S\neq\emptyset,V$. Assuming $|S|\leq n/2$, we have $\left|\overline{S}\right|\geq\frac{n}{2}$ so $\frac{n}{\left|\overline{S}\right|}\leq 2$, hence

$$\lambda_2\geq d-\frac{n\cdot e(S,\overline{S})}{|S|\,\left|\overline{S}\right|}=d-\frac{2e(S,\overline{S})}{|S|}$$

In particular we can choose a set S that maximizes this expression, or equivalently, minimizes $\frac{e(S,\overline{S})}{|S|}$. For such S, the ratio $\frac{e(S,\overline{S})}{|S|}$ is Cheeger's constant, and we get

$$\lambda_2\geq d-2h(G)$$

□

Remark 2.6.5. Calculating $h(G)$ is co-NP-hard. This theorem gives us an approximation using eigenvalues, which can be calculated more easily.

Rapid mixing

An important property of expander graphs is that an SRW converges quickly to the stationary distribution.

Suppose G is connected, d-regular and non-bipartite. The stationary distribution of an SRW starting at some vertex v is the uniform distribution. So as the time t grows, the probability to be in any given vertex gets closer to $\frac{1}{n}$. But early in the process, there may be some vertices that are too far away from v, and their probability is still 0. This gives us a lower bound on the rate of convergence.

Definition 2.6.6. The *distance* between two vertices u,v is the length of a shortest path between them.

Definition 2.6.7. The *diameter* of a graph G, denoted $\mathrm{diam}(G)$, is the maximal distance between two vertices.

If $t<\mathrm{diam}(G)$ then some vertices may not have been reached yet. Only when $t\geq\mathrm{diam}(G)$ we are guaranteed that each vertex has positive probability. We can bound this from below:

Suppose we perform BFS from v. After one step, we reach d neighbors. After each subsequent step, each neighbor can reach at most $d-1$ new vertices (perhaps fewer, if some are not new). Generally, the tth step can reveal about $(d-1)^t$ new vertices. When t is $\mathrm{diam}(G)$ we must have reached the entire graph[9]:

$$\begin{aligned}(d-1)^{\mathrm{diam}(G)}&\geq n\\ \mathrm{diam}(G)&\geq\log_{d-1}n\end{aligned}$$

There exist expander graphs that achieve this lower bound of convergence rate.

[9]This is not a complete calculation. Since $(d-1)^t$ grows exponentially, the sum is dominated by the largest term, and we ignore the rest. The inequalities may need some tweaking to account for all the details.

3 Optimization

3.1 Basics

There are many contexts in which we look for the "best" solution: a shortest path, a minimal spanning tree, and so on. In the most general sense, an optimization problem is defined by a domain D and a function $f : D \to \mathbb{R}$, and we are looking for a variable $x \in D$ that maximizes f (or minimizes f, depending on context). We call f the *objective function*.

Linear programming (LP) is a special case of an optimization problem: we have $D \subseteq \mathbb{R}^n$, and D is defined as all $x \in \mathbb{R}^n$ that satisfy some linear equations and inequalities, and f is linear.

Such a set D is called a polyhedron, and a bounded polyhedron is called a polytope. Throughout, we will go back and forth between geometry and algebra for intuition and formalization.

Remark 3.1.1. Polyhedra may have lower dimension than the space they are in. For example, a 2D hexagon in 3D space is a polytope.

Example 3.1.2. Cows eat several types of fodder (straw, hay, etc.), and require several types of nutrients (proteins, fat, etc.) to be healthy. We want to keep the cows healthy at minimal cost.

We construct a matrix A where each row corresponds to a fodder type, each column corresponds to a nutrient, and each element is the amount of that nutrient contained in that fodder (in some predefined units). Let x be a vector that describes how much of each fodder we choose to buy. Then xA describes how much of each nutrient we obtained. For each nutrient index i, the number $(xA)_i$ must be at least some threshold b_i to keep the cow healthy. We write $xA \geq b$ for short. This is a linear constraint on x. Another relevant constraint is $x \geq \mathbf{0}$.

Let c_i be the cost of one unit of fodder type i. Then the total cost associated with the variable x is $\langle c, x \rangle$, which is a linear objective function we wish to minimize. Together the constraints and objective define a linear program.

Remark 3.1.3. There is a close relationship between finding LP optimum and checking whether the linear constraints are satisfiable. If we can solve LP, then in particular we can obtain some solution. In the other direction, suppose we can check whether a solution exists, and we want to maximize $\langle c, x \rangle$. Add the constraint $\langle c, x \rangle \geq T$ for some T, and check for existence. We can proceed with binary search, assuming we know the possible range of $\langle c, x \rangle$.

Lecture 10
2020-12-22

Example 3.1.4. Suppose we have one set of points on the plane (pluses) and another set of points (minuses). We want a linear separator: a line such that all pluses are on one side of it, and all minuses are on the other side. Additionally, the margins around the line (distance to closest point) should be maximal.

Denote the separating line by $y = ax + b$, and the margin by ε. For every point (x_i, y_i) that should be above the line we have the linear constraint

$$y_i - ax_i - b_i \geq \varepsilon$$

and for every point (x_j, y_j) below the line we have

$$y_j - ax_j - b_j \leq -\varepsilon$$

This is a linear program: our variables are a, b, ε, and we want to maximize ε given these inequalities.

Here are some less obvious examples, that don't seem linear at first.

Example 3.1.5. Suppose we want to separate two sets of points on the plane with a parabola $y = ax^2 + bx + c$. Similarly to above, we would have the constraints:

$$y_i - ax_i^2 - bx_i - c \geq \varepsilon$$
$$y_j - ax_j^2 - bx_j - c \leq -\varepsilon$$

These are quadratics, but not in our variables. The variables are a, b, c, ε, and the constraints are linear with respect to them, so this is a linear program.

Example 3.1.6. Given a convex polygon, what is the maximal radius of a disc that can fit inside? Let $(u, v) \in \mathbb{R}^2$ be the disc center, and let R be the radius. Generally the distance of a point (u, v) from a line $y = ax + b$ is [10]:

$$\frac{v - au - b}{\sqrt{1 + a^2}}$$

This is a signed distance: it is positive if the point is above the line ($v > au + b$), and negative if it is below. Every edge of the polygon, described by some line $y = a_i x + b_i$, imposes a limitation on the location of (u, v) and on the radius R. If the interior of the polygon is above the line, then the signed distance should be at least R:

$$\frac{v - a_i u - b_i}{\sqrt{1 + a_i^2}} \geq R$$

Otherwise, the signed distance should be at most $-R$. The variables are u, v, R. Notice the constraints are linear with respect to them, and we want to maximize R. So this is a linear program.

Example 3.1.7. Given some points $(x_i, y_i)_{i \in I}$ on the plane, we want to draw a line that approximates them in the best way possible. Consider a line $y = ax + b$, so that the error for (x_i, y_i) is $ax_i + b - y_i$. We can define the total error size in various ways. For ℓ_2 (least squares):

$$\sum_{i \in I} ((ax_i + b) - y_i)^2$$

To find a, b that minimize this quantity, we can use calculus. The fact that it is easily differentiable is one of the reasons to consider squares in the first place.

In ℓ_∞, we want the largest $|ax_i + b - y_i|$ to be as small as possible. We can introduce a variable $\varepsilon \geq 0$ and constrain for all i:

$$-\varepsilon \leq ax_i + b - y_i \leq \varepsilon$$

These are linear constraints, and we want to minimize ε, hence this is a linear program.

In ℓ_1, we want to minimize the sum:

$$\sum_{i \in I} |ax_i + b - y_i|$$

We introduce $\varepsilon_i \geq 0$ for all $i \in I$ and constrain:

$$-\varepsilon_i \leq ax_i + b - y_i \leq \varepsilon_i$$

We want to minimize $\sum_{i \in I} \varepsilon_i$, so this is a linear program.

Remark 3.1.8. It is somewhat typical that ℓ_∞ was straightforward, ℓ_2 required some calculus, and ℓ_1 used a trick to take care of sum of absolute values.

[10] https://en.wikipedia.org/wiki/Distance_from_a_point_to_a_line#Another_formula

Integer linear programming (ILP) is a variation where we additionally require all variables to be integers. It is NP-hard: we think of true and false as 1 and 0. A reduction from 3-SAT would constrain each variable $x_1, \ldots, x_n$ to be between 0 and 1, and each clause like $x_1 \vee x_2 \vee \overline{x_3}$ would correspond to a constraint $x_1 + x_2 + (1 - x_3) \geq 1$.

Example 3.1.9. Optimal bipartite matching: we look for a perfect matching in a bipartite graph. Suppose that pairing vertices i and j is worth $c_{i,j}$, and we want to maximize our gains. Let $x_{i,j}$ be an indicator of whether i and j are paired, so we want to maximize $\sum c_{i,j} x_{i,j}$. Since i is uniquely paired, we have the following linear constraint:

$$\sum_j x_{i,j} = 1$$

We also want the constraint $x_{i,j} \in \{0, 1\}$, which is not linear. We can instead allow partial pairings, and only add the constraint $0 \leq x_{i,j} \leq 1$. It turns out that the optimal solution will be integers anyway (a similar phenomenon happens with maximal flow). Another approach is to round solutions to the nearest integers.

3.2 The simplex method

Definition 3.2.1. A *hyperplane* in $\mathbb{R}^n$ is defined by a vector $c \in \mathbb{R}^n$ and a scalar $\gamma \in \mathbb{R}$:

$$H = \{x \mid \langle c, x \rangle = \gamma\}$$

Suppose we have a linear program with the constraints $Ax \leq b$ and $x \geq \mathbf{0}$. To maximize $\langle c, x \rangle$ is to ask what is the largest γ such that H still intersects the polyhedron.

Consider a plane intersecting a room. Changing γ is like moving the plane. With the maximal γ for which the plane still intersects, the intersection will be either a corner of the room, or an entire wall (or ceiling or floor), or an edge. In any case, a corner of the room will be present in the intersection. So geometrically, it is reasonable that if a linear program has a finite optimum, then it is achieved by a vertex of the polyhedron. We show this formally below, and it will give us an algorithm for linear programs.

Lecture 11
2020-12-29

Definition 3.2.2. The *standard form* of a linear program has the constraints $Ax = b$, $x \geq \mathbf{0}$ and the objective is to maximize $\langle c, x \rangle$.

Definition 3.2.3. The *canonical form* of a linear program has the constraints $Ax \leq b$, $x \geq \mathbf{0}$ and the objective is to maximize $\langle c, x \rangle$.

Proposition 3.2.4. *Every linear program can be converted to standard form and to canonical form.*

Proof. To convert to standard form, we need to replace inequalities with equations. For any inequality of the form $\langle a, x \rangle \geq \alpha$, introduce a new variable $\eta \geq 0$ and require that $\langle a, x \rangle - \eta = \alpha$.

To convert to canonical form, replace any equation of the form $\langle a, x \rangle = \alpha$ with two inequalities, $\langle a, x \rangle \geq \alpha$ and $\langle a, x \rangle \leq \alpha$. □

Remark 3.2.5. We will use whichever form is most suitable for our needs. Note that the conversion is not expensive, and only doubles the size of the original problem.

For the following discussion, we assume the linear program is given in standard form.

Definition 3.2.6. A vector $x \in \mathbb{R}^n$ is called a *feasible solution* if it satisfies the constraints: $Ax = b$ and $x \geq \mathbf{0}$.

Definition 3.2.7. The *support* of a vector $x \in \mathbb{R}^n$ is the set of indices of non-zero coordinates:

$$\text{supp}(x) = \{1 \leq i \leq n \mid x_i \neq 0\}$$

Definition 3.2.8. A vector $x \in \mathbb{R}^n$ is called *basic* if the columns in A whose indices are $\text{supp}(x)$ are linearly independent. A basic feasible solution is abbreviated BFS.

Theorem 3.2.9. *Suppose the linear program has at least one feasible solution, and assume the objective function $\langle c, x \rangle$ that we want to maximize is bounded from above (so an optimum exists). Then there exists an optimum x which is also a BFS.*

Proof. Let x be a feasible solution, and consider the columns of A corresponding to $\text{supp}(x)$. If they are linearly independent, then x is BFS and we are done. Otherwise, there exists a non-trivial linear combination of these columns which is $\mathbf{0}$. Let $y \in \mathbb{R}^n$ have coefficients that correspond to this combination, and 0 for the other indices. Then $y \neq \mathbf{0}$ and $Ay = \mathbf{0}$ and $\text{supp}(y) \subseteq \text{supp}(x)$.

We consider a new vector $x + \epsilon y$. We will construct it such that:

1. It is a feasible solution: $x + \epsilon y \geq 0$ and $A(x + \epsilon y) = b$.
2. It has strictly smaller support: $\text{supp}(x + \epsilon y) \subsetneq \text{supp}(x)$.
3. It is not a worse solution: $\langle c, x + \epsilon y \rangle \geq \langle c, x \rangle$.

This will be sufficient: we can repeat the argument and make the support smaller, until it has size 1, which corresponds to a linearly independent set of columns (hence we arrived at a BFS). Since this process can be done for any feasible x, in particular it can be done for an optimal x, which shows there exists an optimum which is also BFS.

The constraint $A(x + \epsilon y) = Ax + \epsilon Ay = b$ is satisfied for any ϵ. To satisfy the rest we choose a specific one: the smallest in absolute value such that $x + \epsilon y \geq \mathbf{0}$ and one of the coordinates in $\text{supp}(x)$ becomes 0. So we have the first two conditions.

For the objective function, we have:

$$\begin{aligned} \langle c, x + \epsilon y \rangle &\geq \langle c, x \rangle \\ \langle c, x \rangle + \epsilon \langle c, y \rangle &\geq \langle c, x \rangle \\ \epsilon \langle c, y \rangle &\geq 0 \end{aligned}$$

If y has both positive and negative coordinates, then there exist both positive and negative choices of ϵ above. In that case we can choose the sign of ϵ to match the sign of $\langle c, y \rangle$, and satisfy this inequality.

Otherwise, suppose $y \geq \mathbf{0}$ (the negative case is symmetric). Then $x + \epsilon y \geq \mathbf{0}$ implies $\epsilon < 0$. If $\langle c, y \rangle \leq 0$, we are done. If $\langle c, y \rangle > 0$, the objective is not bounded: for any k we have

$$\begin{aligned} x + ky &\geq \mathbf{0} \\ A(x + ky) &= b \\ \langle c, x + ky \rangle &= \langle c, x \rangle + k \langle c, y \rangle \end{aligned}$$

So $x + ky$ is feasible, and $\langle c, x + ky \rangle \to \infty$ when $k \to \infty$, contradicting our assumption. $\square$

We now have an algorithm for LP: for every subset of columns of A, check whether they are linearly independent and b is in their span. If so, solve for x and calculate $\langle c, x \rangle$. Return the vector x for which the objective was largest.

Geometric view

We now show that a BFS corresponds to our intuitive notion of a vertex.

Definition 3.2.10. Let $H = \{x \mid \langle a, x\rangle = \alpha\}$ be a hyperplane. It partitions the rest of the space to points above and below H:

$$H^- = \{x \mid \langle a, x\rangle < \alpha\}$$
$$H^+ = \{x \mid \langle a, x\rangle > \alpha\}$$

Each of these sets is called a *half-space*. In some contexts we may consider half-spaces with weak inequality, in which case they are closed sets.

Definition 3.2.11. A *polyhedron* is an intersection of finitely many half-spaces. A *polytope* is a bounded polyhedron.

Theorem 3.2.12. *Equivalently, we can define a polytope to be the convex hull of a finite set X of points in $\mathbb{R}^n$:*

$$\operatorname{conv}(X) = \left\{ \sum_{i=1}^{|X|} \alpha_i x_i \;\middle|\; \alpha_i \geq 0,\ \sum_{i=1}^{|X|} \alpha_i = 1 \right\}$$

where $\operatorname{conv}(X)$ *is the smallest convex set containing X.*

Proposition 3.2.13. *The intersection of convex sets is convex.*

Definition 3.2.14. Let $S = \{x_1, \ldots, x_k\}$ be a finite set of points in $\mathbb{R}^n$. The *affine space* spanned by S is defined as:

$$\operatorname{aff}(S) = \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; \sum_{i=1}^{k} \alpha_i = 1 \right\}$$

This is a linear space which has been translated ("moved"), so it does not necessarily contain $\mathbf{0}$.

Definition 3.2.15. The *affine dimension* of an affine space X is the dimension of the corresponding linear space (after moving X so that it contains the origin). We denote it $\operatorname{aff\,dim}(X)$.

Definition 3.2.16. Let P be a polytope and H be a hyperplane. We say H is a *supporting hyperplane* of P if P is contained in one of its closed half-spaces H^+ or H^-, and also $P \cap H \neq \emptyset$. The set $P \cap H$ is called a *face*. Its dimension is $\operatorname{aff\,dim}(P \cap H)$.

Definition 3.2.17. A 0-dimensional face of a polytope P is called a *vertex*. A 1-dimensional face is called an *edge*. A $(d-1)$-dimensional face is called a *facet*, where $d = \operatorname{aff\,dim}(P)$.

Theorem 3.2.18. *Let $P = \{x \mid Ax = b,\ x \geq 0\}$ be a polytope corresponding to a standard LP, and let $v \in P$. Then v is a vertex if and only if it is BFS.*

Proof. Suppose v is a vertex. Then there is a supporting hyperplane $H = \{x \mid \langle c, x\rangle = \gamma\}$ such that $H \cap P = \{v\}$ and $P \subseteq H^-$. Every other $z \in P$ has $\langle c, z\rangle < \gamma$, so v is a unique maximizer of $\langle c, x\rangle$. By the previous theorem, the maximum is obtained by a BFS. Since v is the only maximizer, it must be a BFS.

In the other direction, suppose v is a BFS. We want to find a supporting hyperplane which intersects P only at v. Define a vector c to have 0 in the indices $\operatorname{supp}(v)$, and -1 elsewhere. We have $\langle c, v\rangle = 0$ directly from definition. For every $z \in P$, if z contains a positive coordinate outside $\operatorname{supp}(v)$ then $\langle c, z\rangle < 0$, in which case z is not a maximizer. Suppose z is 0 outside $\operatorname{supp}(v)$, so $\operatorname{supp}(z) \subseteq \operatorname{supp}(v)$. The coordinates of z describe a linear combination of the columns of A that equals b. Since v is a BFS, these columns are linearly independent, so this representation is unique, thus $z = v$. Therefore the hyperplane $H = \{x \mid \langle c, x\rangle = 0\}$ is a supporting hyperplane which intersects P only at v. $\square$

Algorithm

The optimum of an LP is obtained at a BFS, so we only consider vertices of the polyhedron. The approach is to start at some vertex, then repeatedly move to a neighboring vertex which is better with respect to the objective. This is called Dantzig's simplex algorithm. It is in fact a family of algorithms, which differ in the way they choose neighbors.

Geometrically, we have an intuitive understanding of what a neighbor is. We call $u, v \in P$ neighbors if the segment $[u, v]$ is an edge (1-dimensional face) of P.

Algebraically, consider a subset of the columns of A which forms a basis of the column space. This subset corresponds to a unique vertex which solves $Ax = b$. Two vertices are called neighbors if the subsets only differ in one column. To move to a neighbor, remove one of the columns from the subset and add another column which will complete it to a basis.

For linear programs, there is a theorem that guarantees any local optimum is a global optimum. That is, if none of the neighbors of the current vertex are better, then it is best overall. The proof approach is to assume there is some direction which leads to a better vertex, and notice that this direction is a convex combination of neighboring vertices. The objective is non-increasing in the directions of all neighbors, so it must also be non-increasing for any convex combination of them.

The complexity of the simplex method is not known. In practice there are specific variants (methods of choosing neighbors) that work very well, and solve problems with millions of constraints. There are also variants that have been shown to be not polynomial.

Part II

Recitations

Recitation 1
2020-10-21

4 Probability

4.1 Basics

Definition 4.1.1. A *probability space* is a triplet $(\Omega, \mathcal{F}, \Pr)$ where Ω is the set of possible states of the world, $\mathcal{F}$ is a collection of subsets of Ω called events, and $\Pr : \mathcal{F} \to [0,1]$ assigns probabilities.

For example, take Ω to be a finite set, $\mathcal{F}$ to be all subsets of Ω, and $\{p_\omega\}_{\omega\in\Omega}$ non-negative numbers such that

$$\sum_{\omega\in\Omega} p_\omega = 1$$

and for any $E \in \mathcal{F}$,

$$\Pr[E] = \sum_{\omega\in E} p_\omega$$

For a balanced die, we can take $\Omega = [6]$ and $p_i = \frac{1}{6}$ for all $1 \le i \le 6$.

Two important properties of probability spaces:

- Monotonicity: if $E_1 \subseteq E_2$ are events then $\Pr[E_1] \le \Pr[E_2]$.
- Additivity: if E_1, E_2 are disjoint events then $\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2]$.

They imply the union bound which you should prove yourself:

$$\Pr[E_1 \cup E_2] \le \Pr[E_1] + \Pr[E_2]$$

Definition 4.1.2. *Random variables* are functions X from Ω to some $\mathcal{X}$, for example $\mathbb{R}$.

We often define events using the values of RVs, such as:

$$\Pr[X \ge t] = \Pr[\{\omega \in \Omega \mid X(\omega) \ge t\}]$$

Definition 4.1.3. The *expectation* (*mean*) of X is the weighted average:

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x \Pr[X = x]$$

For continuous RVs, we replace the sum with an integral.

$$\mathbb{E}[X] = \int x f_X(x) dx$$

where the integral is taken over $\mathbb{R}$ and f_X is the density function of X.

Definition 4.1.4. An *indicator* is a random variable with range $\{0,1\}$. For any event E we can define an indicator

$$\mathbb{1}_E(\omega) = \begin{cases} 1 & \omega \in E \\ 0 & \omega \notin E \end{cases}$$

Notice that $\mathbb{E}[\mathbb{1}_E] = \Pr[E]$.

Proposition 4.1.5. *Linearity of expectation: for any RVs X, Y and numbers α, β we have*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$$

Definition 4.1.6. *Variance* is the average squared deviation of an RV:

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

it measures how much X is spread. Useful identity to prove yourself:

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Lemma 4.1.7. $\mathrm{Var}[X] = 0$ *if and only if X is deterministic (constant with probability 1).*

Proof. $\Rightarrow$ Suppose $X = c$ with probability 1. Then $\mathbb{E}[X] = c$, so $(X - \mathbb{E}[X])^2 = (c - c)^2 = 0$, hence $\mathrm{Var}[X] = 0$.

$\Leftarrow$ Suppose $\mathrm{Var}[X] = 0$. Then:

$$0 = \mathrm{Var}[X] = \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 \Pr[X = x]$$

All terms are non-negative so each must be 0. In terms where $x \neq \mathbb{E}[X]$, we have $(x - \mathbb{E}[X])^2 \neq 0$, so the probability factor must be $\Pr[X = x] = 0$. This leaves only the term where $x = \mathbb{E}[X]$. Since the probabilities must sum to 1, we must have $\Pr[X = \mathbb{E}[X]] = 1$. □

Definition 4.1.8. A collection $X_1, \ldots, X_k$ of discrete RVs is called *independent* if

$$\Pr[X_1 = x_1 \wedge \ldots \wedge X_k = x_k] = \Pr[X_1 = x_1] \cdots \Pr[X_k = x_k]$$

for all $x_1 \in \mathcal{X}_1, \ldots, x_k \in \mathcal{X}_k$.

Lemma 4.1.9. *$X_1, \ldots, X_k$ are independent if and only if for every functions $g_1, \ldots, g_k$ such that $g_i : \mathcal{X}_i \to \mathbb{R}$ we have*

$$\mathbb{E}\left[\prod_{i=1}^{k} g_i(X_i)\right] = \prod_{i=1}^{k} \mathbb{E}[g_i(X_i)]$$

The proof is left as an exercise.

Definition 4.1.10. Let X, Y be RVs. Their *covariance* is defined as

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

it measures how much they tend to deviate from their means in the same direction.

Lemma 4.1.11. *If X, Y are independent then* $\mathrm{Cov}(X, Y) = 0$.

This follows from the previous lemma. Note the converse is not true.

Lemma 4.1.12. *Let $X_1, \ldots, X_n$ be RVs. We have:*

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + \sum_{\substack{i,j \in [n] \\ i \neq j}} \mathrm{Cov}(X_i, X_j)$$

Proof. Denote $\tilde{X}_i = X_i - \mathbb{E}[X_i]$. It is X_i with value shifted such that $\mathbb{E}\left[\tilde{X}_i\right] = 0$. We have by definition:

$$\begin{aligned}
\operatorname{Var}[X_i] &= \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] = \mathbb{E}\left[\tilde{X}_i^2\right] \\
\operatorname{Cov}(X_i, X_j) &= \mathbb{E}\left[(X_i - \mathbb{E}[X_i])\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] = \mathbb{E}\left[\tilde{X}_i \tilde{X}_j\right]
\end{aligned}$$

Shifting the values of an RV by a constant does not affect its variance, so

$$\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \operatorname{Var}\left[\sum_{i=1}^{n} \tilde{X}_i\right]$$

and we have by linearity of expectation:

$$\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{n} \tilde{X}_i\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] - \mathbb{E}\left[\sum_{i=1}^{n} \tilde{X}_i\right]^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\tilde{X}_i^2\right] + \sum_{\substack{i,j \in [n] \\ i \neq j}} \mathbb{E}\left[\tilde{X}_i \tilde{X}_j\right] \\
&= \sum_{i=1}^{n} \operatorname{Var}[X_i] + \sum_{\substack{i,j \in [n] \\ i \neq j}} \operatorname{Cov}(X_i, X_j)
\end{aligned}$$

□

Corollary 4.1.13. *If $X_1, \ldots, X_n$ are independent with sum $S_n = \sum_{i=1}^{n} X_i$, and their variance is bounded $\operatorname{Var}[X_i] \leq \sigma^2$ for all i, then:*

$$\operatorname{Var}[S_n] = \sum_{i=1}^{n} \operatorname{Var}[X_i] \leq n\sigma^2$$

Theorem 4.1.14 (Chebyshev's inequality)**.** *We have:*

$$\Pr\left[|Z - \mathbb{E}[Z]| \geq t\right] \leq \frac{\operatorname{Var}[Z]}{t^2}$$

Theorem 4.1.15 (weak law of large numbers)**.** *Let $X_1, \ldots, X_n$ be i.i.d (independent, identically distributed). Let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \operatorname{Var}[X_1]$. Then for any $t > 0$,*

$$\lim_{n \to \infty} \Pr\left[\left|\frac{1}{n} \cdot S_n - \mu\right| \geq t\right] = 0$$

Informally, when we perform an experiment n times and observe some average ($\frac{1}{n} \cdot S_n$), it will approach the theoretical average (μ) as n grows.

Proof. Consider the random variable $\frac{1}{n} S_n$. We have:

$$\operatorname{Var}\left[\frac{1}{n} S_n\right] = \frac{1}{n^2} \operatorname{Var}[S_n] = \frac{\sigma^2}{n}$$

If we put $Z = \frac{1}{n}S_n$ in Chebyshev's inequality,

$$\Pr\left[\left|\frac{1}{n}S_n - \mu\right| \geq t\right] \leq \frac{\sigma^2}{nt^2} \to 0$$

□

Definition 4.1.16. A Bernoulli RV is an indicator X which is 1 with some defined probability α, and 0 with probability $1-\alpha$. We denote this fact by $X \sim \mathrm{Ber}(\alpha)$.

Definition 4.1.17. Fix some natural number n and some real $0 \leq \alpha \leq 1$. A binomial RV is a variable $X : \Omega \to \{0, 1, 2, \ldots\}$ such that for every non-negative integer k,

$$\Pr[X = k] = \binom{n}{k}\alpha^k(1-\alpha)^{n-k}$$

We denote this by $X \sim \mathrm{Bin}(n, \alpha)$.

This definition is motivated by the sum of Bernoulli variables. Suppose we perform n independent experiments, and each has success probability α. The probability that exactly k experiments succeeded is α^k (for the successes), times $(1-\alpha)^{n-k}$ (for the failures), times $\binom{n}{k}$ (for the number of such arrangements).

Formally, if $X_1, \ldots, X_n \sim \mathrm{Ber}(\alpha_n)$, then $S_n = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \alpha_n)$ (here we allow the success probability to depend on n). We have

$$\begin{aligned}\mathbb{E}[S_n] &= \alpha_n n \\ \mathrm{Var}[S_n] &= n\mathrm{Var}[X_i] = n\alpha_n(1-\alpha_n)\end{aligned}$$

How does S_n usually behave? This depends on the way α_n grows.

- Suppose $\alpha_n n \to \infty$. Then
$$\sigma_n = \sqrt{\mathrm{Var}[S_n]} = \sqrt{\alpha_n(1-\alpha_n)n}$$
So in this case
$$\sigma_n = o(\alpha_n n) = o(\mathbb{E}[S_n])$$
Take c_n to be some sequence which is strictly between σ_n and $\sigma_n n$ asymptotically, that is,
$$\omega(\sigma_n) = c_n = o(\alpha_n n)$$
By Chebyshev,
$$\Pr[|S_n - \alpha_n n| \geq c_n] \leq \frac{\sigma_n^2}{c_n^2} \to 0$$
So deviations of size c_n from the mean are unlikely. The mean is large in comparison, so $S_n \approx \mathbb{E}[S_n]$.

- Suppose $\alpha_n n \to 0$. In this case,
$$\sigma_n = \sqrt{\alpha_n(1-\alpha_n)n} \approx \sqrt{\alpha_n n} \gg \alpha_n n = \mathbb{E}[S_n]$$
Take c_n to be a sequence that tends to 0 and is $\omega(\sigma_n)$. Then we can apply Chebyshev as above,
$$\Pr[|S_n - \alpha_n n| \geq c_n] \leq \frac{\sigma_n^2}{c_n^2} \to 0$$

For large enough n we will have $\alpha_n n + c_n < 1$ and also $S_n < \alpha_n n + c_n$ with high probability. Since S_n is a non-negative integer, this means $\Pr[S_n = 0] \to 1$.

Alternatively we could have used a union bound:

$$\begin{aligned}\Pr[S_n > 0] &= \Pr[X_i = 1 \text{ for some } 1 \le i \le n] \\ &\le \sum_{i=1}^{n} \Pr[X_i = 1] \\ &= n\alpha_n \to 0\end{aligned}$$

- Suppose $\alpha_n n = \Theta(1)$. In this case, the standard deviation $\sigma_n = \Theta(1)$ has the same order of magnitude as the mean. So S_n does not concentrate.

  In the special case where $\alpha_n = \frac{\lambda}{n}$ for a constant λ, the Poisson limit theorem states that S_n converges to a Poisson RV with parameter λ. That is, for every fixed k,

$$\lim_{n\to\infty} \Pr[S_n = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

4.2 Connectivity threshold

Recitation 2
2020-10-28

$G(n,p)$ is a model of random graphs on the vertices $[n]$, where every pair $i \neq j$ are connected with independent probability p. It is called the Erdős–Rényi model.

The average degree of a vertex is $p(n-1)$, or $\approx pn$ when n is large.

Suppose p is a function of n. We often ask whether a graph in $G(n,p)$ satisfies some property almost surely, that is: the probability of $G \sim G(n,p)$ satisfying the property tends to 1 when $n \to \infty$. Such properties include whether the graph is connected, whether it contains a copy of K_4, whether it has exactly two connected components, and so on.

Definition 4.2.1. A property $\mathcal{P}$ has a *threshold* $f(n)$ if both of the following hold:

- When $p \ll f(n)$, a random graph satisfies $\mathcal{P}$ with probability tending to 0.
- When $p \gg f(n)$, a random graph satisfies $\mathcal{P}$ with probability tending to 1.

Example 4.2.2. We have seen in the lecture that $f(n) = n^{-2/3}$ is the threshold for containing K_4.

Definition 4.2.3. A property $\mathcal{P}$ is called *monotone* if it is preserved when adding edges. That is, if G has $\mathcal{P}$ and G' is obtained from G by adding edges, then G' has $\mathcal{P}$.

Example 4.2.4. Connectivity and containing K_4 are monotone properties.

Example 4.2.5. G having exactly 2 connected components, and G not having exactly 2 connected components, are non-monotone properties.

Theorem 4.2.6 (Bollobás-Thomason)**.** *Every property which is non-trivial (some graphs satisfy it and some graphs don't) and monotone has a threshold.*

We will study a threshold of connectivity.

Definition 4.2.7. A graph G is *connected* if for any two vertices i, j, there is a path between them.

We will show the threshold of connectivity is $f(n) = \frac{\ln n}{n}$. There is a more precise result, which we will not prove:

Theorem 4.2.8 (Erdős–Rényi)**.** *Suppose $p = \frac{\ln n}{n} + \frac{c}{n}$ where c may depend on n. Then:*

$$\lim_{n\to\infty} \Pr\left[G \text{ is connected}\right] = \begin{cases} 1 & c \to \infty \\ 0 & c \to -\infty \\ e^{-e^{-c}} & c \text{ constant} \end{cases}$$

We will show:

Theorem 4.2.9. *Suppose $p = \lambda \frac{\ln n}{n}$ for a constant $\lambda > 0$. Then:*

- *If $\lambda > 1$ then G is almost surely connected .*
- *If $\lambda < 1$ then G is almost surely not connected.*

Proof. The first step is to define random variables that can capture the connectivity property of G in some way, and yet are easy to study (notice that defining a connectivity indicator does not seem to help much).

Observe that G is not connected iff we can partition its vertices into two subsets, S and $[n]\backslash S$, such that $|S| \leq \frac{n}{2}$ is the smaller one, and there are no edges between S and $[n]\backslash S$.

Define an indicator X_S for the event that S is disconnected from $[n]\backslash S$, and a counter for such events according to the set size[11]:

$$C_k = \sum_{\substack{S\subseteq[n] \\ |S|=k}} X_S$$

And define the sum:

$$C = \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} C_k$$

If $C = 0$ then there are no such ways to partition G, so G is connected. If $C > 0$, then there is at least one way, so G is disconnected. It is left to understand C sufficiently.

For some fixed S with $|S| = k$, there are k vertices in S and $n - k$ outside of S, giving a total of $k(n-k)$ potential edges. S is disconnected from $[n]\backslash S$ iff none of these edges exist:

$$\Pr\left[X_S = 1\right] = (1-p)^{k(n-k)}$$

Therefore,

$$\mathbb{E}\left[C_k\right] = \sum_{\substack{S\subseteq[n] \\ |S|=k}} \mathbb{E}\left[X_S\right] = \binom{n}{k}(1-p)^{k(n-k)}$$

- Suppose $\lambda > 1$. We want to show $C = 0$ almost surely. We use the first moment method, and the inequalities $1 - x \leq e^{-x}$ and $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

$$\begin{aligned}
\mathbb{E}\left[C_k\right] &= \binom{n}{k}(1-p)^{k(n-k)} \leq \left(\frac{ne}{k}\right)^k e^{-pk(n-k)} \\
&= \left(\frac{ne}{k}\right)^k e^{-\lambda \frac{\ln n}{n} k(n-k)} = \left(\frac{ne}{k} e^{-\lambda \frac{\ln n}{n}(n-k)}\right)^k \\
&= \left(\frac{ne}{k} e^{-\lambda \ln n} e^{\lambda \ln n \cdot \frac{k}{n}}\right)^k = \left(en^{-\lambda} \frac{n}{k} e^{\lambda \ln n \cdot \frac{k}{n}}\right)^k \\
&= a_{k,n}^k
\end{aligned}$$

[11] Notice that $C_{n/2}$ counts every such partition twice. This doesn't matter to the proof.

Define $f(x) = \frac{e^{\lambda \ln n \cdot x}}{x}$ so that $a_{k,n} = en^{-\lambda} f(k/n)$. We bound $a_{k,n}$ by bounding f with some calculus. Notice that f is differentiable for any interval $[a, b]$ with $0 < a < b$, and takes at most one minimum and no maxima.

$$f'(x) = e^{\lambda \ln n \cdot x} \left(\frac{\lambda \ln n \cdot x - 1}{x^2} \right)$$
$$f''(x) = e^{\lambda \ln n \cdot x} \left(\frac{\lambda^2 \ln(n)^2 x^2 - 2\lambda \ln n \cdot x + 2}{x^3} \right)$$

Therefore the maximum of f in $[a, b]$ is in one of the interval's endpoints. We are interested in $f(k/n)$ for $1 \leq k \leq n/2$, so consider the interval $\left[\frac{1}{n}, \frac{1}{2}\right]$. For all such k we have:

$$f(k/n) \leq \max\left(f(1/n), f(1/2)\right)$$

and

$$f(1/n) = \frac{e^{\lambda \ln n \cdot \frac{1}{n}}}{\frac{1}{n}} = n \cdot n^{\lambda \cdot \frac{1}{n}} = n^{1+o(1)}$$
$$f(1/2) = \frac{e^{\lambda \ln n \cdot \frac{1}{2}}}{\frac{1}{2}} = 2n^{\lambda \cdot \frac{1}{2}}$$

So

$$\begin{aligned} a_{k,n} &= en^{-\lambda} f(k/n) \\ &= en^{-\lambda} \max\left(n^{1+o(1)}, 2n^{\lambda/2}\right) \\ &= \max\left(en^{-\lambda} n^{1+o(1)}, 2en^{-\lambda} n^{\lambda/2}\right) \\ &= O\left(\max\left(n^{1-\lambda+o(1)}, n^{-\lambda/2}\right)\right) \end{aligned}$$

We have $\lambda > 1$ so this tends to 0. Back to C, we have:

$$\begin{aligned} \mathbb{E}[C] &= \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \mathbb{E}[C_k] \\ &\leq \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} a_{k,n}^k \\ &= \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} O\left(\max\left(n^{1-\lambda+o(1)}, n^{-\lambda/2}\right)\right)^k \\ &= \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} q^k \end{aligned}$$

Bounding with an infinite geometric series for $q \ll 1$,

$$\sum_{k=1}^{\infty} q^k = \frac{q}{1-q} = O(q)$$

and we get

$$\mathbb{E}[C] = O\left(\max\left(n^{1-\lambda+o(1)}, n^{-\lambda/2}\right)\right) \to 0$$

Finally, by Markov's inequality,

$$\Pr[C \geq 1] \leq \mathbb{E}[C] \to 0$$

So when $\lambda > 1$, with high probability there are no subsets S without edges outside, so G is almost surely connected.

- Suppose $\lambda < 1$. We want to show that $C \geq 1$ almost always. We show a stronger result: the number of isolated vertices, C_1, is almost always positive. We have:

$$\mathbb{E}[C_1] = n(1-p)^{n-1} = e^{\ln n + (n-1)\ln(1-p)}$$

From Taylor we have $\log(1-p) = -p + O(p^2)$ so

$$\begin{aligned}\mathbb{E}[C_1] &= e^{\ln n + (n-1)(-p+O(p^2))} \\ &= e^{\ln n - np + p + O(np^2) - O(p^2)}\end{aligned}$$

Notice $p \to 0$ and $np^2 = n\lambda^2 \frac{\ln(n)^2}{n^2} \to 0$ so

$$\begin{aligned}\mathbb{E}[C_1] &= e^{\ln n - np + o(1)} \\ &= e^{\ln n - n\lambda \frac{\ln n}{n} + o(1)} \\ &= e^{(1-\lambda)\ln n + o(1)} \\ &= n^{1-\lambda+o(1)}\end{aligned}$$

and this tends to ∞ because $\lambda < 1$.

Even though $\mathbb{E}[C_1] \to \infty$, this does not yet show that $C_1 \geq 1$ with high probability. Our goal is to use Chebyshev, so we need the variance.

Let X_i be an indicator for the event that vertex i is isolated. Notice that $X_i X_j$ is an indicator that i, j are both isolated. For any $i \neq j$ we have:

$$\mathbb{E}[X_i] = \mathbb{E}\left[X_j\right] = (1-p)^{n-1}$$

For both to be isolated we need i to be disconnected from everyone $(n-1)$ and j to be disconnected from everyone ($n-2$, because we already accounted for $\{i, j\}$). Therefore:

$$\mathbb{E}\left[X_i X_j\right] = (1-p)^{(n-1)+(n-2)} = (1-p)^{2n-3}$$

Hence:

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}[X_i]\,\mathbb{E}\left[X_j\right] \\ &= (1-p)^{2n-3} - (1-p)^{2n-2} \\ &= (1-p)^{2n-3}(1-(1-p)) \\ &= p(1-p)^{2n-3} \\ &\leq pe^{-p(2n-3)} \\ &\approx pe^{-2np} \\ &= \lambda \frac{\ln n}{n} e^{-2n\lambda \frac{\ln n}{n}} \\ &= \lambda \frac{\ln n}{n} n^{-2\lambda} \\ &= \lambda \ln n \cdot n^{-2\lambda-1}\end{aligned}$$

The variance of one RV is:

$$\begin{aligned}
\mathrm{Var}\left[X_i\right] &= \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2 \\
&\leq \mathbb{E}\left[X_i^2\right] \\
&= \mathbb{E}\left[X_i\right] \\
&= (1-p)^{n-1} \\
&\approx e^{-pn} \\
&= n^{-\lambda}
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Var}\left[C_1\right] &= \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \\
&\leq n \cdot O(n^{-\lambda}) + O(n^2) \cdot \lambda \ln n \cdot n^{-2\lambda-1} \\
&= O(n^{1-\lambda}) + O(n^{1-2\lambda} \ln n) \\
&= O(n^{1-\lambda})
\end{aligned}$$

(the last transition breaks down at $\lambda = 0$ but that's a trivial case). Finally with Chebyshev,

$$\begin{aligned}
\Pr\left[C_1 = 0\right] &\leq \frac{\mathrm{Var}\left[C_1\right]}{\mathbb{E}\left[C_1\right]^2} \\
&= \frac{O(n^{1-\lambda})}{\left(n^{1-\lambda+o(1)}\right)^2} \\
&= O\left(\frac{n^{1-\lambda}}{n^{2(1-\lambda)+o(1)}}\right) \\
&= O\left(n^{\lambda-1+o(1)}\right)
\end{aligned}$$

and this tends to 0 since $\lambda < 1$.

□

4.3 Concentration inequalities

Recitation 3
2020-11-04

Generally deviations of an RV X from $\mathbb{E}\left[X\right]$ is quantified by Chebyshev's inequality:

$$\Pr\left[\left|X - \mathbb{E}\left[X\right]\right| \geq t\right] \leq \frac{\mathrm{Var}\left[X\right]}{t^2}$$

We see that for very large t, this becomes unlikely. Can we achieve better bounds if we know more about X? When X is composed of many small independent parts, we sometimes can. Chebyshev bounds the tail with polynomial decay (t^{-2}), but we can sometimes achieve exponential decay if we know more.

Our starting point is Markov's inequality:

Theorem 4.3.1 (Markov's inequality)**.** *For $X \geq 0$ and any $t > 0$ we have*

$$\Pr\left[X \geq t\right] \leq \frac{\mathbb{E}\left[X\right]}{t}$$

Proof. Observe that $X \geq X \cdot \mathbb{1}_{X \geq t} \geq t \cdot \mathbb{1}_{X \geq t}$, and take expectation:

$$\mathbb{E}[X] \geq \mathbb{E}[X \cdot \mathbb{1}_{X \geq t}] \geq \mathbb{E}[t \cdot \mathbb{1}_{X \geq t}] = t\mathbb{E}[\mathbb{1}_{X \geq t}] = t\Pr[X \geq t]$$

□

This has many uses when $\mathbb{E}[X]$ is easy to calculate. But when it's not, we want to bound the expectation from above using some friendly function f. Suppose f is non-decreasing and non-negative. The event $X \geq t$ implies $f(X) \geq f(t)$, so Markov gives

$$\Pr[X \geq t] \leq \Pr[f(X) \geq f(t)] \leq \frac{\mathbb{E}[f(X)]}{f(t)}$$

This may be useful in some situations where $f(X)$ and $\mathbb{E}[f(X)]$ are much nicer. To see how far the bound is, notice that (for finite Ω)

$$\Pr[X \geq t] = \sum_{\omega \in \Omega} \mathbb{1}_{X(\omega) \geq t} \Pr[\omega]$$

while the upper bound is

$$\frac{\mathbb{E}[f(X)]}{f(t)} = \sum_{\omega \in \Omega} \frac{f(X(\omega))}{f(t)} \Pr[\omega]$$

so the closer $\frac{f(x)}{f(t)}$ is to $\mathbb{1}_{x \geq t}$, the tighter the bound. Since f should be nice, we will have some trade-off between how close the ratio is to 0 when $x < t$ and how close it is to 1 when $x \geq t$.

Method 1: moments

We can take f to be the pth absolute moment:

$$f(x) = |x|^p$$

This gives

$$\mathbb{E}[|X| \geq t] \leq \frac{\mathbb{E}\left[|X|^p\right]}{t^p}$$

Notice that for larger p, the ratio $\frac{f(x)}{f(t)} = \left|\frac{x}{t}\right|^p$ is closer to 0 for $x < t$, which is good. However, we pay for this when $x > t$, as it will be much higher than 1 and thus a very loose bound.

Remark 4.3.2. In practice, it's not obvious which p would work best, and it depends on the specifics of the problem.

Remark 4.3.3. The existence of the pth moment is not guaranteed. When it does exist, it already means the tail $\Pr[X \geq t]$ tends to 0 like $\frac{1}{t^p}$ (or faster).

Method 2: exponentials

We can take some real parameter $\lambda > 0$ and define

$$f_\lambda(x) = e^{\lambda x}$$

This gives

$$\Pr[X \geq t] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right]$$

and on the other side,

$$\Pr[X \leq -t] = \Pr[-X \geq t] = \Pr\left[e^{-\lambda X} \geq e^{\lambda t}\right] \leq e^{-\lambda t}\mathbb{E}\left[e^{-\lambda X}\right]$$

For the absolute size of X,

$$\begin{aligned}\Pr[|X| \geq t] &= \Pr[X \geq t] + \Pr[X \leq -t] \\ &\leq e^{-\lambda t}\left(\mathbb{E}\left[e^{\lambda X}\right] + \mathbb{E}\left[e^{-\lambda X}\right]\right)\end{aligned}$$

Definition 4.3.4. The function $M_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$ is called the *moment generating function (MGF)* of X.

It is not obvious for which values of λ this function can be defined (except the trivial case of $\lambda = 0$). If it does exist for some interval $\lambda \in [-\lambda_0, \lambda_0]$, then we have

$$\Pr[|X| \geq t] \leq (M_X(\lambda_0) + M_X(-\lambda_0))\, e^{-\lambda_0 t}$$

The tail here goes to 0 with exponential speed.

For any $\lambda \in (-\lambda_0, \lambda_0)$, we can expand M_X as a power series:

$$M_X(\lambda) = \mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{\lambda^k X^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[X^k\right]}{k!}\lambda^k$$

Notice that the kth derivative of $M_X(\lambda)$ is exactly $\mathbb{E}\left[X^k\right]$ at $\lambda = 0$, so in a way M_X defines every moment (hence its name).

Hoeffding's inequality

Suppose M_X exists for all $\lambda \in \mathbb{R}$. Then we have:

$$\Pr[X \geq t] \leq M_X(\lambda)e^{-\lambda t} = e^{\ln M_X(\lambda) - \lambda t}$$

Definition 4.3.5. The *cumulant generating function* of X is $\psi_X(\lambda) = \ln M_X(\lambda)$.

The tightest version of this bound occurs when λ minimizes the expression:

$$\Pr[X \geq t] \leq \min_{\lambda > 0} e^{\psi_X(\lambda) - \lambda t} = e^{\min_{\lambda > 0}(\psi_X(\lambda) - \lambda t)}$$

So bounding M_X, or alternatively ψ_X, will give a bound for the tail.

Suppose now that $X = \sum_{i=1}^{n} Z_i$ is a sum of independent variables. Then by independence,

$$M_X(\lambda) = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} Z_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda Z_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z_i}\right] = \prod_{i=1}^{n} M_{Z_i}(\lambda)$$

Therefore taking logarithms we get:

$$\psi_X(\lambda) = \sum_{i=1}^{n} \psi_{Z_i}(\lambda)$$

which is convenient.

Lemma 4.3.6 (Hoeffding's lemma)**.** *If* $\mathbb{E}[Z] = 0$ *and* $|Z| \leq a$ *with probability* 1*, then*

$$\psi_Z(\lambda) \leq \frac{1}{2}a^2\lambda^2$$

The proof can be found in TA3 (it was not shown in class).

Theorem 4.3.7 (Hoeffding's inequality)**.** *If* $Z_1, \ldots, Z_n$ *are independent,* $\mathbb{E}[Z_i] = 0$ *and* $|Z_i| \leq a$ *for all* i*, then*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq t\right] \leq 2e^{-\frac{nt^2}{2a^2}}$$

Proof. Let $X = \sum_{i=1}^{n} Z_i$. We want to bound $\Pr[|X| \geq nt]$. We have from the lemma:

$$\psi_X(\lambda) = \sum_{i=1}^{n} \psi_{Z_i}(\lambda) \leq \sum_{i=1}^{n} \frac{1}{2}a^2\lambda^2 = \frac{1}{2}na^2\lambda^2$$

and as we have seen,

$$\Pr[X \geq nt] \leq M_X(\lambda)e^{-\lambda nt} = e^{\psi_X(\lambda)-\lambda nt}$$
$$\Pr[X \leq -nt] \leq M_X(-\lambda)e^{-\lambda nt} = e^{\psi_X(-\lambda)-\lambda nt}$$

Bounding $\psi_X(\lambda), \psi_X(-\lambda)$ by $\frac{1}{2}na^2\lambda^2$ we have:

$$\begin{aligned}\Pr[|X| \geq nt] &\leq e^{\psi_X(\lambda)-\lambda nt} + e^{\psi_X(-\lambda)-\lambda nt} \\ &\leq e^{\frac{1}{2}na^2\lambda^2-\lambda nt} + e^{\frac{1}{2}na^2\lambda^2-\lambda nt} \\ &= 2e^{\frac{1}{2}na^2\lambda^2-\lambda nt} \\ &= 2e^{-n\left(\lambda t-\frac{1}{2}a^2\lambda^2\right)}\end{aligned}$$

To minimize the expression, we want to maximize the quadratic $f(\lambda) = \lambda t - \frac{1}{2}a^2\lambda^2$. This happens when $\lambda = \frac{t}{a^2}$, and the expression's value is

$$f\left(\frac{t}{a^2}\right) = \frac{t}{a^2}t - \frac{1}{2}a^2\frac{t^2}{a^4} = \frac{t^2}{a^2} - \frac{1}{2}\cdot\frac{t^2}{a^2} = \frac{t^2}{2a^2}$$

Therefore:

$$\Pr[|X| \geq nt] \leq 2e^{-n\left(\lambda t-\frac{1}{2}a^2\lambda^2\right)} \leq 2e^{-n\cdot\frac{t^2}{2a^2}}$$

□

Remark 4.3.8. If we think of a as fixed and n large, notice that $t = \frac{1}{\sqrt{n}}$ cancels out the n in the bound, resulting in a non-negligible constant. The bound goes to 0 for any $t \gg \frac{1}{\sqrt{n}}$. In other words, the typical deviations of $\frac{1}{n}X$ from 0 are going to be on the order of magnitude of $\frac{1}{\sqrt{n}}$ (this is comparable to the central limit theorem).

Remark 4.3.9. Hoeffding's inequality doesn't make use of the variance (it only relies on $|Z_i| \leq a$). This is a drawback: if the variance is very small, the tail's decay may be much faster than this bound. There are more complicated variations, that bound ψ_X more carefully.

Example 4.3.10. Suppose $S \sim \text{Bin}(n, \alpha)$, like we saw in the first recitation, and assume $\alpha = o\left(\frac{1}{n}\right)$. The expected value of S is αn, and we want to bound the probability that $|S - \alpha n|$ is large. Define Z_i to be Bernoulli$(\alpha) - \alpha$, that is,

$$\Pr[Z_i = 1 - \alpha] = \alpha$$
$$\Pr[Z_i = -\alpha] = 1 - \alpha$$

Then define the sum:

$$X = \sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} (\text{Bernoulli}(\alpha) - \alpha) = S - n\alpha$$

Notice that $\mathbb{E}[Z_i] = 0$ and $|Z_i| \leq 1$, so we can use Hoeffding with $a = 1$:

$$\Pr[|S - \alpha n| \geq nt] = \Pr[X \geq nt] \leq 2e^{-\frac{nt^2}{2}}$$

For convenience set $t = \frac{c}{n}$,

$$\Pr[|S - \alpha n| \geq c] \leq 2e^{-\frac{c^2}{2n}} = e^{-\Theta\left(\frac{c^2}{n}\right)}$$

The bound is small for $c \gg \sqrt{n}$, but a constant for $c = \sqrt{n}$. So Hoeffding only tells us that the deviations of S from αn are probably at most $O(\sqrt{n})$. As we have shown previously, S is actually 0 with high probability, so in this example Hoeffding is very weak.

Remark 4.3.11. Generally, the choice of bound depends on the problem we are solving. Hoeffding's inequality is sometimes easier to use than Chernoff. Notice that Chernoff assumes the Z_i are indicators, and has a more complicated expression.

5 Linear algebra

Recitation 4
2020-11-11

5.1 Basics

Matrices and subspaces

Definition 5.1.1. We denote by $M_{n\times m}(\mathbb{F})$ the vector space of all $n \times m$ matrices with elements in a field $\mathbb{F}$, such as $\mathbb{R}, \mathbb{C}$. A matrix $A \in M_{n\times m}(\mathbb{F})$ defines a linear map $T : \mathbb{F}^m \to \mathbb{F}^n$ given by $T(x) = Ax$.

Definition 5.1.2. For every $1 \leq i \leq m$ we denote by $e_i \in \mathbb{F}^m$ the vector with 1 at index i and 0 elsewhere: $(e_i)_i = 1$, $(e_i)_j = 0$ for $j \neq i$. The vectors $e_1, \ldots, e_m$ are called the standard basis.

Note that Ae_i gives column i of A. By definition of matrix multiplication, for any $x \in \mathbb{F}^n$ and $1 \leq k \leq n$ we have:

$$(Ax)_k = \sum_{i=1}^{m} x_i A_{k,i} = \sum_{i=1}^{m} x_j (Ae_i)_k = \left(\sum_{i=1}^{m} x_i Ae_i\right)_k$$

So we can write Ax as a linear sum of columns:

$$Ax = \sum_{i=1}^{m} x_i (Ae_i) \in \mathbb{F}^n$$

Similarly, if $y \in \mathbb{F}^n$ then $y^T A$ is a vector, and a linear combination of the rows of A. Notice that $(y^T A)^T = A^T y$, and we can apply the above.

Definition 5.1.3. The *kernel* of a matrix $A \in M_{n\times m}(\mathbb{F})$ is the subspace

$$\ker A = \{x \in \mathbb{F}^m \mid Ax = \mathbf{0}\}$$

Generally for $b \neq \mathbf{0}$, the set of vectors x such that $Ax = b$ is not a subspace. However, note that if two different vectors x, y satisfy $Ax = Ay = b$, then $A(x - y) = \mathbf{0}$ so $x - y \in \ker A$. Therefore the set of x such that $Ax = b$ can be generated with an initial solution x_0 which satisfies $Ax_0 = b$, and the kernel:

$$\{x \mid Ax = b\} = \{x_0 + v \mid v \in \ker A\} = x_0 + \ker A$$

This is an affine space: a linear space with an additive shift (translation).

Definition 5.1.4. The *image* (*range*) of $A \in M_{n\times m}(\mathbb{F})$ is

$$\mathrm{Im}(A) = \{Ax \mid x \in \mathbb{F}^m\}$$

Equivalently, it is the subspace spanned by the columns of A.

Definition 5.1.5. The *span* of a set of vectors $v_1, \ldots, v_m$ is the set of all their linear combinations.

$$\mathrm{span}(v_1, \ldots, v_m) = \left\{ \sum_{i=1}^{m} \alpha_i v_i \;\middle|\; \alpha_i \in \mathbb{F} \right\}$$

Definition 5.1.6. Vectors $v_1, \ldots, v_n$ are called *linearly independent* if none of them is a linear combination of the others. Equivalently, we can require that any linear combination $\sum_{i=1}^{n} \alpha_i v_i$ is 0 iff all α_i are 0.

Definition 5.1.7. A *basis* of a vector space V is a set of vectors that span v and are linearly independent. All bases of V have the same size, and it is called the *dimension* of V. For example, $\dim \mathbb{F}^n = n$.

Theorem 5.1.8. *For any $A \in M_{n\times m}(\mathbb{F})$ we have*

$$\dim(\mathrm{Im}(A)) = m - \dim(\ker(A))$$

Definition 5.1.9. We define an inner product on $\mathbb{C}^m$:

$$\langle x, y\rangle = \sum_{i=1}^{m} x_i^* y_i$$

where * denotes complex conjugate.

Definition 5.1.10. Let V be an inner space product, and W be a subspace. The *orthogonal complement* of W is the set of all vectors that are orthogonal to W:

$$W^\perp = \{v \in V \mid \langle w, v\rangle = 0 \text{ for all } w \in W\}$$

We have:

$$\dim V = \dim W + \dim W^\perp$$

Definition 5.1.11. The *rank* of a matrix $A \in M_{n\times m}(\mathbb{F})$ is the dimension of its column span, $\mathrm{col}(A)$ as a subspace of $\mathbb{F}^n$, or equivalently, its row span $\mathrm{row}(A)$ as a subspace of $\mathbb{F}^m$.

Proposition 5.1.12. *We have* $\dim \mathrm{col}(A) = \dim \mathrm{row}(A)$.

Proof. We have seen above that the image of A is the span of its columns, so $\dim \operatorname{col}(A) = \dim \operatorname{Im}(A)$. We have $Ax = \mathbf{0}$ iff x is orthogonal to each row (we are assuming $\mathbb{F} = \mathbb{R}$ here for simplicity), so $\ker A = \operatorname{row}(A)^{\perp}$. Hence

$$\begin{aligned}\dim \operatorname{row}(A) &= m - \dim \operatorname{row}(A)^{\perp} \\ &= m - \dim \ker A \\ &= \dim \operatorname{Im}(A) \\ &= \dim \operatorname{col}(A)\end{aligned}$$

□

Lemma 5.1.13. *For $B \in M_{n\times k}(\mathbb{F})$, $C \in M_{k\times m}(\mathbb{F})$, we have*

$$\operatorname{rank}(BC) \leq \min(\operatorname{rank}(B), \operatorname{rank}(C))$$

and so $\operatorname{rank}(BC) = \min(n, k, m)$ *(see exercise 4).*

Theorem 5.1.14. *For $A \in M_{n\times m}(\mathbb{F})$, the number $r = \operatorname{rank}(A)$ is the smallest number such that there exist $B \in M_{n\times r}(\mathbb{F})$, $C \in M_{r\times m}(\mathbb{F})$ with $A = BC$.*

Definition 5.1.15. The *determinant* is defined for square matrices, $\det : M_{n\times n}(\mathbb{F}) \to \mathbb{F}$, as

$$\det(A) = \sum_{\sigma\in S_n} (-1)^{\operatorname{sign}(\sigma)} \sum_{i=1}^{n} A_{i,\sigma(i)}$$

The following properties of the determinant are particularly important:

- $\det(A) = 0$ iff A is singular (not invertible).
- $\det(AB) = \det(A)\det(B)$.
- If A is triangular then the determinant is the product of the main diagonal. $\det(A) = \prod_{i=1}^{n} A_{i,i}$.

Linear transformations

Let U, V be vector spaces with finite dimensions n and m respectively. Let $T : V \to U$ be a linear map. Suppose $v_1, \ldots, v_m \in V$ and $u_1, \ldots, u_n \in U$ are bases. For any $x \in V$ we denote by $[x]_{v_1,\ldots,v_m}$ the vector of coefficients of x's representation in the basis $v_1, \ldots, v_m$. We may write just $[x]$, if the basis is obvious from context. We have:

$$x = \sum_{j=1}^{m} [x]_j v_j$$

This representation of x as a linear combination of the given basis is unique.

Definition 5.1.16. Suppose V is an inner product space. A basis $v_1, \ldots, v_m$ is called *orthonormal* if $\langle v_i, v_j\rangle = 0$ for all $i \neq j$ and $\langle v_i, v_i\rangle = 1$ for all i.

In the case of an orthonormal basis we have

$$\langle x, v_i\rangle = \left\langle \sum_{j=1}^{m} [x]_j v_j, v_i \right\rangle = \sum_{j=1}^{m} [x]_j \langle v_j, v_i\rangle = [x]_i$$

We can think of $[\cdot]_{v_1,\ldots,v_m}$ as an isomorphism from V to $\mathbb{F}^m$, and similarly $[\cdot]_{u_1,\ldots,u_n}$ as an isomorphism from U to $\mathbb{F}^n$. What is the relationship between $[x]$ and $[Tx]$?

Define an $n \times m$ matrix as follows: the i, j entry is $\left([T(v_j)]_{u_1,\ldots,u_n}\right)_i$. That is, the ith coefficient of $T(v_j)$ when represented with the basis $u_1, \ldots, u_n$. We denote this matrix by $[T]^{u_1,\ldots,u_n}_{v_1,\ldots,v_m}$, or $[T]$ when the bases are implied from context. We have:

$$\begin{aligned}
T(x) &= T\left(\sum_{j=1}^{m} [x]_j v_j\right) \\
&= \sum_{j=1}^{m} [x]_j T(v_j) \\
&= \sum_{j=1}^{m} [x]_j \sum_{i=1}^{n} [T]_{i,j} u_i \\
&= \sum_{i=1}^{n} u_i \sum_{j=1}^{m} [T]_{i,j} [x]_j \\
&= \sum_{i=1}^{n} u_i [T(x)]_i u_i
\end{aligned}$$

Therefore $[T(x)] = [T][x]$, or more explicitly

$$[T(x)]_{u_1,\ldots,u_n} = [T]^{u_1,\ldots,u_n}_{v_1,\ldots,v_m} [x]_{v_1,\ldots,v_m}$$

Generally, composition of linear maps behaves like matrix multiplication. Suppose U, V, W have bases $\mathcal{B}_U, \mathcal{B}_V, \mathcal{B}_W$ and we have two linear maps $T : U \to V$ and $S : V \to W$. The composition $ST : U \to W$ is defined as $ST(x) = S(T(x))$. We have:

$$[ST]^{\mathcal{B}_W}_{\mathcal{B}_U} = [S]^{\mathcal{B}_W}_{\mathcal{B}_V} [T]^{\mathcal{B}_V}_{\mathcal{B}_U}$$

Change of basis

Definition 5.1.17. Let $\mathcal{B} = \{v_1, \ldots, v_n\}, \mathcal{B}' = \{v'_1, \ldots, v'_n\}$ be bases of V. Define $[\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}$ to be the change of basis from $\mathcal{B}$ to $\mathcal{B}'$, such that for all $x \in V$,

$$[x]_{\mathcal{B}'} = [\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}} [x]_{\mathcal{B}}$$

The entry at i, j of $[\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}$ is the coefficient of v'_i when v_j is represented with $\mathcal{B}'$:

$$v_j = \sum_{i=1}^{n} \left([\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}\right)_{i,j} v'_i$$

We have $[\mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}} = I$ and

$$I = [\mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}} = [\mathrm{Id} \circ \mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}} = [\mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}'} [\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}$$

so $[\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}$ and $[\mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}'}$ are inverses of each other.

Let $T : V \to V$. To change basis we have

$$[T]^{\mathcal{B}'}_{\mathcal{B}'} = [\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}} [T]^{\mathcal{B}}_{\mathcal{B}} [\mathrm{Id}]^{\mathcal{B}}_{\mathcal{B}'} = P [T]^{\mathcal{B}}_{\mathcal{B}} P^{-1}$$

where $P = [\mathrm{Id}]^{\mathcal{B}'}_{\mathcal{B}}$

Definition 5.1.18. Two square matrices A, B are called *similar* if there is an invertible matrix P with $A = PBP^{-1}$.

Two matrices are similar iff they represent the same linear transformation in two bases (see exercise 4).

Diagonalization

Recitation 5
2020-11-18

Definition 5.1.19. A matrix $A \in M_{n\times n}(\mathbb{F})$ is *diagonalizable* over $\mathbb{F}$ if there is a basis of $\mathbb{F}^n$ where all elements are eigenvectors of A. Equivalently, A is similar to a diagonal matrix, $A = PDP^{-1}$ where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. The columns $v_1, \ldots, v_n$ of P are then the eigenvectors corresponding to the eigenvalues.

Theorem 5.1.20 (Spectral theorem for real symmetric matrices)**.** *For any $A \in M_{n\times n}(\mathbb{R})$ there is an orthonormal basis of eigenvectors. That is, there is an orthogonal matrix U such that $A = UDU^T$ where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ contains the eigenvalues of A.*

The corresponding eigenvectors are the columns $u_1, \ldots, u_n$ of U. Note that generally $u_1, \ldots, u_n$ form an orthonormal basis of $\mathbb{R}^n$ iff $I = U^T U = UU^T$.

We interpret the spectral theorem as a decomposition of A into simple operations. Given some vector z, the operation $z \mapsto Az$ is done as follows:

- Rotate/reflect with an isometry: $z \mapsto U^T z$.
- Stretch each coordinate by the corresponding eigenvalue: $z \mapsto Dz$.
- Reverse the isometry $z \mapsto Uz$.

Another interpretation of the spectral theorem: we decompose the vector space $\mathbb{R}^n$ into orthogonal subspaces, in which A acts by multiplication. Suppose $u_1, \ldots, u_n$ is an orthonormal basis of eigenvectors. A vector x can be written as

$$x = \sum_{i=1}^{n} \langle x, u_i \rangle \, u_i$$

So A multiplies each coefficient:

$$Ax = \sum_{i=1}^{n} \langle x, u_i \rangle \, Au_i = \sum_{i=1}^{n} \langle x, u_i \rangle \, \lambda_i u_i$$

Note that we can write $A = UDU^T$ as

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T = \sum_{i=1}^{n} \lambda_i P_i$$

$P_i = u_i u_i^T$ is the orthogonal projection onto the space spanned by u_i.

Polynomials

Suppose $A = PDP^{-1}$ is diagonalizable. Notice that

$$A^2 = PDP^{-1}PDP^{-1} = PD^2P^{-1}$$

and generally $A^k = PD^kP^{-1}$. Raising a diagonal matrix to k^{th} power is easy, $D^k = \text{diag}(\lambda_1^k, \ldots, \lambda_n^k)$, so this may be an easier way of raising A to a power. For any polynomial $\phi(x) = a_0 + a_1x + \ldots + a_kx^k$ we have:

$$\begin{aligned}
\phi(A) &= a_0I + a_1A + \ldots + a_kA^k \\
&= a_0PP^{-1} + a_1PDP^{-1} + \ldots + a_kPD^kP^{-1} \\
&= P(a_0 + a_1D + \ldots + a_kD^k)P^{-1} \\
&= P\phi(D)P^{-1}
\end{aligned}$$

so this can be generalized from powers to polynomials as well. $\phi(A)$ is diagonalizable using the same basis, and its eigenvalues are $\phi(\lambda_1), \ldots, \phi(\lambda_n)$. This is true for any polynomial, and in particular for the characteristic polynomial $\phi_A(x) = \det(xI - A)$.

Theorem 5.1.21 (Cayley-Hamilton)**.** *Any matrix A satisfies $\phi_A(A) = \mathbf{0}$.*

5.2 Graphs

Definition 5.2.1. The adjacency matrix $A = A_G$ of a graph G on n vertices is an $n \times n$ symmetric matrix where $A_{i,j} = 1$ if i and j are neighbors, 0 otherwise. Usually we consider graphs with no self-edges, so we set $A_{i,i} = 0$.

Lemma 5.2.2. *Let G be d-regular. Then*

1. *d is an eigenvalue of A with corresponding eigenvector $\mathbf{1} = (1, \ldots, 1)$.*
2. *If G is connected and not bipartite, then d has multiplicity 1, and all other eigenvalues satisfy $|\lambda| < d$.*

Proof. For the first part,

$$(A_G\mathbf{1})_j = \sum_{i=1}^{n}(A_G)_{j,i} = \deg(j) = d$$

so $A_G\mathbf{1} = d \cdot \mathbf{1}$. For the second part see exercise 5. □

Lemma 5.2.3. *Let $i, j \in [n]$. Then $(A_G^k)_{i,j}$ (equivalently, $(A_G^k)_{j,i}$) is the number of paths of length k between i and j.*

Proof. By induction on k. For $k = 1$, paths of length 1 are just edges, so this is true. Take some $k > 1$ and assume the claim for $k - 1$. Let $p_k[i, j]$ be the number of paths from i to j of length k. Every such path from i to j has a second-to-last vertex v, such that the v, j are neighbors and the path from i to v has length $k - 1$. So we have:

$$\begin{aligned}
p_k[i, j] &= \sum_{v=1}^{n} \mathbb{1}_{\{v,j \text{ neighbors}\}} p_{k-1}[i, v] \\
&= \sum_{v=1}^{n} (A_G)_{j,v} p_{k-1}[i, v] \\
&= \sum_{v=1}^{n} (A_G)_{j,v} (A_G^{k-1})_{v,i} \\
&= (A_G^k)_{i,j}
\end{aligned}$$

□

Lemma 5.2.4. *Let G be connected, d-regular and not bipartite with n vertices. The number of paths between two vertices a, b of length k is*

$$\frac{d^k}{n}\left(1 + o(1)\right)$$

where n is constant and $k \to \infty$.

Intuitively, each step in a path can take us to any of d vertices, so there are d^k paths of length k. This lemma states that the endpoints of these paths are roughly uniformly distributed across all vertices when k is large (each gets about $\frac{d^k}{n}$ of the possibilities).

Proof. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of A_G with corresponding orthonormal basis of eigenvectors $u_1, \ldots, u_n$, so that $A_G = \sum_{i=1}^n \lambda_i u_i u_i^T$. By a previous lemma, $\lambda_1 = d$ with corresponding eigenvector $\mathbf{1}$, which we normalize as $u_1 = \frac{1}{\sqrt{n}} \cdot \mathbf{1}$. The rest of the eigenvalues are $|\lambda_i| < d$.

$$\begin{aligned}
(A_G^k)_{a,b} &= \left(\sum_{i=1}^n \lambda_i^k u_i u_i^T\right)_{a,b} \\
&= \left(\lambda_1^k u_1 u_1^T\right)_{a,b} + \sum_{i=2}^n \lambda_i^k \left(u_i u_i^T\right)_{a,b} \\
&= \left(d^k \frac{1}{\sqrt{n}} \cdot \mathbf{1} \cdot \frac{1}{\sqrt{n}} \mathbf{1}^T\right)_{a,b} + \sum_{i=2}^n \lambda_i^k \left(u_i u_i^T\right)_{a,b} \\
&= \frac{d^k}{n} + \sum_{i=2}^n \lambda_i^k \left(u_i u_i^T\right)_{a,b} \\
&= d^k \left(\frac{1}{n} + \sum_{i=2}^n \left(\frac{\lambda_i}{d}\right)^k \left(u_i u_i^T\right)_{a,b}\right)
\end{aligned}$$

It remains to show that the sum asymptotically negligible. From $|\lambda_i| < d$ we have $\left(\frac{\lambda_i}{d}\right)^k \to 0$, so the entire sum is $o(1)$. □

5.3 Norms

Recitation 6
2020-11-25

Definition 5.3.1. For any $1 \leq p \leq \infty$ the ℓ_p norm on $\mathbb{R}^n$ is defined as:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$$

where ℓ_∞ is

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

We justify the ∞ notation:

Proposition 5.3.2. *We have*

$$\lim_{p \to \infty} \|x\|_p = \|x\|_\infty$$

Proof. We sandwich $\|x\|_p$. From above:

$$\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p \leq n \cdot \max_{1 \leq i \leq n} |x_i|^p$$

so $\|x\|_p \leq n^{\frac{1}{p}} \|x\|_\infty$ which tends to $\|x\|_\infty$ as $p \to \infty$ because n is constant. From below:

$$\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p \geq \max_{1 \leq i \leq n} |x_i|^p = \|x\|_\infty^p$$

so $\|x\|_p \geq \|x\|_\infty$. □

Matrix norms

Definition 5.3.3. Let $A \in M_{m \times n}(\mathbb{R})$. The *$\ell_p$-to-$\ell_q$ operator norm* is defined as:

$$\|A\|_{p,q} = \max_{\|x\|_p = 1} \|Ax\|_q$$

Note this is the smallest number such that $\|Ax\|_q \leq \|A\|_{p,q} \|x\|_p$ for all $x \in \mathbb{R}^n$.

If $\frac{1}{p} + \frac{1}{q} = 1$ then ℓ_p and ℓ_q are dual norms:

$$\|x\|_p = \max_{\|y\|_q = 1} \langle x, y \rangle$$
$$\|x\|_q = \max_{\|y\|_p = 1} \langle x, y \rangle$$

Lemma 5.3.4. *Let $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then:*

$$\|A\|_{p,q} = \left\|A^T\right\|_{p,q}$$
$$\|A\|_{p,p} = \left\|A^T\right\|_{q,q}$$

and we can generalize this to any two dual norms.

Proof. We have:

$$\begin{aligned}
\|A\|_{p,q} &= \max_{\|x\|_p = 1} \|Ax\|_q \\
&= \max_{\|x\|_p = 1} \max_{\|y\|_p = 1} \langle y, Ax \rangle \\
&= \max_{\|y\|_p = 1} \max_{\|x\|_p = 1} \left\langle A^T y, x \right\rangle \\
&= \max_{\|y\|_p = 1} \left\|A^T y\right\|_q \\
&= \left\|A^T\right\|_{p,q}
\end{aligned}$$

and similarly,

$$\begin{aligned}
\|A\|_{p,p} &= \max_{\|x\|_p=1} \|Ax\|_p \\
&= \max_{\|x\|_p=1} \max_{\|y\|_q=1} \langle y, Ax \rangle \\
&= \max_{\|y\|_q=1} \max_{\|x\|_p=1} \left\langle A^T y, x \right\rangle \\
&= \max_{\|y\|_q=1} \left\| A^T y \right\|_q \\
&= \left\| A^T \right\|_{q,q}
\end{aligned}$$

□

Lemma 5.3.5. *For any $A \in M_{n\times m}(\mathbb{R})$ we have:*

$$\begin{aligned}
\|A\|_{\infty,\infty} &= \max_{1\le i\le n} \sum_{j=1}^{m} |A_{i,j}| \\
\|A\|_{1,1} &= \max_{1\le j\le m} \sum_{i=1}^{n} |A_{i,j}| \\
\|A\|_{1,\infty} &= \max_{\substack{1\le i\le n \\ 1\le k\le m}} |A_{i,j}|
\end{aligned}$$

See exercise 6.

Theorem 5.3.6 (SVD)**.** *Every $A \in M_{n\times m}(\mathbb{R})$ can be decomposed as*

$$A = U\Sigma V^T$$

where U, V are orthogonal matrices of sizes $n \times n$ and $m \times m$ respectively, and Σ is $n \times m$ and contains a diagonal of $r = \min(n, m)$ singular values denoted $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_r$.

Definition 5.3.7. The *Hilbert-Schmidt inner product* for matrices is defined as:

$$\langle A, B\rangle_{HS} = \mathrm{tr}(AB^T) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}B_{i,j}$$

Definition 5.3.8. The *Frobenius norm* of A is defined as:

$$\|A\|_F^2 = \langle A, A\rangle_{HS} = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}^2$$

Lemma 5.3.9. *We have $\|A\|_{2,2} = \sigma_1(A)$ and $\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i(A)^2$. See exercise 6.*

Equivalent formulation: if we denote $\sigma(A) = (\sigma_1(A), \ldots, \sigma_r(A))$, then the ℓ_2-to-ℓ_2 operator norm of A is $\|\sigma(A)\|_\infty$, and the Frobenius norm is $\|\sigma(A)\|_2$.

Definition 5.3.10. For any $1 \le p \le \infty$ the *Schatten-p norm* is defined as:

$$\|A\|_p = \|\sigma(A)\|_p$$

and $\|\sigma(A)\|_1$ is called the *nuclear norm*. We will not prove that these are norms.

Vector norms

We like ℓ_2 because it corresponds to an inner product structure of $\mathbb{R}^n$. From this we have:

- Notion of orthogonality: $\mathbb{R}^n = V + V^\perp$ for any subspace V. So any vector x can be decomposed into its V-component and its $V^\perp$-component, $x = v + v^\perp$, and the distance from x to V is $\left\|v^\perp\right\|_2$.
- Isometries of ℓ_2 can be characterized as orthogonal matrices.
- Results such as the spectral theorem and SVD.

So why care about other ℓ_p norms with $p \neq 2$? In some applications, different notions of distance may be relevant[12]. Since we have intuitive understanding for ℓ_2, a natural question is how to imagine ℓ_p. If a vector has $\|x\|_2 = 1$ and we find that $\|x\|_p$ is small, what do we learn about it?

For $p < 2$, the answer is that such vectors are localized, with relatively sparse coordinates, as we will show.

Lemma 5.3.11. *For any $x \in \mathbb{R}^n$, $\|x\|_p$ is decreasing in $p \in [1, \infty]$ (see exercise 6).*

So for $p < 2$ we have the lower bound $\|x\|_p \geq \|x\|_2$. To get an upper bound, let $x^p = \left(|x_1|^p, \ldots, |x_n|^p\right)$. From Hölder's inequality,

$$\|x\|_p^p = \langle \mathbf{1}, x^p \rangle \leq \|\mathbf{1}\|_{\frac{1}{1-\frac{p}{2}}} \|x^p\|_{\frac{2}{p}}$$

because the dual of $\frac{2}{p}$ is $\frac{1}{1-\frac{p}{2}}$. This is useful because we have

$$\|x^p\|_{\frac{2}{p}} = \left(\sum_{i=1}^n \left(|x_i|^p\right)^{\frac{2}{p}}\right)^{\frac{p}{2}} = \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{p}{2}} = \|x\|_2^p$$

$$\|1\|_{\frac{1}{1-\frac{p}{2}}} = \left(\sum_{i=1}^n 1^{\frac{1}{1-\frac{p}{2}}}\right)^{1-\frac{p}{2}} = n^{1-\frac{p}{2}}$$

Therefore we get the bound

$$\|x\|_p^p \leq n^{1-\frac{p}{2}} \|x\|_2^p$$

Combining both bounds, we have:

$$\|x\|_2 \leq \|x\|_p \leq n^{\frac{1}{p}-\frac{1}{2}} \|x\|_2$$

We can check that the bounds are tight by considering extreme cases. When x has only one non-zero coordinate, we have equality with the lower bound, $\|x\|_2 = \|x\|_p$. When all coordinates of x are $\frac{1}{\sqrt{n}}$ (for normalization), we have $\|x\|_2 = 1$ and equality with the upper bound:

$$\|x\|_p = \left(\sum_{i=1}^n \left(n^{-\frac{1}{2}}\right)^p\right)^{\frac{1}{p}} = \left(n \cdot n^{-\frac{p}{2}}\right)^{\frac{1}{p}} = n^{\frac{1}{p}-\frac{1}{2}}$$

This suggests that small ℓ_p means sparse non-zero coordinates, and large $\|x\|_p$ means we cannot approximate x with a sparse vector.

[12] For example, ℓ_1 is relevant for driving in Manhattan.

Definition 5.3.12. For any s and $x \in \mathbb{R}^n$, let $x_s^* \in \mathbb{R}^n$ be the vector x with all entries 0 except the largest s coordinates.

Then $\|x - x_s^*\|_2$ is the distance of x from a sparse vector that has only s non-zero coordinates.

Theorem 5.3.13. *There is a constant C_p such that for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ we have*

$$\|x - x_s^*\|_2 \leq \frac{C_p}{s^{\frac{1}{p}-\frac{1}{2}}} \|x\|_p$$

See exercise 6.

For example, if $\|x\|_p$ is smaller than $s^{\frac{1}{p}-\frac{1}{2}}$, then x and x_s^* are close, up to some distance C_p that only depends on p. Most of the size of x is concentrated in the largest s coordinates in that case. Even more extreme, when $\|x\|_p = o(n^{\frac{1}{p}-\frac{1}{2}})$ (as a function of n) we get the same result for $s = o(n)$ of the coordinates.

5.4 Variational characterization of eigenvalues

Recitation 7
2020-12-02

Let $A \in M_{n\times n}(\mathbb{R})$ be symmetric. Denote its eigenvalues by $\lambda_1 \geq \ldots \geq \lambda_n$, and let $u_1, \ldots, u_n$ be an orthonormal basis of eigenvectors that correspond to the eigenvalues ($Au_i = \lambda_i u_i$).

Theorem 5.4.1 (Rayleigh-Ritz)**.** *Let $V_{k-1} = \operatorname{span}(u_1, \ldots, u_{k-1})$. Then*

$$\lambda_k = \max_{\substack{\|x\|=1 \\ x \perp V_{k-1}}} x^T A x$$

Let $W_{k+1} = \operatorname{span}(u_{k+1}, \ldots, u_n)$. Then

$$\lambda_k = \min_{\substack{\|x\|=1 \\ x \perp W_{k+1}}} x^T A x$$

In each case, the x that maximizes or minimizes the expression is an eigenvector corresponding to λ_k.

We can use this theorem to calculate eigenvalues and eigenvectors by solving a sequence of optimization problems. We also have a non-sequential alternative:

Theorem 5.4.2 (Courant-Fischer)**.** *We have:*

$$\lambda_k = \min_{\dim F = k-1} \max_{\substack{\|x\|=1 \\ x \perp F}} x^T A x$$

$$\lambda_k = \max_{\dim G = n-k} \min_{\substack{\|x\|=1 \\ x \perp G}} x^T A x$$

Before we prove it we need a lemma:

Lemma 5.4.3. *For any subspaces $V, W \subseteq \mathbb{R}^n$ we have*

$$\dim(V + W) = \dim(V) + \dim(W) - \dim(V \cap W)$$

In particular, if $\dim(V) + \dim(W) > n$ then $\dim(V \cap W) \geq 1$.

Proof. Suppose the dimensions of $V, W, V \cap W$ are k, m, d respectively. Let $\varphi_1, \ldots, \varphi_d$ be a basis of $V \cap W$. Complete it to a basis of V by adding $v_{d+1}, \ldots, v_k$ and to a basis of W by adding $w_{d+1}, \ldots, w_m$. We claim that the union of bases $\varphi_1, \ldots, \varphi_d, v_{d+1}, \ldots, v_k, w_{d+1}, \ldots, w_m$ is a basis of $V + W$. This will be enough, because then

$$\dim(V + W) = d + (k - d) + (m - d) = k + m - d = \dim(V) + \dim(W) - \dim(V \cap W)$$

Clearly the vectors span $V + W$, so we need to show minimality. Suppose that a linear combination of them is $\mathbf{0}$:

$$\sum_{i=1}^{d} \alpha_i \varphi_i + \sum_{i=d+1}^{k} \beta_i v_i + \sum_{i=d+1}^{m} \gamma_i w_i = \mathbf{0}$$

Then write this as

$$\sum_{i=1}^{d} \alpha_i \varphi_i + \sum_{i=d+1}^{k} \beta_i v_i = - \sum_{i=d+1}^{m} \gamma_i w_i$$

On the left we have a vector in V, and on the right we have a vector in W, and they are equal. So this vector must be in $V \cap W$. It is spanned by $\varphi_1, \ldots, \varphi_d$, meaning that the β_i are 0. We get

$$\sum_{i=1}^{d} \alpha_i \varphi_i + \sum_{i=d+1}^{m} \gamma_i w_i = \mathbf{0}$$

This is a linear combination of a basis of W, so all coefficients must be 0. □

Proof of Courant-Fischer. We will show the first equality (the other one is similar). We bound the expression by λ_k from below and from above. For the upper bound, choose $F = V_{k-1}$ and use Rayleigh-Ritz:

$$\min_{\dim F = k-1} \max_{\substack{\|x\|=1 \\ x \perp F}} x^T A x \leq \max_{\substack{\|x\|=1 \\ x \perp V_{k-1}}} x^T A x = \lambda_k$$

For the lower bound, consider some F with $\dim F = k - 1$. Then $\dim F^\perp = n - k + 1$, and we have

$$\dim V_k + \dim F^\perp = n + 1$$

By the lemma, there exists a non-trivial vector $x \in V_k \cap F^\perp$. Take one with $\|x\| = 1$ and write it using the orthonormal basis vectors in V_k:

$$x = \sum_{i=1}^{k} \langle x, u_i \rangle u_i$$

Then:

$$x^T A x = \sum_{i=1}^{k} \lambda_i \langle x, u_i \rangle^2 \geq \sum_{i=1}^{k} \lambda_k \langle x, u_i \rangle^2 = \lambda_k$$

Therefore for any such F, the maximum of $x^T A x$ is bounded from below by λ_k. □

Eigenvalue interlacing

We show a corollary of Courant-Fischer. Suppose now that $B = A_{-i,-i}$ is A without the ith row and column. We assume that $A \in M_{n \times n}(\mathbb{R})$ is symmetric, so $B \in M_{(n-1)\times(n-1)}(\mathbb{R})$ is symmetric as well. There is a relationship between the eigenvalues of A and B:

Theorem 5.4.4. *For every $1 \leq k \leq n-1$ we have*

$$\lambda_{k+1}(A) \leq \lambda_k(B) \leq \lambda_k(A)$$

Proof. To be able to apply Courant-Fischer, we relate the expressions $x^T Ax$ and $x^T Bx$. For $x \in \mathbb{R}^{n-1}$, we can add a 0 in the ith position to make a vector in $\mathbb{R}^n$ that would give the same result for A and B. Formally, define

$$\widehat{x} = (x_1, \ldots, x_{i-1}, 0, x_i, \ldots, x_{n-1}) \in \mathbb{R}^n$$

Then we have:

$$\widehat{x}^T A\widehat{x} = x^T Bx$$

because in both sides the ith row and column of A is not counted.

For any subspace $F \subseteq \mathbb{R}^{n-1}$ we can define

$$\widehat{F} = \{\widehat{x} \mid x \in F\} \subseteq \mathbb{R}^n$$

Now, write Courant-Fischer as follows. Instead of choosing F and x orthogonal to it, we choose $W = F^\perp$ and x in it:

$$\lambda_{k+1}(A) = \min_{\substack{\dim F = k \\ F \subseteq \mathbb{R}^n}} \max_{\substack{\|x\|=1 \\ x \perp F}} x^T Ax = \min_{\substack{\dim W = n-k \\ W \subseteq \mathbb{R}^n}} \max_{\substack{\|x\|=1 \\ x \in W}} x^T Ax$$

Suppose that instead of minimizing over all $W \subseteq \mathbb{R}^n$ with dimension $n-k$, we restrict ourselves to such W that have the form $W = \widehat{V}$ for some $V \subseteq \mathbb{R}^{n-1}$ (that is, we only consider subspaces with i^{th} coordinate 0). This gives an upper bound, since we have fewer opportunities to minimize. Then:

$$\lambda_{k+1}(A) \leq \min_{\substack{\dim V = n-k \\ V \subseteq \mathbb{R}^{n-1}}} \max_{\substack{\|x\|=1 \\ x \in V}} \widehat{x}^T A\widehat{x} = \min_{\substack{\dim V = n-k \\ V \subseteq \mathbb{R}^{n-1}}} \max_{\substack{\|x\|=1 \\ x \in V}} x^T Bx = \lambda_k(B)$$

where we used Courant-Fischer again, this time for $\lambda_k(B)$.

The other direction is similar, using the max-min formulation:

$$\lambda_k(A) = \max_{\substack{\dim F = n-k \\ F \subseteq \mathbb{R}^n}} \min_{\substack{\|x\|=1 \\ x \perp F}} x^T Ax = \max_{\substack{\dim W = k \\ W \subseteq \mathbb{R}^n}} \min_{\substack{\|x\|=1 \\ x \in W}} x^T Ax$$

Maximizing only over W of the form $\widehat{V}$ gives a lower bound:

$$\lambda_k(A) \geq \max_{\substack{\dim V = k \\ V \subseteq \mathbb{R}^{n-1}}} \min_{\substack{\|x\|=1 \\ x \in V}} \widehat{x}^T A\widehat{x} = \max_{\substack{\dim V = k \\ V \subseteq \mathbb{R}^{n-1}}} \min_{\substack{\|x\|=1 \\ x \in V}} x^T Bx = \lambda_k(B)$$

□

5.5 Markov chains

Recitation 8
2020-12-09

Definition 5.5.1. Let S be a finite set of states. A Markov chain of values in S is a stochastic process $X_1, X_2, \ldots$ such that for all $t = 1, 2, \ldots$:

$$\Pr[X_{t+1} = v \mid X_t = u] = P_{u,v}$$

where P is some transition matrix. Usually we will consider $S = [n]$.

Let x_t be a column vector with the distribution of probabilities at time t:

$$x_t = (\Pr[X_t = 1], \ldots, \Pr[X_t = n])$$

Notice that $x_{t+1}^T = x_t^T P$ and so $x_t^T = x_0^T P^t$, where x_0 is some initial distribution.

Theorem 5.5.2. *Let P be a transition matrix of a Markov chain which is irreducible and aperiodic. Then:*

1. *The spectral radius of P is 1, and the eigenvalue 1 has multiplicity 1.*
2. *There is a left eigenvector π such that $\pi^T P = \pi^T$ and π is a distribution (non-negative and sums to 1). It is called the stationary distribution.*
3. *We have:*
$$\lim_{t\to\infty} P^t = \mathbf{1}\pi^T$$

Definition 5.5.3. An irreducible, aperiodic Markov chain is called *ergodic*.

In ergodic chains, the initial distribution does not matter for the asymptotic behavior: for any x_0,

$$\lim_{t\to\infty} x_t^T = \lim_{t\to\infty} x_0^T P^t = x_0^T \mathbf{1}\pi^T = \pi^T$$

We are interested in measuring the speed of convergence, so the following definition will be useful:

Definition 5.5.4. The *total variation distance* between distributions p, q is:

$$d_{TV}(p,q) = \frac{1}{2}\|p-q\|_1 = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i|$$

Proposition 5.5.5. *For a set $S \subseteq [n]$, let $p(S) = \sum_{i\in S} p_i$. Then we have (see exercise 8):*

$$d_{TV}(p,q) = \max_S (p(S) - q(S))$$

Proposition 5.5.6. *The sequence $d_{TV}(x_t, \pi)$ is non-increasing (see exercise 8).*

Definition 5.5.7. Let x_0 be some initial distribution. We denote by $\tau(\varepsilon \mid x_0)$ the earliest time t such that the distribution is within ε of π:

$$\tau(\varepsilon \mid x_0) = \min\{t \mid d_{TV}(x_t, \pi) \le \varepsilon\}$$

Remark 5.5.8. The $\varepsilon \mid x_0$ notation is not necessarily standard.

Definition 5.5.9. The *mixing time* of a chain is defined, with parameter ε, as the earliest time that guarantees to get within ε of π no matter which distribution we start with:

$$\tau(\varepsilon) = \max_{x_0} \tau(\varepsilon \mid x_0)$$

Proposition 5.5.10. *$\tau(\varepsilon)$ is finite for any $\varepsilon > 0$, and a "worst" initial distribution is one of $e_1, \ldots, e_n$:*

$$\tau(\varepsilon) = \max_{1\le i\le n} \tau(\varepsilon \mid e_i)$$

See exercise 8.

Random walks

Definition 5.5.11. A simple random walk (SRW) on a graph G is a Markov chain where the states are vertices, and X_{t+1} is the result of moving from X_t to one of its neighbors with uniform probability. The transition matrix is:

$$P_{i,j} = \begin{cases} \frac{1}{\deg(i)} & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

Proposition 5.5.12. *The stationary distribution π of P is:*

$$\pi_i = \frac{\deg(i)}{2\,|E|}$$

For the rest of this recitation, assume that G is d-regular, connected and non-bipartite. The transition matrix is $P = \frac{1}{d} A_G$ where A_G is the adjacency matrix. The stationary distribution is uniform:

Proposition 5.5.13. *If the transition matrix P is symmetric, then the uniform distribution $\pi = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ is stationary (see exercise 8).*

Definition 5.5.14. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the spectrum of P. The *spectral gap* of G is defined as:

$$\gamma = 1 - \max\left(\lambda_2, |\lambda_n|\right)$$

We are assuming G is connected, so $1 = \lambda_1 > \lambda_2$, and non-bipartite so $\lambda_n > -1$. Therefore $\gamma > 0$.

Theorem 5.5.15. *Let G be d-regular, connected and non-bipartite. Then the mixing time of the SRW is:*

$$\tau(\varepsilon) \leq \frac{\log\left(\frac{\sqrt{n}}{2\varepsilon}\right)}{\gamma}$$

Suppose ε is fixed. We say that the SRW is *rapidly mixing* if $\tau(\varepsilon)$ is at most polylogarithmic in n, that is, it is bounded above by a polynomial in $\log n$. In the above expression, if $\gamma = \frac{1}{\text{polylog}(n)}$ then we have rapid mixing.

Proof. We show that for any initial distribution x_0, we have $d_{TV}(x_t, \pi) \leq \varepsilon$ for the claimed t. We have:

$$d_{TV}(x_t, \pi) = \frac{1}{2} \sum_{i=1}^{n} |(x_t)_i - \pi_i|$$

We can write this sum as an inner product of the vector $x_t - \pi$ with a vector of ± 1s, where the sign of each element matches the corresponding element of $x_t - \pi$. Then we can use Cauchy-Schwarz:

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^{n} \left| \left(x_0^T P^t\right)_i - \pi_i \right| &= \frac{1}{2} \left\langle x_t - \pi, (\pm 1, \ldots, \pm 1) \right\rangle \\ &\leq \frac{1}{2} \left\| x_t - \pi \right\|_2 \left\| (\pm 1, \ldots, \pm 1) \right\|_2 \\ &= \frac{\sqrt{n}}{2} \left\| x_t - \pi \right\|_2 \end{aligned}$$

This is useful because with the ℓ_2 norm we can use the spectral theorem. P is real and symmetric, so there is an orthonormal basis of eigenvectors $u_1, \ldots, u_n$ corresponding to $\lambda_1 \geq \ldots \geq \lambda_n$. By above, u_1 is the constant vector so

$$u_1 = \frac{1}{\sqrt{n}} \mathbf{1}$$

Write x_0 in the orthonormal basis:

$$\begin{aligned}
x_0 &= \sum_{i=1}^{n} \langle x_0, u_i \rangle u_i \\
&= \langle x_0, u_1 \rangle u_1 + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i \\
&= \left\langle x_0, \frac{1}{\sqrt{n}}\mathbf{1} \right\rangle \cdot \frac{1}{\sqrt{n}}\mathbf{1} + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i \\
&= \langle x_0, \mathbf{1} \rangle \frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i \\
&= \pi + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i
\end{aligned}$$

where the last transition is because $\langle x_0, \mathbf{1} \rangle = 1$ (since x_0 is a distribution its elements sum to 1) and $\pi = \frac{1}{n}\mathbf{1}$.

So we have:

$$\begin{aligned}
x_t^T &= x_0^T P^t \\
&= \left(\pi^T + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i^T \right) P^t \\
&= \pi^T P^t + \sum_{i=2}^{n} \langle x_0, u_i \rangle u_i^T P^t \\
&= \pi^T + \sum_{i=2}^{n} \langle x_0, u_i \rangle \lambda_i^t u_i^T
\end{aligned}$$

where the last transition is because $\pi^T = \pi^T P$ and $u_i^T P^t = (P^t u_i)^T = \lambda_i^t u_i^T$. Hence:

$$\begin{aligned}
\|x_t - \pi\|_2^2 &= \left\| \pi + \sum_{i=2}^{n} \langle x_0, u_i \rangle \lambda_i^t u_i - \pi \right\|_2^2 \\
&= \left\| \sum_{i=2}^{n} \langle x_0, u_i \rangle \lambda_i^t u_i \right\|_2^2 \\
&= \sum_{i=2}^{n} |\lambda_i|^{2t} \langle x_0, u_i \rangle^2
\end{aligned}$$

By definition of the spectral gap, we have $|\lambda_i| \leq 1 - \gamma$, so this sum is at most

$$\sum_{i=2}^{n} (1-\gamma)^{2t} \langle x_0, u_i \rangle^2 = (1-\gamma)^{2t} \sum_{i=2}^{n} \langle x_0, u_i \rangle^2 = (1-\gamma)^{2t} \|x_0\|_2^2$$

We have $\|x_0\|_2 \leq \|x_0\|_1 = 1$, so this is bounded by $(1-\gamma)^{2t}$, which in turn is bounded by $e^{-\gamma 2t}$. Overall we get:

$$\|x_t - \pi\|_2 \leq e^{-\gamma t}$$

which bounds the total variation distance:

$$d_{TV}(x_t, \pi) \leq \frac{\sqrt{n}}{2} \|x_t - \pi\|_2 \leq \frac{\sqrt{n}}{2} e^{-\gamma t}$$

We are guaranteed that $d_{TV}(x_t, \pi) \leq \varepsilon$ when this is at most ε:

$$\begin{aligned}
\frac{\sqrt{n}}{2} e^{-\gamma t} &\leq \varepsilon \\
\ln \frac{\sqrt{n}}{2} - \gamma t &\leq \ln \varepsilon \\
\ln \frac{\sqrt{n}}{2} - \ln \varepsilon &\leq \gamma t \\
\frac{\ln \frac{\sqrt{n}}{2\varepsilon}}{\gamma} &\leq t
\end{aligned}$$

□

Random walk on a cycle

Let C_n be a cycle on n vertices. Assume n is odd, so it is not bipartite. The graph is 2-regular, and its adjacency matrix is:

$$A = \begin{pmatrix} & 1 & & & 1 \\ 1 & & 1 & & \\ & 1 & & 1 & \\ & & 1 & & 1 \\ 1 & & & 1 & \end{pmatrix}$$

We can write it as the sum of two inverse matrices:

$$S = \begin{pmatrix} & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \\ 1 & & & & \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} & & & & 1 \\ 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \end{pmatrix}$$

Notice that S performs a cyclic shift, mapping $(x_0, \ldots, x_{n-1})$ to $(x_1, \ldots, x_{n-1}, x_0)$ and S^{-1} does the opposite shift, so they are indeed inverses.

For every $\ell \in \{0, \ldots, n-1\}$ define the vector f_ℓ such that for $k = 0, \ldots, n-1$,

$$(f_\ell)_k = \frac{1}{\sqrt{n}} e^{2\pi i \frac{k\ell}{n}}$$

The vectors f_ℓ are an orthonormal basis of $\mathbb{C}^n$. We have:

$$(Sf_\ell)_k = \frac{1}{\sqrt{n}} e^{2\pi i \frac{(k+1 \bmod n)\ell}{n}} = e^{2\pi i \frac{\ell}{n}} (f_\ell)_k$$

So the f_ℓ are eigenvectors of S and S^{-1}, with eigenvalues $e^{2\pi i \frac{\ell}{n}}$ and $e^{-2\pi i \frac{\ell}{n}}$ respectively.

Let P be the Markov chain transition matrix of an SRW on C_n. We have:

$$P = \frac{1}{2} A$$

and so,

$$P(f_\ell) = \frac{1}{2}\left(S + S^{-1}\right) f_\ell = \frac{1}{2}\left(e^{2\pi i \frac{\ell}{n}} + e^{-2\pi i \frac{\ell}{n}}\right) f_\ell = \cos\left(\frac{2\pi\ell}{n}\right) f_\ell$$

This is the spectrum of P, for $\ell = 0, \ldots, n-1$. We have:

$$\lambda_2 = \cos\left(\frac{2\pi}{n}\right), \quad \lambda_n = \cos\left(\frac{2\pi \left\lfloor \frac{n}{2} \right\rfloor}{n}\right) = \cos\left(\pi - \frac{\pi}{n}\right)$$

By expanding cosine to a Taylor series around 0 and around π, we can see that the spectral gap is

$$\gamma = \Theta\left(\frac{1}{n^2}\right)$$

as $n \to \infty$. By the above theorem, the mixing time is

$$\tau(\varepsilon) = O\left(\frac{\log\left(\frac{\sqrt{n}}{2\varepsilon}\right)}{\gamma}\right) = O\left(\frac{\log n}{\gamma}\right) = O\left(n^2 \log n\right)$$

for any fixed ε.

We show that this bound is tight up to a $\log n$ factor, specifically that $\tau\left(\frac{1}{4}\right) = \Omega(n^2)$. The idea is to consider the half of the cycle farthest from the starting point (between $\frac{1}{4}n$ and $\frac{3}{4}n$), and show that it takes $\Omega(n^2)$ steps until those vertices get close to the uniform distribution.

We have:

$$d_{TV}(x_t, \pi) = \max_S \left(\pi(S) - x_t(S)\right)$$

Choosing a specific S gives a lower bound. When S is the far half cycle, it contains $> \frac{n}{2}$ vertices[13], so $\pi(S) > \frac{n}{2} \cdot \frac{1}{n} = \frac{1}{2}$. It remains to bound $x_t(S)$ from above.

Let $Z_1, Z_2, \ldots$ be the step sizes, each selected uniformly and independently from $\{1, -1\}$. Define:

$$X_t = \sum_{i=1}^{t} Z_i$$

The position on the cycle at time t is $X_t \bmod n$. By Hoeffding,

$$\Pr\left[|X_t| \geq \frac{n}{4}\right] \leq 2e^{-c\frac{\left(\frac{n}{4}\right)^2}{t}}$$

for some constant $c > 0$. The right hand side increases with t, and it reaches $\frac{1}{4}$ when t is on the order of magnitude of n^2. For any smaller t, the probability of being in S is $\leq \frac{1}{4}$ so we have:

$$x_t(S) \leq \frac{1}{4}$$

and the total variation distance is bounded from below:

$$d_{TV}(x_t, \pi) \geq \pi(S) - x_t(S) > \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Random walk on a cube

Recitation 9
2020-12-16

Let $H_n = \{0,1\}^n$ be the vertices of an n-dimensional unit cube. Consider the following lazy walk: on every step, choose $1 \leq i \leq n$ uniformly, then set the coordinate x_i to be 0 or 1 uniformly. There

[13] n is odd, we can round up so that $|S| = \frac{n+1}{2}$.

is probability $\frac{1}{2}$ to stay in place. This is an ergodic Markov chain, and the stationary distribution is uniform.

We study the mixing time. Note that $|H_n| = 2^n$, so rapid mixing here means polynomial in n (polylogarithmic in 2^n). After every coordinate has been chosen at least once, the position is random, which reduces the question to the coupon collector problem. This happens after $n \log n + O(n)$ time, so this is an upper bound on the mixing time.

Formally, we want t such that $d_{TV}(x_t, \pi) \leq \varepsilon$ for all initial distributions x_0. Let T be the earliest time after which all coordinates have been chosen at least once. X_t is uniform for $t \geq T$. For any vertex v we have, by total probability:

$$\begin{aligned}(x_t)_v &= \Pr[X_t = v] \\ &= \Pr[X_t = v \mid T \leq t] \Pr[T \leq t] + \Pr[X_t = v \mid T > t] \Pr[T > t] \\ &= \pi_v \Pr[T \leq t] + \Pr[X_t = v \mid T > t] \Pr[T > t]\end{aligned}$$

Write $\pi_v = \pi_v \Pr[T \leq t] + \pi_v \Pr[T > t]$ so:

$$\begin{aligned}|(x_t)_v - \pi_v| &= |\pi_v \Pr[T \leq t] + \Pr[X_t = v \mid T > t] \Pr[T > t] - (\pi_v \Pr[T \leq t] + \pi_v \Pr[T > t])| \\ &= |\Pr[X_t = v \mid T > t] \Pr[T > t] - \pi_v \Pr[T > t]| \\ &= |(\Pr[X_t = v \mid T > t] - \pi_v) \Pr[T > t]|\end{aligned}$$

Let q_t be a vector defined by $(q_t)_v = \Pr[X_t = v \mid T > t]$. Then:

$$|(x_t)_v - \pi_v| = |(q_v - \pi)_v| \Pr[T > t]$$

Summing we get:

$$d_{TV}(x_t, \pi) = \frac{1}{2} \|x_t - \pi\|_1 = \frac{1}{2} \|q - \pi\|_1 \Pr[T > t] \leq \Pr[T > t]$$

Therefore $\tau(\varepsilon)$ is bounded above by the T such that $\Pr[T > t] \leq \varepsilon$, which is $n \log n + O(n)$ when ε is a fixed constant.

Ergodic theorem

Let $X_1, X_2, \ldots$ be independent and identically distributed with values in some finite set S. Let $f : S \to \mathbb{R}$ be a function, so that $f(X_1), f(X_2), \ldots$ is a sequence of bounded real numbers. By the weak law of large numbers, the average converges to the expectation:

$$\lim_{T \to \infty} \Pr\left[\left|\frac{1}{T} \sum_{t=1}^{T} f(X_t) - \mathbb{E}[f(X)]\right| \geq \varepsilon\right] = 0$$

Suppose $X_1, X_2, \ldots$ are states in an ergodic Markov process, which has stationary distribution π. After a long time τ, the variable X_τ should be distributed by π, and similarly for $X_{2\tau}, X_{3\tau}, \ldots$. So it makes sense that the average value of f on these variables is the expectation of f on π:

$$\frac{1}{N} \sum_{n=1}^{N} f(X_{n\tau}) \to \mathbb{E}_{X \sim \pi}[f(X)]$$

The same should hold for a shifted sequence $X_{\tau+k}, X_{2\tau+k}, \ldots$ so that

$$\begin{aligned}\frac{1}{N\tau} \sum_{t=1}^{N\tau} f(X_t) &= \frac{1}{\tau} \sum_{k=0}^{\tau-1} \frac{1}{N} \sum_{n=0}^{N-1} f(X_{\tau n+k+1}) \\ &\approx \frac{1}{\tau} \sum_{k=0}^{\tau-1} \mathbb{E}_{X \sim \pi}[f(X)] \\ &= \mathbb{E}_{X \sim \pi}[f(X)]\end{aligned}$$

This was informal. We will not prove the theorem:

Theorem 5.5.16 (Ergodic theorem for Markov chains)**.** *Let $X_1, X_2, \ldots$ be samples from an ergodic Markov chain, with set of states S and stationary distribution π. Let $f : S \to \mathbb{R}$ and define*

$$A_T = \frac{1}{T} \sum_{t=1}^{T} f(X_t)$$

Then for every $\varepsilon > 0$ we have:

$$\lim_{T \to \infty} \Pr\left[|A_T - \mathbb{E}_{X \sim \pi}[f(X)]| \geq \varepsilon\right] = 0$$

Monte Carlo

An application of Markov chains is to sample from a distribution which is too complicated for direct methods. The idea is to find a Markov chain for which this distribution is stationary, and where it's easy to compute X_{t+1} from X_t. We succeed if the chain converges quickly enough to the desired distribution. We will see an example.

Let G be a graph with n vertices, and suppose the maximal degree is at most Δ. A coloring of G in q colors is a labeling of the vertices with q possible colors such that neighbors have different colors. Formally, it is a function $c : V \to \{1, \ldots, q\}$ such that if $u \sim v$ then $c(u) \neq c(v)$. Such a coloring does not always exist. If $q \geq \Delta + 1$, then we can color the vertices arbitrarily and we cannot fail, since there are at most Δ constraints on each vertex.

Suppose we want a random q-coloring, uniformly chosen from the set Ω of all possible q-colorings. It is not obvious how to do this, since finding colorings may be difficult, and there may be too many to consider.

Define a Markov chain on Ω. A state is a coloring. We move to the next state as follows: choose a random vertex uniformly, and choose a random color uniformly. If the vertex can legally be changed to that color, this is the new state; otherwise we stay in place.

It may be unfeasible to write down the transition matrix, but notice that moving from a given state to the next is computationally easy. If it is ergodic and rapidly mixing, then this gives us a random coloring selected according to the stationary distribution. The stationary distribution is uniform, like we want (because the transition matrix is symmetric, see exercise 8). All we have to do is find one initial coloring, then perform the local modifications several times.

It can be shown that the chain is rapidly mixing for $q \geq 2\Delta + 1$, in time $O(n \log n)$, but we will not do it here. We will show the following:

Lemma 5.5.17. *The Markov chain is ergodic for $q \geq \Delta + 2$.*

Proof. The chain is aperiodic because there is positive probability to stay in place. Therefore there is a path with length 1 from every state to itself, and so the GCD of the path lengths is 1.

To show it is irreducible, we need to prove any coloring can be turned to any other coloring by modifying single vertices. Let c, c' be two colorings, and we want to turn c to c'. Suppose they disagree on a vertex v: $c(v) \neq c'(v)$. If $c(v)$ can be changed to $c'(v)$ legally, we do so. Otherwise, it has some neighbor w with $c(w) = c'(v)$. The vertex w is constrained by at most Δ vertices, and we want to change it to a different color; since $q \geq \Delta + 2$, there is at least one remaining option, so we change $c(w)$ to it. Do this for each problematic neighbor of v, until $c(v)$ can be changed to $c'(v)$.

In any case we decreased the number of vertices that c, c' disagree on by at least 1 (we certainly fixed $c(v)$, and we may have fixed some $c(w)$ by chance. We did not ruin any $c(w)$, because we only modify it if $c(w) = c'(v)$, in which case $c'(w) \neq c(w)$). We repeat this until we reach c'. $\square$

6 Optimization

6.1 Basics

Recitation 10
2020-12-23

Definition 6.1.1. A hyperplane is defined by a vector $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$:

$$H(a,b) = \left\{ x \,\middle|\, a^T x = b \right\}$$

Definition 6.1.2. A *half-space* is defined by a vector $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$:

$$HS(a,b) = \left\{ x \,\middle|\, a^T x \leq b \right\}$$

Definition 6.1.3. A *polyhedron* is a set $P \subseteq \mathbb{R}^n$ defined by finitely many linear inequalities. Formally, for $a_1, \ldots, a_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$:

$$P = \left\{ x \,\middle|\, a_i^T x \leq b_i \text{ for all } 1 \leq i \leq n \right\}$$

Equivalently we can define:

$$P = \bigcap_{i=1}^{m} HS(a_i, b_i)$$

Let $A \in M_{m \times n}(\mathbb{R})$ be a matrix with rows $a_1, \ldots, a_m$, and let $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$. We can also define:

$$P = \{ x \mid Ax \leq b \}$$

where the inequality is per-coordinate.

Definition 6.1.4. A *polytope* is a bounded polyhedron.

Proposition 6.1.5. *A polyhedron P is convex. That is, if $x, y \in P$ then P contains the entire segment between them: $\alpha x + (1-\alpha)y \in P$ for all $0 \leq \alpha \leq 1$.*

Proof. We need to show that $\alpha x + (1-\alpha)y$ satisfies the constraints of P. We have:

$$a_i^T(\alpha x + (1-\alpha)y) = \alpha a_i^T x + (1-\alpha) a_i^T y$$

Since $a_i^T x \leq b_i$ and $a_i^T y \leq b_i$, this is at most

$$\alpha b_i + (1-\alpha) b_i = b_i$$

□

Example 6.1.6. Consider a triangle $T \subseteq \mathbb{R}^2$ with vertices in $(0,0), (0,1), (1,0)$. The points in T are the ones satisfying:

$$\begin{aligned} x &\geq 0 \\ y &\geq 0 \\ x + y &\leq 1 \end{aligned}$$

So we can write:

$$T = \left\{ (x,y) \,\middle|\, A \begin{pmatrix} x \\ y \end{pmatrix} \leq b \right\}$$

where the matrix A and vector b are:

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Definition 6.1.7. Let P be a polyhedron and let $H(a,b)$ be a hyperplane. We say $H(a,b)$ is a *supporting hyperplane* of P if the following conditions hold:

- They intersect: $P \cap H(a,b) \neq \emptyset$.
- P is entirely on one side: $P \subseteq HS(a,b)$.

Remark 6.1.8. If P has the constraint $\langle a_i, x\rangle \leq b_i$, it doesn't necessarily mean that $H(a_i, b_i)$ is supporting. In the above example, we can add to T the constraint $x \geq -1$. Nothing would change, and the hyperplane $x = -1$ does not intersect T.

Definition 6.1.9. A *face* of P is the intersection of P with a supporting hyperplane.

Remark 6.1.10. P itself is a face, since $P = P \cap H(0,0)$. Notice that $H(0,0) = \mathbb{R}^n$.

Definition 6.1.11. A *vertex* of P is a face with exactly one point.

Example 6.1.12. The vertices of T are $(0,0), (1,0), (0,1)$, the 1-dimensional faces are $x = 0, y = 0, x + y = 1$. and T itself is a 2-dimensional face.

Lemma 6.1.13. *A vector $v \in P$ is a vertex if and only if there is some $c \in \mathbb{R}^n$ such that v is the unique maximizer of $f(x) = c^T x$ over the domain $x \in P$.*

Remark 6.1.14. We could replace maximization with minimization, since maximizing $c^T x$ is the same as minimizing $(-c)^T x$.

Remark 6.1.15. Consider a linear program where we want to maximize $c^T x$ with constraints $Ax \leq b$. The lemma implies that if there is a unique maximum, then it is at a vertex of P.

Proof. Suppose v is a vertex. Then there is a supporting hyperplane $H(a,b)$ with $a^T v = b$, and all other points in P have strict inequality $a^T x < b$, by definition. Hence v maximizes the function $a^T x$.

Suppose $v \in P$ is a unique maximizer of $c^T x$ for some c, and consider the hyperplane $H(c, c^T v)$. We have $v \in P \cap H(c, c^T v)$ and all other vertices $x \in P$ have $c^T x < c^T v$ by definition. So it is a supporting hyperplane with $P \cap H(c, c^T v) = \{v\}$, hence v is a vertex. □

Proposition 6.1.16. *The vertices of the cube $C_n = [0,1]^n$ are $\{0,1\}^n$.*

Proof. We show equality by two-sided containment:

- $\{0,1\}^n \subseteq \text{vertices}(C_n)$: let $v \in \{0,1\}^n$. Define a vector c to have 1s in the same indices as v, and -1s elsewhere:
$$c_i = \begin{cases} 1 & v_i = 1 \\ -1 & v_i = 0 \end{cases}$$
We claim v uniquely maximizes $c^T x$. Notice that $c^T v$ is the number of 1s in v, and for $x \neq v$ the quantity $c^T x$ decreases by 1 for every index where $x_i \neq v_i$.
$$\begin{aligned} c^T x &= \sum_{i \mid v_i = 1} x_i - \sum_{i \mid v_i = 0} x_i \\ &= |\{i \mid x_i = v_i = 1\}| - |i \mid x_i \neq v_i = 0| \\ &< |\{i \mid v_i = 1\}| \\ &= c^T v \end{aligned}$$

- vertices$(C_n) \subseteq \{0,1\}^n$: we show that for any c, some $v \in \{0,1\}^n$ is a maximizer of $c^T x$. This means there cannot be unique maximizers outside $\{0,1\}^n$, so all vertices are in $\{0,1\}^n$. Take v such that $v_i = 1$ if $c_i \geq 0$ and $v_i = 0$ otherwise. Then $c^T v$ is the sum of positive coordinates of c. Any other vector can only decrease this quantity, by including negative coordinates or excluding positive ones. Formally, for any $x \in C_n$ and $c \in \mathbb{R}^n$:

$$c^T x = \sum_{i|c_i \geq 0} c_i x_i + \sum_{i|c_i < 0} c_i x_i \leq \sum_{i|c_i \geq 0} c_i x_i \leq \sum_{i|c_i \geq 0} c_i = c^T v$$

□

6.2 Convexity

Recitation 11
2020-12-30

Definition 6.2.1. A set $C \subseteq \mathbb{R}^d$ is called *convex* if for every pair of points $x, y \in C$, the set also contains the entire segment $[x, y]$. That is, $\alpha x + (1-\alpha) y \in C$ for all $0 \leq \alpha \leq 1$.

Definition 6.2.2. Let $x_1, \ldots, x_m \in \mathbb{R}^d$. The *convex hull* is the set of all linear combinations with non-negative coefficients $\alpha_1, \ldots, \alpha_m$ that sum to 1:

$$\operatorname{conv}(x_1, \ldots, x_m) = \left\{ \sum_{i=1}^m \alpha_i x_i \;\middle|\; \alpha_i \geq 0,\ \sum_{i=1}^m \alpha_i = 1 \right\}$$

Lemma 6.2.3. *Let $P = \operatorname{conv}(x_1, \ldots, x_m)$. Then P is a polytope and* vertices$(P) \subseteq \{x_1, \ldots, x_m\}$ *(see exercise 10).*

Theorem 6.2.4 (Radon)**.** *Let $S \subseteq \mathbb{R}^d$ be a set of size $|S| \geq d+2$. Then it can be partitioned into subsets of points with overlapping convex hulls. That is, there exists $T \subseteq S$ such that*

$$\operatorname{conv}(T) \cap \operatorname{conv}(S \backslash T) \neq \emptyset$$

Proof. We will assume $|S| = d+2$. If S is larger, we can include the extra points in T (this can only add points to $\operatorname{conv}(T)$, so the intersection does not get smaller).

Let $S = \{y_1, \ldots, y_{d+2}\}$. We claim there is a non-zero vector $\gamma = (\gamma_1, \ldots, \gamma_{d+2})$ such that:

$$\sum_{i=1}^{d+2} \gamma_i y_i = \mathbf{0}$$
$$\sum_{i=1}^{d+2} \gamma_i = 0$$

The second sum is a linear equation in $\gamma_1, \ldots, \gamma_{d+2}$. The first sum implies d linear equations (one per coordinate). We have more variables ($d+2$) than equations ($d+1$), so the kernel is non-trivial and such γ exists.

Let I be the set of indices i for which $\gamma_i > 0$. We claim that I partitions the vectors in the way we want, that is, we can take $T = \{y_i\}_{i \in I}$. Notice that T and $S \backslash T$ are both non-empty: γ is not zero, and its sum is zero, so it has both positive and negative coordinates. We have:

$$\mathbf{0} = \sum_{i=1}^{d+2} \gamma_i y_i = \sum_{i \in I} \gamma_i y_i + \sum_{i \notin I} \gamma_i y_i$$

So:

$$\sum_{i \in I} \gamma_i y_i = \sum_{i \notin I} (-\gamma_i) y_i$$

There exist positive coordinates in γ, so we can divide both sides by $\sum_{i' \in I} \gamma_{i'} > 0$:

$$\sum_{i \in I} \left(\frac{\gamma_i}{\sum_{i' \in I} \gamma_{i'}} \right) y_i = \sum_{i \notin I} \left(\frac{(-\gamma_i)}{\sum_{i' \in I} \gamma_{i'}} \right) y_i$$

In both sides the coefficients are non-negative and sum to 1. So on the left we have an element in $\operatorname{conv}(T)$, and on the right we have an element in $\operatorname{conv}(S \backslash T)$, hence they intersect. □

Theorem 6.2.5 (Carathéodory)**.** *Let $S \subseteq \mathbb{R}^d$ and $x \in \operatorname{conv}(S)$. Then there exist $d+1$ points $y_1, \ldots, y_{d+1} \in S$ such that $y \in \operatorname{conv}(y_1, \ldots, y_{d+1})$.*

Proof. Since $x \in \operatorname{conv}(S)$ there are $y_1, \ldots, y_m \in S$, $m \geq 1$, such that $x = \sum_{i=1}^m \alpha_i y_i$ with $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i = 1$. Assume $\alpha_i > 0$, since otherwise we can omit y_i from the sum.

If $m \leq d+1$, we are done. From now assume $m \geq d+2$. We show that one of the y_i can be omitted, so that only $m-1$ points are required. By repeating the argument we can decrease m to $d+1$.

Similarly to the previous proof, there is a non-zero vector $\gamma = (\gamma_1, \ldots, \gamma_m)$ such that

$$\begin{aligned} \sum_{i=1}^m \gamma_i y_i &= \mathbf{0} \\ \sum_{i=1}^m \gamma_i &= 0 \end{aligned}$$

In this case there are $d+1$ equations and $m > d+1$ variables. Let $\gamma_j \neq 0$ be some non-zero coordinate. We try to write x as a convex combination without y_j, and we will conclude there is a specific j for which this can be done.

We have:

$$\begin{aligned} \sum_{i=1}^m \gamma_i y_i &= \mathbf{0} \\ \gamma_j y_j + \sum_{i \neq j} \gamma_i y_i &= \mathbf{0} \\ y_j &= \sum_{i \neq j} \left(-\frac{\gamma_i}{\gamma_j} \right) y_i \end{aligned}$$

So we can substitute this into the expression for x:

$$\begin{aligned} x &= \sum_{i=1}^m \alpha_i y_i \\ &= \sum_{i \neq j} \alpha_i y_i + \alpha_j y_j \\ &= \sum_{i \neq j} \alpha_i y_i + \alpha_j \sum_{i \neq j} \left(-\frac{\gamma_i}{\gamma_j} \right) y_i \\ &= \sum_{i \neq j} \left(\alpha_i - \frac{\gamma_i}{\gamma_j} \alpha_j \right) y_i \end{aligned}$$

This is a linear combination without y_j. If we can show it is a convex combination, we are done. The sum of coefficients is:

$$\sum_{i \neq j} \left(\alpha_i - \frac{\gamma_i}{\gamma_j} \alpha_j \right) = \left(\sum_{i \neq j} \alpha_i \right) - \left(\frac{\alpha_j}{\gamma_j} \sum_{i \neq j} \gamma_i \right) = (1 - \alpha_j) - \left(\frac{\alpha_j}{\gamma_j} \cdot (-\gamma_j) \right) = 1$$

We also want the coefficients to be non-negative:

$$\begin{aligned} 0 &\leq \alpha_i - \frac{\gamma_i}{\gamma_j} \alpha_j \\ 0 &\leq 1 - \frac{\gamma_i}{\gamma_j} \cdot \frac{\alpha_j}{\alpha_i} \\ \frac{\gamma_i / \alpha_i}{\gamma_j / \alpha_j} &\leq 1 \end{aligned}$$

This holds for j such that $\left| \gamma_j / \alpha_j \right|$ is maximal. □

Theorem 6.2.6 (Helly)**.** *Let $A_1, \ldots, A_m \subseteq \mathbb{R}^d$ be a family of $m \geq d+1$ convex sets. Suppose that every $d+1$ of these sets intersect, that is:*

$$A_{i_1} \cap \ldots \cap A_{i_{d+1}} \neq \emptyset$$

for all choices of $i_1, \ldots, i_{d+1} \in [m]$. Then the intersection of the entire family is non-empty:

$$A_1 \cap \ldots \cap A_m \neq \emptyset$$

The proof uses Radon's theorem (see exercise 11).