

Lecture 1

Lecturer: Nati Linial

Scribe: Robby Lampert

Last Update: 19 Feb 2009 9:39 a.m.

This lecture is about the probabilistic method. We demonstrate how using probabilistic arguments, we can derive theorems that do not seem to have a clear relation to probability theory. We start with basic definitions from probability, and proceed with examples of the probabilistic method.

1 Basic Definitions

- A *discrete probability space* is a (finite or countable) set Ω equipped with a non-negative function $\Pr : \Omega \rightarrow \mathbb{R}_+$ of total sum one, i.e., $\sum_{\omega \in \Omega} \Pr \omega = 1$. The set Ω is called also the *sample space* and the function $\Pr \cdot$ is called *probability function*.
- An *event* is a subset $A \subseteq \Omega$ and its *probability* is defined as $\Pr A = \sum_{\omega \in A} \Pr \omega$.
- For two events $A, B \subseteq \Omega$, we define the *conditional probability* of A given B to be $\Pr A|B \stackrel{def}{=} \Pr A \cap B / \Pr B$. Conditioning on the event B induces the sample space B and its probability function is defined to be $\Pr \cdot|B$.
- Two events $A, B \subseteq \Omega$ are called *independent* if $\Pr A \cap B = \Pr A \cdot \Pr B$, or equivalently $\Pr A|B = \Pr A$ or $\Pr B|A = \Pr B$. Intuitively two events are independent if they do not affect each other, i.e., the occurrence of one event does not change the probability of the occurrence of the other.
- A *discrete random variable* X is a real function $X : \Omega \rightarrow \mathbb{R}$ with a finite range $\{r_1, \dots, r_n\}$. Given a set $A \subseteq \{r_1, \dots, r_n\}$ we denote by $\Pr X \in A$ the probability of the set $\{\omega \in \Omega : X(\omega) \in A\}$.
- Two discrete random variables X, Y are called *independent* if for every r, s in their respective ranges, we have $\Pr X = r \wedge Y = s = \Pr X = r \cdot \Pr Y = s$.
- The *expectation* of a random variable X is the weighted average of its values, i.e., $\mathbb{E}[X] = \sum_{\omega \in \Omega} \Pr \omega \cdot X(\omega)$. The expectation of a random variable $\mathbb{E}[X]$ is also called its mean and is often denoted by μ . It is easy to see that the expectation is linear, i.e., for any two constants $\alpha, \beta \in \mathbb{R}$, and two random variables X, Y , the expectation of the random variable $\alpha X + \beta Y$ is $\alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$.
- The **variance** of a random variable X is defined to be $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. The variance quantifies how scattered the values of X 's are around their average, and this is expressed in terms of the l_2 norm.

The union bound follows from our basic definitions, and is stated without a proof:

Proposition 1 (union bound) For a family of events $\{A_i\}_{i \in I}$ in a probability space Ω there holds $\Pr \cup_{i \in I} A_i \leq \sum_{i \in I} \Pr A_i$

we also state two basic inequalities which imply that under appropriate assumptions a random variable tends not to deviate much from its expectation:

Theorem 2 Markov inequality

For a non-negative random variable $X \geq 0$ and a constant $c \geq 1$ there holds

$$\Pr X > c \cdot \mathbb{E}[X] \leq \frac{1}{c}$$

Theorem 3 Chebyshev inequality

For a random variable X with a mean $\mathbb{E}[X] = \mu$ and a variance $\text{Var}(X) = \sigma^2$, and for any constant c , we have

$$\Pr |X - \mu| > (c \cdot \sigma) \leq 1/c^2$$

2 Cycles in random permutations

A *permutation* is a bijection from a finite set onto itself. We define $S_n = \{\pi : [n] \rightarrow [n]\}$ the set of permutations over $[n] = \{1, 2, \dots, n\}$. A *fixed point* of a permutation π is an i such that $\pi(i) = i$. We use probabilistic arguments to show that the average number of fixed points of a permutation is one.

Proposition 4 Let $X : S_n \rightarrow \mathbb{N}$ be a function that counts the number of fixed points in a permutation π , i.e., $X(\pi) = |\{i : \pi(i) = i\}|$. Assume that π is sampled from S_n according to the uniform distribution, i.e., $\Pr \pi = 1/n!$ for every $\pi \in S_n$, then $\mathbb{E}[X] = 1$.

Proof: To calculate $\mathbb{E}[X]$ we use the linearity of expectation by representing X as a sum of random variables. For $1 \leq i \leq n$ define the random variable X_i as

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

This kind of random variable is called a *characteristic random variable* or *indicator random variable*. It uniquely corresponds to a certain event. Namely, its value is 1 exactly when the event occurs, and is 0 otherwise. Here, X_i corresponds to the event in which i is a fixed point of π . Since $X(\pi) = \sum_{i=1}^n X_i(\pi)$ its expectation equals to $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i]$. Now, for all $1 \leq i \leq n$,

$$\begin{aligned} \mathbb{E}[X_i] &= \frac{1}{n!} \cdot |\{\pi \in S_n : \pi(i) = i\}| \quad (\text{because } X_i \text{ indicates the event " } i \text{ is a fixed point"}) \\ &= \frac{1}{n!} \cdot |S_{n-1}| \quad (\text{the number of permutations over } [n] \setminus \{i\}) \\ &= \frac{(n-1)!}{n!} = \frac{1}{n}. \end{aligned}$$

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

A cycle of length 2 in a permutation π is a pair $1 \leq i \neq j \leq n$ such that $\pi(i) = j$ and $\pi(j) = i$. Similarly, a cycle of length k is an ordered k -tuple of distinct elements $(i_1, \dots, i_k) \in [n]$ such that $\pi(i_j) = i_{j+1}$ for all $1 \leq j \leq k-1$ and $\pi(i_k) = i_1$. Set \mathcal{F}_k be the family of all possible k -cycles in $[n]$, then its size is $|\mathcal{F}_k| = \binom{n}{k}(k-1)!$ since every set of k elements has exactly $(k-1)!$ cyclic ordering.

Proposition 5 Let $X : S_n \rightarrow \mathbb{N}$ be a function that counts the number of k -cycles in a permutation. Assume that π is sampled from S_n according to the uniform distribution, then $\mathbb{E}[X] = 1/k$.

Proof: There is an equivalence relation on ordered k -tuples where $(i_1, \dots, i_k) \sim (i_2, \dots, i_k, i_1) \sim (i_3, \dots, i_k, i_1, i_2) \sim \dots$ which represents the same cycle. Let \mathcal{F}_k be the family of all possible k -cycles under the above equivalence relation. For every $(i_1, \dots, i_k) \in \mathcal{F}_k$ let $X_{i_1, \dots, i_k} : S_n \rightarrow \{0, 1\}$ be the indicator random variable corresponding to the event $\pi(i_j) = i_{j+1}$ for $j = 1, \dots, k-1$ and $\pi(i_k) = i_1$. For every $\pi \in S_n$, the random variable $X(\pi)$ is represented as $\sum_{i_1, \dots, i_k \in \mathcal{F}_k} X_{i_1, \dots, i_k}(\pi)$. Since X_{i_1, \dots, i_k} is a characteristic random variable, its mean is the probability for a random permutation to have the k -cycle (i_1, \dots, i_k) . For clarity, we maintain the above equivalence relation among ordered k -tuples. Therefore,

$$\mathbb{E}[X_{i_1, \dots, i_k}] = \frac{|S_{n-k}|}{|S_n|} = \frac{(n-k)!}{n!},$$

and

$$\mathbb{E}[X] = \sum_{i_1, \dots, i_k \in \mathcal{F}_k} \frac{(n-k)!}{n!} = |\mathcal{F}_k| \frac{(n-k)!}{n!} = \binom{n}{k} (k-1)! \frac{(n-k)!}{n!} = \frac{1}{k},$$

■

3 Ramsey numbers and the probabilistic method

Another application for the probabilistic method yields a lower bound for Ramsey numbers.

Theorem 6 (Ramsey Theorem) For every $k, l \in \mathbb{N}$ there exists an integer $R = R(k, l)$ such that if we color the edges of K_R (the complete graph on R vertices) blue and red, it will necessarily contain a blue K_k or a red K_l .

The standard proof of Ramsey's Theorem actually provides an upper bound on R , namely $R(k, l) \leq \binom{k+l-2}{k-1}$ which for $k = l$ yields: $R(k, k) \leq 2^{2k} = 4^k$. It implies that in every graph with n vertices there exists a clique or an anti-clique of size $\geq \frac{1}{2} \log_2 n$.

We show that there are graphs on n vertices without a clique or an anti-clique of size $> 2 \log_2 n$, thus proving the lower bound $R(k, k) \geq 2^{k/2}$. The key idea is to consider the set of all graphs on n vertices as a probability space Ω of size $2^{\binom{n}{2}}$. We now need to specify the probability of every element in Ω , namely, every n vertex graph. An important example of such a probability space is $G(n, p)$ - where each edge is present with probability p and absent with probability $1 - p$, independently among edges. If we set $e(G)$ to be the number of edges in a graph G then the probability of the graph G is $\Pr G = p^{e(G)} \cdot (1-p)^{\binom{n}{2}-e(G)}$. Alternatively, we can describe a sampling of a graph from $G(n, p)$ as a process in which for every $1 \leq i < j \leq n$, independently, we put the edge (i, j) in the graph with probability p . Note that in $G(n, \frac{1}{2})$ all the graphs have the same probability, i.e., $G(n, \frac{1}{2})$ is the

probability space consisting of all n -vertex graphs under the uniform distribution .

To prove the desired lower bound we define two random variables:

$$\begin{aligned} X(G) &= && \text{the number of } k\text{-cliques in } G. \\ Y(G) &= && \text{the number of } k\text{-anti-cliques in } G. \end{aligned}$$

We show that for the choice of $k = 2 \log_2 n$ we get $\mathbb{E}[X] < 1/2$ and $\mathbb{E}[Y] < 1/2$, which implies $\mathbb{E}[X + Y] < 1$. It follows that the random variable $X + Y$ takes the value of 0. In other words, there are graphs in the probability space for which $X + Y = 0$, that is, they have neither a clique nor an anti-clique of size k . We represent X as a sum of characteristic random variables X_S , for all $S \subseteq [n]$ of size k , where for each such S ,

$$X_S = \begin{cases} 1 & S \text{ is a clique} \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$\begin{aligned} \mathbb{E}[X] &= \sum_S \mathbb{E}[X_S] = \sum_S \Pr S \text{ is a clique} = \sum_S \left(\frac{1}{2}\right)^{\binom{k}{2}} = \binom{n}{k} 2^{-\frac{k(k-1)}{2}} \\ &\leq \left(\frac{ne}{k}\right)^k \left(2^{-\frac{k-1}{2}}\right)^k = \left(\frac{ne}{k2^{\frac{k-1}{2}}}\right)^k \end{aligned}$$

where the inequality is derived using Stirling approximation for factorials. If we take $k > 2 \log_2 n$ then $\mathbb{E}[X] < 1/2$.

Lecture 2

Lecturer: Nati Linial

Scribe: Ofer Neiman

Last Update: 19 Feb 2009 9:39 a.m.

In the previous lecture we saw how the linearity of expectation is used in certain situations. In this lecture we establish another important technique, known as the second moment method. We use the second moment of a random variable, i.e., $\mathbb{E}X^2$, to analyze the existence of K_4 , a clique with 4 vertices, in a random graph in $G(n, p)$. We also present the *birthday paradox* and the *coupon collector* problem, and demonstrate another use of the second moment method.

4 The notion of almost surely

We often deal with problems in discrete probability where there is often an underlying parameter n that grows to infinity. If some event A holds with probability $1 - \epsilon = 1 - \epsilon(n)$ where $\epsilon(n)$ tends to zero as $n \rightarrow \infty$, we say that A holds "almost surely" often abbreviated to *a.s.* It is true that the exact phrase should in fact be "asymptotically almost surely". However, this abuse of language is already being used by almost everyone..., so we use it as well.

5 Second Moment Method

Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable. We are interested in the distribution of the values of X , however, only very rarely can learn all the information about the distribution. Usually it is easy to calculate first moments, e.g., $\mathbb{E}X, \mathbb{E}X^2$, and we show that already from this limited information we can derive interesting conclusions.

In the previous lecture we saw that if X is a non-negative random variable and $\mathbb{E}X = o(1)$ then $X = 0$ almost always. We use the second moment method to derive a similar statement: If $\mathbb{E}X \rightarrow \infty$ and $\text{Var}(X)$ is not too big then $\Pr[X = 0] = o(1)$. In words X is almost surely strictly positive. To show this we use Chebyshev's inequality: $\Pr[|X - \mathbb{E}X| > c\sigma] \leq \frac{1}{c^2}$, and the fact that $\{\omega : X(\omega) = 0\} \subseteq \{\omega : |X(\omega) - \mathbb{E}X| \leq \mathbb{E}X\}$ to conclude

$$\Pr[X = 0] \leq \frac{\text{Var}(X)}{\mathbb{E}^2 X} \quad (1)$$

6 Existence of K_4 in $G(n, p)$

As mentioned already K_4 is a clique of 4 vertices. $G(n, p)$ is a probability space of all the graphs with $|V| = n$ vertices such that every edge is present with probability p independent of the other edges. We use the second moment method to prove the following:

Theorem 7 *The threshold function for the existence of K_4 in $G(n, p)$ is $p = n^{-2/3}$. Namely,*

1. If $p = o(n^{-2/3})$ then G almost surely does not contain K_4 .
2. If $p = \omega(n^{-2/3})$ then G almost surely contains K_4 .

Proof:

1. The first part is proved using the linearity of the expectation: Let $X : G(n, p) \rightarrow \mathbb{N}_+$ be a random variable counting the number of copies of K_4 in G . We express it as a sum of simple variables:

$$X = \sum_{|T|=4, T \subseteq [n]} X_T,$$

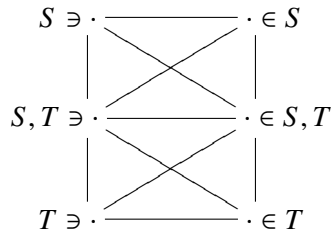
where X_T is an indicator to the event that T is a clique. $\mathbb{E}X = \binom{n}{4} p^6 \leq n^4 p^6 = o(1)$. Consequently $X = 0$ almost surely. In words, a random graph drawn from the distribution $G(n, p)$ has, almost surely, no K_4 .

2. To prove the second part we note that this time $\mathbb{E}X > \left(\frac{n-3}{4!}\right) p^6 \rightarrow \infty$, thus to prove that $\Pr[X = 0] = o(1)$ we use the estimation in equation (1). To estimate $\text{Var}(X)$ we compute $\mathbb{E}X^2 - \mathbb{E}^2X$ while noting that for every two sets $T \neq S \subset [n]$ of size 4 the random variables X_T, X_S are independent unless $|S \cap T| = 2, 3$:

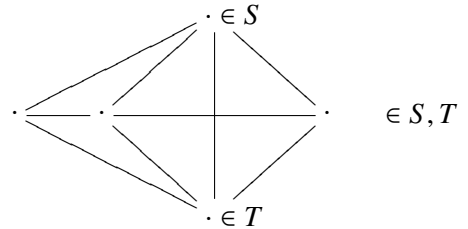
$$\text{Var}(X) = \sum_{|T|=4, T \subseteq [n]} (\mathbb{E}X_T - \mathbb{E}^2X_T) + 2 \underbrace{\sum_{S \neq T} (\mathbb{E}[X_S X_T] - \mathbb{E}X_S \mathbb{E}X_T)}_{\text{vanishes unless } |S \cap T|=2,3}$$

The expression $\mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$ is called the *covariance* of X, Y and denoted by $\text{cov}(X, Y)$. Whenever $X = Y$ it reduces to the variance of X .

- The term $\sum_{|T|=4, T \subseteq [n]} (\mathbb{E}X_T - \mathbb{E}^2X_T)$ is clearly less than $\mathbb{E}X = \sum_{|T|=4} \mathbb{E}X_T$
- $\sum_{|S \cap T|=2} \mathbb{E}[X_S X_T] = \binom{n}{4} \cdot \binom{4}{2} \cdot \binom{n-4}{2} p^{11}$, where $\binom{n}{4}$ are the number of possibilities to choose the set T , $\binom{4}{2}$ are the possibilities to choose the two elements in $S \cap T$ from the 4 elements of T , $\binom{n-4}{2}$ are the possible ways to choose the remaining two elements of S and p^{11} is the probability that all 11 edges appear between the vertices in T and S .



- Similarly $\sum_{|S \cap T|=3} \mathbb{E}[X_S X_T] = \binom{n}{4} \cdot \binom{4}{3} \cdot \binom{n-4}{1} p^9$.



- $\mathbb{E}X_T = \mathbb{E}X_S = p^6$.

Therefore substituting in the formula result in $\text{Var}(X) \leq \mathbb{E}X + n^6 p^{11} + n^5 p^9 = o(n^8 p^{12}) = o(\mathbb{E}^2 X)$.

■

7 Balls and Bins problems

We start throwing balls into n bins and ask:

1. **The birthday paradox:** What is the time at which we expect to see first two balls in the same bin? The answer is about \sqrt{n} .
2. **The coupons collector problem:** What is the first time at which we expect to see no empty bins? Answer: $n \log n$.

7.1 The Birthday Paradox

We first ask for the probability that we throw r balls into n bins and each of them falls into a different bin. This probability equals

$$1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{r-1}{n}\right) \simeq \exp\left(-\sum_{i=1}^{r-1} \frac{i}{n}\right) = \exp\left(-\frac{r(r-1)}{2n}\right)$$

It follows that if $r = o(\sqrt{n})$, then this probability is $1 - o(1)$. On the other hand if $r \gg \sqrt{n}$, then this probability is $o(1)$.

Let us also look at the following closely related question: Given $A, B \subseteq [n]$ two randomly chosen sets of given coordinates $|A| = k, |B| = l$, What is the probability that $A \cap B = \emptyset$? Without loss of generality we may fix A and consider only B as chosen randomly. Then,

$$\begin{aligned} \Pr(A \cap B = \emptyset) &= \binom{n-k}{n} \cdots \binom{n-k-l+1}{n} \\ &= \left(1 - \frac{k}{n}\right) \cdots \left(1 - \frac{k+l-1}{n}\right) \\ &\simeq \exp\left(-\sum_{i=0}^{l-1} \frac{k+i}{n}\right) = \exp\left(-\frac{(2k+l-1)l}{2n}\right) = \exp\left(-\frac{kl(1+o(1))}{n}\right) \end{aligned}$$

- If $kl \gg n$ ($m = \omega(n)$), then $\Pr(A \cap B = \emptyset) \leq o(1)$, and so $A \cap B = \emptyset$ almost surely.
- If $kl \ll n$ ($m = o(n)$), then $\Pr(A \cap B = \emptyset) \geq 1 - o(1)$, and so $A \cap B \neq \emptyset$ almost surely.
- If $kl \sim n$ ($m = \theta(n \log n)$), then both $\Pr(A \cap B = \emptyset)$ and $\Pr(A \cap B \neq \emptyset)$ are bounded away from zero.

7.2 The Coupons Collector

Assume we throw m balls into n bins, then what is the probability that none of the bins remains empty?

Claim 8

1. If $m \gg n \log n$ then almost surely no bin will be empty.
2. The expected time till no bin is empty is $n \log_e n + O(n)$.

Proof:

1. Let X be the random variable that counts the number of empty bins. We write $X = \sum_{i=1}^n X_i$, where X_i is the characteristic random variable of the event "the i -th bin is empty". The random variables X_i are identically distributed and their expectation is $\mathbb{E}[X_i] = \Pr(X_i = 1) = \left(1 - \frac{1}{n}\right)^m$. By the linearity of the expectation of $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = n \left(1 - \frac{1}{n}\right)^m$. If $m = n(\log n + \Delta_n)$, where $\lim_{n \rightarrow \infty} \Delta_n = \infty$, then

$$\mathbb{E}X = n \left(1 - \frac{1}{n}\right)^{n(\log n + \Delta_n)} \simeq n e^{(-\log n - \Delta_n)} = e^{-\Delta_n} \rightarrow 0.$$

Consequently the expected number of empty bins is $o(1)$, and thus almost surely, there is no empty bin.

2. We throw balls into the bins until no bin is empty. Let Y be the random variable that counts the time until no bin remains empty. We write $Y = \sum_{i=1, \dots, n} Y_i$, where Y_i is the time since we had $i - 1$ nonempty bins, until we have i nonempty bins. What is the expectation of Y_{t+1} ? The situation is that t bins are nonempty and $n - t$ are empty. We perform an experiment with $p = \frac{n-t}{n}$ probability for success (success is throwing a ball into an empty bin). This is a geometric random variable with probability of success p . Therefore the expected time until success is $\frac{1}{p} = \frac{n}{n-t}$. Thus,

$$\mathbb{E}Y = \sum_{i=1}^n \frac{n}{n-i+1} = n \sum_{j=1}^n \frac{1}{j} = n(\log_e n + O(1))$$

■

We would like to make a more precise statement, namely, that at time exceeding $n \log_e n + O(n)$ it is almost certain that no bin remains empty. This, however, requires some second moment considerations. How can we prove a statement like this? Markov's inequality allows us to deduce only weaker claims, such as

$$\Pr\{\text{It takes } > 10n \log_e n \text{ until no bin is empty}\} < \frac{1}{10}$$

It is understandable that this method yields only such unsatisfactory bounds, since Markov's inequality relies only on the crudest information concerning a random variable, namely, its expectation. To derive the conclusion we want, we need to incorporate the variance as well. We first make the following simple observation:

Claim 9 If X_1, \dots, X_n are random variables that are pairwise independent, then $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$

We next observe that if Z is a geometric random variable with probability of success p , then $\text{Var}(Z) = \frac{1-p}{p^2} \leq \frac{1}{p^2}$. Therefore in our case $p_i = \frac{n-i+1}{n}$ and the geometric random variable X_i with parameter p_i is the waiting time for the i 'th bin to fill, and $X = \sum_{i=1}^n X_i$. Consequently

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) \leq \sum_{i=1}^n \left(\frac{n}{n-i+1} \right)^2 \leq n^2 \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2 n^2}{6}$$

Corollary 10 $\Pr\{X > nH_n + c \frac{\pi n}{\sqrt{6}}\} \leq \frac{1}{c^2}$

We can pick any $c = c_n$ that tends to infinity with n , e.g. $c_n = \sqrt{\log n}$, to conclude

Theorem 11 *With almost certainty, the time until no empty bins remain is $(1 + o(1))n \log n$.*

Lecture 3

Lecturer: Nati Linial

Scribe: Benny Kimelfeld

Last Update: 19 Feb 2009 9:39 a.m.

8 Measure Concentration

Consider a probability space (Ω, \Pr) . When we analyze a random variable $X : \Omega \rightarrow \mathbb{R}$, we are often interested in knowing, in addition to the expectation, how *concentrated* X is. Namely, the likelihood of X attaining values that are far from the mean. More formally, we would like to estimate values of the form $\Pr[X > \mathbb{E}[X] + \Delta]$, $\Pr[X < \mathbb{E}[X] - \Delta]$ and/or $\Pr[|X - \mathbb{E}[X]| > \Delta]$. Theorems that provide such estimations (e.g., upper bounds) are called *measure-concentration* theorems. Note that we have already seen two such theorems, namely:

- **Markov's inequality:** If X is nonnegative and $c > 0$ then $\Pr[X \geq c\mathbb{E}[X]] \leq \frac{1}{c}$.
- **Chebyshev's inequality:** If $c > 0$ then $\Pr[|X - \mathbb{E}[X]| \geq c] \leq \frac{\text{Var}[X]}{c^2}$.

One can show that the two inequalities above are *tight*, that is, there are examples in which equalities hold. However, the literature contains many other, much stronger bounds (e.g., Chernoff's, Hoeffding's, etc.) that apply to special cases of interest, namely, random variables which satisfy some additional conditions. Such a bound is studied in the next section.

9 Case Study - Sum of n coin flips / Random walk on the line

Assume a walk (e.g. of a particle) on the discrete line such that

- At time $t = 0$ it starts on $x = 0$.
- At each steps it moves either to the left (-1) or to the right (+1) with equal probability ($\frac{1}{2}$)

Such a walk is known in the literature as random walk or drunkard's walk

There are several question that might interest us:

1. What is the (distribution over) position of the particle at time $t = N$ (for large N)?
2. How many times will the particle visit $x = 0$ until time $t = N$ (for large N)?
3. What is the farthest point the particle will visit until time $t = N$ (for large N)?
4. Similar questions when the walk is in higher dimensions or the walk is biased (p for left, $(1-p)$ for right when $p \neq \frac{1}{2}$).

We will define the RV X_1, X_2, \dots s.t $\Pr[X_1 = 1] = \Pr[X_1 = -1] = \frac{1}{2}$. (One RV for each step) and $X = \sum_{i=1}^N X_i$ for the position of the particle after N steps. We are interested in the distribution of X .

Claim 12 $\Pr[X \geq a] \leq e^{-\frac{a^2}{2N}}$

Proof: In this case we know the distribution of X . The event $[X = t]$ occurs only if the particle moved l steps to the left, $N - l$ steps to the right, and $N - 2l = N - l - l = t$. Therefore

$$\Pr[X = t] = \frac{\binom{N}{\frac{N-t}{2}}}{2^N}$$

$$\begin{aligned} \Pr[X \geq a] &= \sum_{t>a} \Pr[X = t] \\ &= \sum_{t>a} \frac{\binom{N}{\frac{N-t}{2}}}{2^N} \\ &= \frac{1}{2^N} \sum_{t>a} \binom{N}{\frac{N-t}{2}} \end{aligned}$$

But it is hard to approximate this expression and hence we will try to find a more convenient approximation.

$$\begin{aligned} \text{For any } t > 0: \Pr[X \geq a] &= \Pr[e^{tX} \geq e^{ta}] \\ &\leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} && \text{[Markov Inequality]} \\ &= e^{-ta} \mathbb{E}[e^{t \sum_{i=1}^N X_i}] \\ &= e^{-ta} \prod_{i=1}^N \mathbb{E}[e^{tX_i}] && \text{[The } X_i \text{ are independent RV]} \\ \mathbb{E}[e^{tX_i}] &= \frac{1}{2} (e^t + e^{-t}) \\ &= \frac{1}{2} \left(\sum_{i=0}^{\infty} \frac{t^i}{i!} + \sum_{i=0}^{\infty} \frac{(-t)^i}{i!} \right) && \text{[Taylor expansion]} \\ &= \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \\ &= \sum_{k=0}^{\infty} \frac{(t^2)^k}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} \frac{(t^2)^k}{2^k k!} \\ &= \sum_{k=0}^{\infty} \frac{\left(\frac{t^2}{2}\right)^k}{k!} \\ &= e^{\frac{t^2}{2}} && \text{[Taylor expansion]} \\ \Pr[X \geq a] &\leq e^{-ta} \prod_{i=1}^N \mathbb{E}[e^{tX_i}] \\ &\leq e^{-ta} \prod_{i=1}^N \left(e^{\frac{t^2}{2}} \right) \\ &= e^{-\frac{Nt^2}{2} - ta} \end{aligned}$$

Minimizing the expression $e^{-\frac{Nt^2}{2} - ta}$ by choosing $t = \frac{a}{N}$ will get us the bound $\Pr[X \geq a] \leq e^{-\frac{a^2}{2N}}$ ■

By symmetry we also have $\Pr[X \leq -a] \leq e^{-\frac{a^2}{2N}}$ and

$$\begin{aligned} \Pr[|X| \geq a] &\leq 2e^{-\frac{a^2}{2N}} \\ \Pr[|X| \geq k\sqrt{N}] &\leq 2e^{-0.5k^2}. \end{aligned}$$

For instance after N steps the particle will be within distance from its starting point of at most $1.7\sqrt{N}$ with probability $\geq 1 - 2e^{-0.5(1.7)^2} \approx 0.53$ and within distance from its starting point of at most $5\sqrt{N}$ with probability $\geq 1 - 2e^{-25} \approx 0.999$.

A useful combinatorial lemma for approximating expressions like we saw above:

Lemma 13 $\forall \alpha \in (0, 1) \binom{n}{\alpha n} = 2^{n(H(\alpha)+o(1))}$ (assuming αn is natural) when the entropy function is defined to be $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$ (See figure 13)

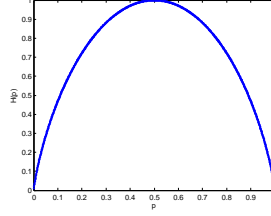


Figure 1: Entropy $H(p)$

Proof of Lemma $1 = (\alpha + (1 - \alpha))^n = \sum_{k=0}^n \binom{n}{k} \alpha^k (1 - \alpha)^{n-k}$

- The maximal addend in the sum is for $k = \alpha n$

$$\begin{aligned} & \binom{n}{j} \alpha^j (1 - \alpha)^{n-j} \geq \binom{n}{j-1} \alpha^{j-1} (1 - \alpha)^{n-j+1} \\ \iff & \frac{n!}{j!(n-j)!} \alpha^j (1 - \alpha)^{n-j} \geq \frac{n!}{(j-1)!(n-j+1)!} \alpha^{j-1} (1 - \alpha)^{n-j+1} \quad \text{Therefore (if } \alpha n \text{ is natural)} \\ \iff & (n - j + 1) \alpha \geq j(1 - \alpha) \\ \iff & j \leq (n + 1) \alpha = \alpha n + \alpha \end{aligned}$$

and $\alpha \in (0, 1)$ the maximal element is for $k = \alpha n$.

$$\begin{aligned} \Rightarrow & \binom{n}{\alpha n} \alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n} \leq 1 \leq (n + 1) \binom{n}{\alpha n} \alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n} \\ & ((n + 1) \alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n})^{-1} \leq \binom{n}{\alpha n} \leq (\alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n})^{-1} \\ & \frac{1}{n+1} (\alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n})^{-1} \leq \binom{n}{\alpha n} \leq (\alpha^{\alpha n} (1 - \alpha)^{1 - \alpha n})^{-1} \\ & \alpha^\alpha (1 - \alpha)^{(1-\alpha)} = 2^{-H(\alpha)} \\ & \frac{2^{nH(\alpha)}}{n+1} \leq \binom{n}{\alpha n} \leq 2^{nH(\alpha)} \\ & \frac{2^{nH(\alpha)}}{n+1} = 2^{nH(\alpha) - \log_2(n+1)} = 2^{n(H(\alpha) - o(1))} \end{aligned}$$

- And hence we got that $2^{n(H(\alpha) - o(1))} \leq \binom{n}{\alpha n} \leq 2^{nH(\alpha)}$.

■

10 Chernoff Bound

Chernoff's inequality, or *Chernoff bound*, is one of the most useful measure-concentration theorems. This inequality provides an exponentially small bound on the probability of deviating from the expectation. It applies to an important special case, namely, when the random variable at hand is the sum of independent indicator variables.¹ This inequality is formulated as follows.

Theorem 14 (Chernoff Bound) *Let X_1, \dots, X_n be independent indicator random variables, $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then*

1. For all $\delta > 0$,

$$\Pr[X \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu. \quad (2)$$

2. For all $0 < \delta < 1$,

$$\Pr[X \leq (1 - \delta)\mu] \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\mu. \quad (3)$$

Proof: Only Part 1 is proved. The proof of Part 2 is similar. Consider two numbers a and t and suppose that $t > 0$. For a given point ω of the sample space, $X(\omega) \geq a$ holds if and only if $e^{tX(\omega)} \geq e^{ta}$ holds. Therefore, $\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}]$. So the goal is to appropriately bound (from above) $\Pr[e^{tX} \geq e^{ta}]$ for $a = (1 + \delta)\mu$. Since e^{tX} is positive, we can use the Markov bound, i.e.,

$$\Pr[e^{tX} \geq e^{ta}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}. \quad (4)$$

The idea is to find the value t that minimizes the right side of this inequality. We first analyze the value $\mathbb{E}[e^{tX}]$.

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

From the independence of X_1, \dots, X_n it follows that $e^{tX_1}, \dots, e^{tX_n}$ are independent as well. Therefore, the expectation of the above product is the product of the individual expectations:

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] \quad (5)$$

For $1 \leq i \leq n$, let $p_i = \Pr[X_i = 1]$ (and, so, $\mu = \sum_{i=1}^n p_i$). Note that e^{tX_i} can take only two values: it is e^t with probability p_i and $e^0 (= 1)$ with probability $(1 - p_i)$. So we can easily compute the expectation of e^{tX_i} :

$$\mathbb{E}[e^{tX_i}] = p_i e^t + 1 - p_i = 1 + p_i(e^t - 1)$$

¹There are actually several different inequalities that go under the name ‘‘Chernoff (or Hoeffding) inequality.’’ They all have a similar flavor and what we do here is a good illustration for the whole area.

Since $1 + x \leq e^x$ for $x \geq 0$, it follows that $\mathbb{E}[e^{tX_i}] \leq e^{p_i(e^t - 1)}$. Thus, from (5) we obtain the following.

$$\mathbb{E}[e^{tX}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] \leq \prod_{i=1}^n e^{p_i(e^t - 1)} = e^{\sum_{i=1}^n p_i(e^t - 1)} = e^{\mu(e^t - 1)}$$

We now return to (4) and replace a with $(1 + \delta)\mu$.

$$\Pr[e^{tX} \geq e^{t(1+\delta)\mu}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\delta)\mu}} \leq \frac{e^{\mu(e^t - 1)}}{e^{t(1+\delta)\mu}} = e^{\mu(e^t - 1 - t(1+\delta))}$$

Recall that t is arbitrary. To obtain the best upper bound, we need to fix t so as to minimize $e^t - 1 - t(1 + \delta)$. By a straightforward calculus, we conclude that the minimum is obtained for $t = \ln(1 + \delta)$. After applying this assignment, we get

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

as required. ■

Notes about the proof. The crux of the above proof is the transition from $\Pr[X \geq a]$ to $\Pr[e^{tX} \geq e^{ta}]$. The rest of the proof is rather expected and straightforward. The idea behind that transition is as follows. As we have seen in the past, one can deduce some interesting properties of a random variable X (e.g., regarding its concentration) by looking at the *moments* of X , where the i th moment of X is defined as $\mathbb{E}[X^i]$. For example, we used the first and second moments in proving Markov and Chebyshev inequalities, respectively. By considering $\mathbb{E}[e^{tX}]$, we are actually using *all* the moments, packed together in the form of a *generating function*,² since

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(tX)^j}{j!}\right] = \sum_{j=0}^{\infty} \frac{t^j}{j!} \mathbb{E}[X^j].$$

As a result of Theorem 14, we obtain the following simplified (yet useful) bounds by a rather simple calculus (that is omitted).

Corollary 15 *Let X_1, \dots, X_n be independent indicator random variables, $X = \sum_{i=1}^n X_i$, and $\mu = \mathbb{E}[X]$. Then,*

- For all $0 < \delta < 1$, $\Pr[X \geq (1 + \delta)\mu] \leq e^{-\frac{\mu\delta^2}{3}}$,
- For all $0 < \delta < 1$, $\Pr[X \leq (1 - \delta)\mu] \leq e^{-\frac{\mu\delta^2}{2}}$, and
- For all $R \geq 6\mu$, $\Pr[X \geq R] \leq 2^{-R}$.

²The generating function of a series $\{a_j\}$ is the function $F(t) = \sum_j a_j t^j$. It is a useful tool for analyzing finite and infinite sequences. Occasionally, one uses the *exponential generating function*, namely, $\sigma(t) = \sum_j \frac{a_j}{j!} t^j$.

11 Uncountable Probability Spaces and a Taste of Measure Theory

Up until now, we have considered only *discrete* probability spaces. In many common scenarios, such spaces are not satisfactory. For example, already fairly simple problems (e.g. the analysis of geometric distributions) make it necessary to work with probability spaces over the set of all (infinite) strings of 0s and 1s. Clearly discrete spaces are inadequate here since this set has the cardinality \aleph of the real numbers. Thus, we are also interested in studying *uncountable* probability spaces. But for such spaces it is impossible to define the probability of events in terms of their individual members, as we did in the discrete case. As an example, it makes perfect sense to say that if we sample a point from the interval $[0, 1]$, then with probability $\frac{1}{10}$ it falls between 0.3 and 0.4. At the same time, the probability that it falls on 0.3721 is clearly zero, so the previous claim is not gotten by adding the probabilities of individual points.

We therefore need a more general notion of probability. In this more general framework we will have to consider as events only some (rather than all) of the subsets of the sample space. Put differently, given an uncountable sample space Ω , the set of events is a subset of 2^Ω that has a certain structure that guarantees several necessary properties. A structure with such properties is a σ -algebra. The events are also called the *measurable subsets*.

Formally, given a set Ω , a collection $\mathcal{J} \subseteq 2^\Omega$ is called a σ -algebra over Ω if \mathcal{J} contains the empty set and is closed under complementation and countable unions of its members; that is,

1. $\emptyset \in \mathcal{J}$,
2. $\Omega \setminus A \in \mathcal{J}$ for all $A \in \mathcal{J}$, and
3. If A_1, A_2, \dots is a (possibly infinite) sequence of subsets from \mathcal{J} , then $\bigcup_i A_i \in \mathcal{J}$.

A *measure* over a σ -algebra \mathcal{J} is a mapping $\mu : \mathcal{J} \rightarrow \mathbb{R}^+$ that satisfies the following two properties.

1. $\mu(\emptyset) = 0$
2. If A_1, A_2, \dots is a (possibly infinite) sequence of pairwise-disjoint subsets from \mathcal{J} , then

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i).$$

If, in addition, $\mu(\Omega) = 1$, then μ is said to be a *probability measure*. Now, we formally define a probability space as a triplet $(\Omega, \mathcal{J}, \mu)$, where Ω is a set (sample space), \mathcal{J} is a σ -algebra (events) and μ is a probability measure.

Essentially everything we have done so far with discrete probability spaces carries through to the more general settings, but more additional work is required to this end.

11.1 Defining measure (and probability) spaces in terms of a basis

It is very often convenient to define a σ -algebra by specifying only a subset \mathcal{E} of the σ -algebra that is called a *basis*. Then, the σ -algebra is implicitly defined by *extending* \mathcal{E} . The most natural way

of doing this is by introducing the σ -algebra that is *generated by* \mathcal{E} , that is, the intersection of all the σ -algebras that contain \mathcal{E} . (The reader can easily verify the intersection of every collection of σ -algebras over the same space Ω is itself a σ -algebra. Furthermore, there is at least one σ -algebra that contains \mathcal{E} , namely, 2^Ω .) Some of the standard examples of this procedure are:

- $\Omega = \mathbb{R}$ and \mathcal{E} consists of all open intervals.
- $\Omega = \prod_{i=1}^n \Omega_i$ where $\Omega_1, \dots, \Omega_n$ are σ -algebras. \mathcal{E} consists of all the *cylinders*, i.e., sets of the form $\Omega_1 \times \dots \times \Omega_{i-1} \times E_i \times \Omega_{i+1} \times \dots \times \Omega_n$ where E_i is a measurable set of Ω_i .

One advantage of this procedure is that it makes it possible to define the probability measure implicitly by specifying only the probability of the events of the basis \mathcal{E} , and calculate the probability for the rest of the events in the σ -algebra \mathcal{J} . As an example, suppose that Ω is the interval $[0, 1]$, and that \mathcal{E} consists of all the open intervals in $[0, 1]$ and $\mu((a, b)) = b - a$ for all $0 \leq a < b \leq 1$. We would like to extend μ , in some way, to the rest of \mathcal{J} , e.g., the sets $(0.1, 0.2) \cup (0.8, 0.9)$ or $[0.1, 0.2]$. There are some standard ways of extending μ . Let $A \in \mathcal{J}$ be given. The *outer measure* of A is

$$\mu^*(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu(E_i) : E_1, E_2, \dots \in \mathcal{E} \text{ and } A \subseteq \bigcup_{i=1}^{\infty} E_i \right\}.$$

Also, the *inner measure* of A is defined by

$$\mu_*(A) = \sup \left\{ \sum_{i=1}^{\infty} \mu(E_i) : E_1, E_2, \dots \in \mathcal{E}, \text{ where } E_i \text{ are disjoint, and } \bigcup_{i=1}^{\infty} E_i \subseteq A \right\}.$$

Whenever the inner measure and the outer measure agree we say A is measurable and its measure is $\mu(A)$

12 The Central Limit Theorem

The central limit theorem describe the limiting distribution of average of independent and identically distributed random variables. The theorem states that regardless of the distribution of the random variables their average is distributed according to the normal distribution. Recall that the standard *normal* (or *Gaussian*) distribution is defined by the density function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

(this function is illustrated in Figure 2) and the cumulative distribution function

$$\Phi(x) = \int_x^{\infty} \varphi(t) dx.$$

Now, consider a sequence X_1, \dots, X_n of random variables that are independent and identically distributed (i.i.d. for short). The central limit theorem says that the sum of these variables, appropriately normalized, has approximately a normal distribution.

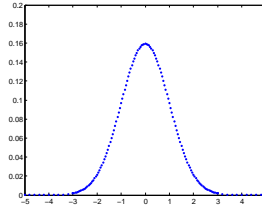


Figure 2: The density function of the normal distribution in the range $[-5,5]$

Theorem 16 (Central Limit Theorem) *Let X_1, X_2, \dots be a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define $Y_n = \sum_{i=1}^n X_i$ and $Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$. For all $a \in \mathbb{R}$,*

$$\lim_{n \rightarrow \infty} \Pr[Z_n > a] = \Phi(a).$$

Lecture 4

Lecturer: Nati Linial

Scribe: Menachem Fromer

Last Update: 19 Feb 2009 9:39 a.m.

These notes present basic properties of linear algebra metric spaces that we will encounter during the course.

13 Linear Algebra

The fundamental object of linear algebra is the vector space. We define matrices and their relation to linear transformation between vector spaces, and then mention eigenvectors as vectors on which a linear transformation acts in a particularly simple way. We discuss the cases where a linear transformation can be understood by observing the eigenvectors and eigenvalues of the corresponding matrix. From here our discussion advances to the topic of matrix similarity.

13.1 Vector Space

Let \mathcal{F} be a field, \mathcal{V} a vector space over \mathcal{F} . We will often take $\mathcal{F} = \mathbb{R}$, $\mathcal{V} = \mathbb{R}^n$; or alternatively $\mathcal{F} = \mathbb{Z}_p \equiv \{0, \dots, p - 1\}$ for a prime number p , $\mathcal{V} = \mathbb{Z}_p^n$.

The axioms of vector spaces include:

1. closure under vector addition
2. closure under scalar multiplication (multiplication by elements in \mathcal{F})

We define the *dimension* of $\mathcal{V} = \dim(\mathcal{V})$ as the minimal number of vectors which (linearly) span \mathcal{V} . Such a set of vectors is known as a *basis* of \mathcal{V} and can be shown to be linearly independent. A set of vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is said to be *linearly dependent* iff there exist scalars a_1, \dots, a_k (where at least one $a_i \neq 0$) s.t.

$$a_1 \mathbf{v}_1 + \dots + a_k \mathbf{v}_k = \mathbf{0}$$

Otherwise, the set is said to be *linearly independent*.

13.2 Matrices

A matrix is a two-dimensional array of elements from the field \mathcal{F} . Specifically, a linear transformation T s.t. $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be fully described by a unique matrix of dimension $m \times n$, $A_{m \times n}$:

$$\forall \mathbf{x} \in \mathbb{R}^n \text{ there holds } T(\mathbf{x}) = A\mathbf{x}$$

Thus, linear transformations and matrices can be thought of interchangeably. For a given vector space of a basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$, T and A are uniquely defined by the values $T(\mathbf{v}_1), \dots, T(\mathbf{v}_n)$ and $A\mathbf{v}_1, \dots, A\mathbf{v}_n$, respectively. This is so since any vector $\mathbf{v} \in \mathcal{V}$ can be uniquely represented as

$$c_1\mathbf{v}_1 + \dots + c_n\mathbf{v}_n$$

So by linearity:

$$A\mathbf{v} = \sum_{i=1}^n c_i A\mathbf{v}_i$$

Let \mathcal{V} and \mathcal{W} be vector spaces over \mathcal{F} , and let $T : \mathcal{V} \rightarrow \mathcal{W}$ be a linear transformation. Then the following equality holds:

$$\dim(\mathcal{V}) = \dim(\text{Ker}(T)) + \dim(\text{Im}(T)) \quad (6)$$

where

$$\text{Ker}(T) \equiv \{\mathbf{v} \in \mathcal{V} : T(\mathbf{v}) = 0\}$$

and

$$\text{Im}(T) \equiv \{\mathbf{w} \in \mathcal{W} : \exists \mathbf{v} \in \mathcal{V}, T(\mathbf{v}) = \mathbf{w}\}$$

The value $\dim(\text{Im}(T))$ is known as the *rank* of T , and is equal to the rank of the matrix A , which represents T . The *rank* of a matrix A can be defined in three equivalent ways:

1. the *row rank* of A . The row rank is the largest number of linearly independent row vectors of A .
2. the *column rank* of A . The column rank is the largest number of linearly independent column vectors of A .
3. the minimal number of matrices of rank 1, whose sum equals A . Matrix B is said to have rank 1 if $B = \mathbf{x} \otimes \mathbf{y} = \mathbf{xy}^T$ for some vectors \mathbf{x} and \mathbf{y} . In other words for all i, j there holds $B_{ij} = x_i y_j$.

13.3 Eigenvalues and Eigenvectors

Definition 17 Let $A \in M_n(\mathbb{C})$ be an $n \times n$ matrix with complex entries. We say that \mathbf{x} is an *eigenvector* of A , with *eigenvalue* λ iff:

$$A\mathbf{x} = \lambda\mathbf{x}$$

13.3.1 Characteristic Polynomial

$A\mathbf{v} = \lambda\mathbf{v}$ iff $(A - \lambda I)\mathbf{v} = 0$. Such a vector exists iff $A - \lambda I$ is singular, i.e., iff $\det(A - \lambda I) = 0$. This determinant can be considered as a polynomial in the variable λ . This polynomial is called the *characteristic polynomial* of A is defined as a polynomial in λ :

$$p(\lambda) = \det(\lambda I - A) \quad (7)$$

The roots of this polynomial are the eigenvalues of A . The claims below follow easily:

1. A has at most n eigenvalues of A (since they are the roots of an n^{th} degree polynomial).
2. A and A^T have the same eigenvalues (since the determinant is invariant under the transpose operation).

13.4 Eigenvector basis / Diagonalization

Lemma 18 Let $\mathbf{v}_1, \dots, \mathbf{v}_r$ be eigenvectors of A with corresponding distinct eigenvalues $\lambda_1, \dots, \lambda_r$. Then $\mathbf{v}_1, \dots, \mathbf{v}_r$ are linearly independent.

Proof: Suppose that there are k vectors of the \mathbf{v}_i that are linearly dependent, but no $k - 1$ or fewer of the \mathbf{v}_i 's are linearly dependent. Consider a linear dependency among the \mathbf{v}_i 's:

$$\sum_{j=1}^k c_j \mathbf{v}_{i_j} = \mathbf{0}. \quad (8)$$

Set $\mathbf{0}$ to be the all zeros vector, and apply A to both sides:

$$\mathbf{0} = A(\mathbf{0}) = A \left(\sum_{j=1}^k c_j \mathbf{v}_{i_j} \right) = \sum_{j=1}^k c_j \lambda_{i_j} \mathbf{v}_{i_j} \quad (9)$$

Multiply equation (8) by λ_{i_1} and subtract from equation (9) to derive a shorter non-trivial linear dependency among the \mathbf{v}_i , a contradiction. (Where have we used the assumption that the λ_i are distinct?) ■

Thus, if A has n distinct eigenvalues, then its eigenvectors form a basis for the space \mathfrak{R}^n .

13.4.1 Matrix similarity

Definition 19 A is similar to B if there exists a non-singular matrix P s.t. $A = PBP^{-1}$.

Definition 20 A is diagonalizable if it is similar to a diagonal matrix.

If the eigenvectors of A form a basis for the space, then $AV = V\Lambda$ where the i^{th} column of V is the i^{th} eigenvector \mathbf{v}_i , and Λ is a diagonal matrix where $\Lambda_{ii} = \lambda_i$ is the i^{th} eigenvalue of A . Equivalently $A = V\Lambda V^{-1}$ and we say that A is **similar** to a diagonal matrix Λ . Geometrically, if we can select for our space \mathcal{V} a basis that consists of eigenvectors of A , then the linear transformation corresponding to A simply performs contractions and expansions of the vectors in this basis, whence we can conveniently describe the behavior of the linear transformation arbitrary vector in the vector space.

Example: A Matrix that cannot be diagonalized

Not always is there a vector space basis of eigenvectors. For example, for the matrix

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

the only eigenvalue is 0. However, the eigenspace V_0 is spanned solely by $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Thus, here, there is no basis of \mathfrak{R}^2 composed of eigenvectors of C . Nevertheless for every matrix in $C \in M_n(\mathbb{C})$ we can find a matrix C_ϵ that has n distinct eigenvalues and is ϵ close to C (This is not difficult to prove, but does require some basic familiarity with measure theory, so we do not provide a proof). Therefore there are numerous problems on matrices which can be solved with the following general strategy: Deal first with the case of a matrix with n distinct eigenvalues and solve the general case by a limit argument.

13.4.2 Unitary equivalence

We state some definitions and facts which are useful

Definition 21 A matrix U is called unitary if $UU^* = I$, where $U^* = \overline{U^T}$

Theorem 22 U is unitary \iff The transformation $[x \rightarrow Ux]$ is an isometry (in l_2)

Proof:

\implies :

If U is an unitary matrix then

$$\|Ux\|^2 = \langle Ux, Ux \rangle = \langle U^*(Ux), x \rangle = \langle U^*Ux, x \rangle = \langle x, x \rangle = \|x\|^2$$

I.e. $[\forall x \|Ux\| = \|x\|]$. I.e. U is an isometry.

\Leftarrow :

Let U be an isometry. Then,

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle y, y \rangle + \langle x, y \rangle + \langle y, x \rangle \\ &= \|x\|^2 + \|y\|^2 + 2\operatorname{Re}(\langle x, y \rangle) \\ \|x + y\|^2 &= \|U(x + y)\|^2 \\ &= \|Ux + Uy\|^2 \\ &= \|Ux\|^2 + \|Uy\|^2 + 2\operatorname{Re}(\langle Ux, Uy \rangle) \\ &= \|x\|^2 + \|y\|^2 + 2\operatorname{Re}(\langle U^*Ux, Uy \rangle) \end{aligned}$$

Hence for any x, y

$$\operatorname{Re}(\langle x - U^*Ux, y \rangle) = \operatorname{Re}(\langle x, y \rangle) - \operatorname{Re}(\langle U^*Ux, y \rangle) = 0$$

Especially for any x and $y = x - U^*Ux$,

$$\|y\|^2 = \langle x - U^*Ux, y \rangle = 0 \quad (\|y\| \in \mathbb{R})$$

Hence $x = U^*Ux$ for any x and $U^*U = I$.

■

Definition 23 A is unitary equivalent to B if there exists a unitary matrix U s.t. $A = UBU^*$.

Theorem 24 For complex-valued matrices, $A^*A = AA^* \iff A = U\Lambda U^*$, for a unitary matrix U and diagonal matrix Λ . Such a matrix A is called normal.

Especially, Real symmetric matrices are normal and Hermitian matrices are normal.

Corollary 25 (The spectral theorem for real symmetric matrices) Every real symmetric matrix $A \in \mathbb{M}_{n \times n}$ can be expressed as $A = UDU^T$, where $U \in \mathbb{M}_{n \times n}$ is orthogonal and $D \in \mathbb{M}_{n \times n}(\mathbb{R})$ is "diagonal", i.e., the diagonal entries of D are $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ and all other entries of D are zero.

Theorem 26 If a complex-valued matrix A is Hermitian, i.e., $A = A^*$, then all of its eigenvalues are real.

Theorem 27 If the matrix A is real and symmetric and symmetric, then all its eigenvalues are real.

Definition 28 A matrix A is positive semi-definite if it is Hermitian and all its eigenvalues are non-negative.

Definition 29 A matrix A is positive definite if it is Hermitian and all its eigenvalues are positive.

14 Metric Structures defined on the vector space

Besides the concepts of linear dependencies, vector bases, and ranks, there are additional useful structures that can be associated with a vector space:

14.1 Metric Space

A *metric space* is a set X with a "distance function" d . For every two elements $\mathbf{v}, \mathbf{x} \in X$, $d(\mathbf{v}, \mathbf{x})$ is a non-negative real number to be thought as "distance" between \mathbf{v} and \mathbf{x} . A metric d must satisfy the following conditions:

1. Positivity: $d(\mathbf{v}, \mathbf{x}) \geq 0$ for every $\mathbf{v}, \mathbf{x} \in X$, and $d(\mathbf{v}, \mathbf{x}) = 0$ iff $\mathbf{v} = \mathbf{x}$
2. Symmetry: $d(\mathbf{v}, \mathbf{x}) = d(\mathbf{x}, \mathbf{v})$
3. Triangle Inequality: $\forall \mathbf{y}, \mathbf{x}, \mathbf{v}$ there holds $d(\mathbf{v}, \mathbf{x}) \leq d(\mathbf{v}, \mathbf{y}) + d(\mathbf{y}, \mathbf{x})$

14.2 Normed Spaces (Banach Spaces)

In many cases it is useful to associate a "length" to each vector in the space. This length is called a norm. The norm of \mathbf{v} is denoted as $\|\mathbf{v}\|$. A norm must satisfy the following conditions:

1. Linearity: $\forall \mathbf{v} \in \mathcal{V}$, and every $\lambda \in \mathfrak{R}$ there holds $\|\lambda \mathbf{v}\| = |\lambda| \|\mathbf{v}\|$
2. Triangle Inequality: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}$ there holds $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
3. Positivity: $\forall \mathbf{v} \in \mathcal{V} \|\mathbf{v}\| \geq 0$, and $\|\mathbf{v}\| = 0$ iff $\mathbf{v} = \mathbf{0}$

Examples of norms: For every $1 \leq p < \infty$ the norm $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ is called the l_p - norm. The l_∞ - norm is defined to be $\|x\|_\infty = \max_i |x_i|$. Two particularly important instances of norms are the Euclidean norm l_2 with which you are already familiar is defined via $\|x\|_2 = \sqrt{\sum_i x_i^2}$, and the l_1

norm $\|x\|_1 = \sum_i |x_i|$.



Figure 3: unit sphere on \mathfrak{R}^2 of l_p for several values of p

A normed space is also a metric since $d(\mathbf{v}, \mathbf{x}) = \|\mathbf{v} - \mathbf{x}\|$ yields a metric on \mathcal{V} .

14.3 Inner Product Spaces (Hilbert Spaces)

In a vector space over \mathbb{C} we define the *inner product* of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_i u_i \bar{v}_i$$

For real vectors spaces this take the form $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_i u_i v_i$. There is a close connection between inner products and l_2 norms as follows:

$$\|\mathbf{v}\|_2 = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \tag{10}$$

This relation implies that the l_2 norm satisfies the *parallelogram law*:

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$$

the converse also holds: every normed space which satisfies the *parallelogram law* is an inner product space .

14.3.1 Matrix Norm

Similarly to vectors, we would like a definition of the "size" of a matrix. One application using such quantities is *Singular Value Decomposition* (SVD), which has as input a matrix $A_{m \times n}$ and a natural number r . As output, we would like a matrix $B_{m \times n}$ s.t. $\text{rank}(B) \leq r$ and B is "close" to A . For these tasks, we need to define a norm on matrices. There are quite a few possibilities as to how to define such a norm:

1. Consider a matrix to simply be a vector (though the entries are arranged in a 2-dimensional array) and use any known vector norms. For example: Frobenius Norm = Hilbert-Schmidt Norm = $\| \cdot \|_F$. This is the special case where we apply the l_2 norm to the entries of the matrix

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

2. When matrices are viewed as linear transformations, then a natural question to ask is "to what extent does the transformation expand vector lengths?" To make sense out of this we need the above definitions of vector lengths. As an illustration, consider the $n \times m$ matrix A as a linear map $A : \mathfrak{R}^n \rightarrow \mathfrak{R}^m$ given by $\mathbf{x} \rightarrow A\mathbf{x}$. Now let us equip \mathfrak{R}^n with the l_p norm and \mathfrak{R}^m with the l_q nor. This induces the following norm on A :

$$\|A\|_{p \rightarrow q} \equiv \max_{\mathbf{x} \in \mathbb{C}^n} \frac{\|A\mathbf{x}\|_q}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|A\mathbf{x}\|_q$$

Norms of this type on matrices are called operator norms.

Lecture 5

Lecturer: Nati Linial

Scribe: Shai Shalev-Shwartz

Last Update: 19 Feb 2009 9:39 a.m.

The structure of this lecture is as follows. We start with a short review of the equivalence between norms and convex bodies which are symmetric relative to the origin. We then define the notion of dual norms from several perspectives.

15 Norms and Dual Norms

15.1 Norms and unit balls

In the previous lecture we defined a vector norm, $\|\mathbf{x}\|$, as a function on vectors in \mathbb{R}^n that satisfies:

1. Linearity: $\forall \lambda \in \mathbb{R}, \|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$
2. Triangle inequality: $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n, \|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
3. Positivity: $\forall \mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\| \geq 0$ and equality holds iff $\mathbf{v} = \mathbf{0}$.

A norm $\|\cdot\|$ defines a *unit ball*, $B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}$. B is a convex set, (means that if \mathbf{x}_1 and \mathbf{x}_2 are in B then the interval defined by $\{\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2 : \lambda \in [0, 1]\}$ is included in B as well). This property follows from the triangle inequality and from the linearity of the norm,

$$\|\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2\| \leq \|\lambda \mathbf{x}_1\| + \|(1 - \lambda) \mathbf{x}_2\| = \lambda \|\mathbf{x}_1\| + (1 - \lambda) \|\mathbf{x}_2\| \leq \lambda + (1 - \lambda) = 1.$$

In addition, B is symmetric with respect to the origin. That is, if $\mathbf{x} \in B$ then $-\mathbf{x} \in B$ (this follows directly from the linearity of the norm and the definition of B). Finally, from the positivity condition we get that B is full dimensional — This can be stated in two equivalent ways:

- B contains some small Euclidean ball, i.e. $B \supset \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq \epsilon\}$ for some $\epsilon > 0$.
- For every $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$ there is some $\alpha > 0$ such that $\alpha \mathbf{x} \in B$.

The other implication is also true. That is, if the set B is convex and symmetric around the origin, and full dimensional, then it defines a norm as follows. We set $\|\mathbf{0}\|_B = 0$ and for each $\mathbf{x} \neq \mathbf{0}$, we define $\|\mathbf{x}\|_B = \frac{1}{\max\{\lambda > 0 : \lambda \mathbf{x} \in B\}}$.

15.2 Dual norms and dual sets

Given a norm $\|\cdot\|$ we define its dual norm as,

$$\|\mathbf{x}\|^\star = \max_{\mathbf{v} \neq 0} \frac{|\langle \mathbf{x}, \mathbf{v} \rangle|}{\|\mathbf{v}\|} . \quad (11)$$

Explanation for this definition: As we saw, it is a good idea to measure how "large" a transformation is by observing how much it stretches the elements in its domain. We view every $\mathbf{x} \in \mathbb{R}^n$ as a linear functional, namely we associate to each $\mathbf{x} \in \mathbb{R}^n$ a map $\mathbb{R}^n \rightarrow \mathbb{R}$ via $\mathbf{y} \rightarrow \langle \mathbf{x}, \mathbf{y} \rangle$. In this view the definition in Eqn 11 is very natural.

From the linearity of the norm and of the inner product, we can rewrite the above as,

$$\|\mathbf{x}\|^\star = \max_{\mathbf{v} : \|\mathbf{v}\|=1} |\langle \mathbf{x}, \mathbf{v} \rangle| . \quad (12)$$

First, let us prove that this indeed defines a norm. The linearity property follows from the linearity of the inner product and the positivity property follows directly from the definition. To prove the triangle inequality, we use the triangle inequality on scalars as follows,

$$\begin{aligned} \|\mathbf{x}_1 + \mathbf{x}_2\|^\star &= \max_{\mathbf{v} : \|\mathbf{v}\|=1} |\langle \mathbf{x}_1 + \mathbf{x}_2, \mathbf{v} \rangle| \\ &= \max_{\mathbf{v} : \|\mathbf{v}\|=1} |\langle \mathbf{x}_1, \mathbf{v} \rangle + \langle \mathbf{x}_2, \mathbf{v} \rangle| \\ &\leq \max_{\mathbf{v} : \|\mathbf{v}\|=1} (|\langle \mathbf{x}_1, \mathbf{v} \rangle| + |\langle \mathbf{x}_2, \mathbf{v} \rangle|) \\ &\leq \max_{\mathbf{v} : \|\mathbf{v}\|=1} |\langle \mathbf{x}_1, \mathbf{v} \rangle| + \max_{\mathbf{v} : \|\mathbf{v}\|=1} |\langle \mathbf{x}_2, \mathbf{v} \rangle| \\ &= \|\mathbf{x}_1\|^\star + \|\mathbf{x}_2\|^\star . \end{aligned}$$

Recall that a norm $\|\cdot\|$ is completely defined by its unit ball B and that B is a convex set. In general, for a convex set B we define its *polar*,

$$B^\star = \{\mathbf{x} : \forall \mathbf{v} \in B, \langle \mathbf{x}, \mathbf{v} \rangle \leq 1\} .$$

For instance, recall that the unit ball of the norm l_∞ is the cube $[-1, 1]^n$ and the unit ball of the norm l_1 is $\text{conv}(\{\pm e_i\}_{i=1}^n)$. The two sets are polar to each other.

The following lemma connects the definitions of dual sets and norms.

Lemma 30 *Let $\|\cdot\|$ be a norm and let B be its unit ball. Let B^\star be the polar of B . Then, B^\star is the unit ball of the dual norm $\|\cdot\|^\star$.*

Proof: Let $\mathbf{x} \in B^\star$. Then, for all $\mathbf{v} \in B$ we have that $\langle \mathbf{x}, \mathbf{v} \rangle \leq 1$. In particular, for all \mathbf{v} such that $\|\mathbf{v}\| = 1$ we have that $\langle \mathbf{x}, \mathbf{v} \rangle \leq 1$. We therefore get from (12) that $\|\mathbf{x}\|^\star \leq 1$. Now, assume that $\mathbf{x} \notin B^\star$. Then, there exists a vector $\mathbf{v} \in B$ and a scalar $a > 1$ such that $\langle \mathbf{x}, \mathbf{v} \rangle = a$. From the linearity of the norm we can assume that $\|\mathbf{v}\| = 1$. Therefore, from (12) we get that $\|\mathbf{x}\|^\star \geq a > 1$. In summary, a vector \mathbf{x} is in B^\star iff $\|\mathbf{x}\|^\star \leq 1$, which concludes our proof. ■

The above lemma leads to a generalization of Cauchy-Schwarz inequality: Hölder inequality.

Lemma 31 Let $\|\cdot\|$ be a norm and let $\|\cdot\|^\star$ be its dual norm. Then, for any two vectors \mathbf{x} and \mathbf{v} we have that,

$$|\langle \mathbf{x}, \mathbf{v} \rangle| \leq \|\mathbf{x}\| \|\mathbf{v}\|^\star .$$

Proof: Let B and B^\star be the unit balls of $\|\cdot\|$ and $\|\cdot\|^\star$. First note that from the definition of B we have that $\mathbf{x}/\|\mathbf{x}\| \in B$ and $\mathbf{v}/\|\mathbf{v}\|^\star \in B^\star$. The definition of B^\star now implies that

$$\left| \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{v}}{\|\mathbf{v}\|^\star} \right\rangle \right| \leq 1 .$$

The claim in the lemma now follows from the linearity of the inner product. ■

15.3 The p -norm

For each $p \geq 1$ we have defined the p -norm of a vector $\mathbf{v} \in \mathbb{R}^d$ as

$$\|\mathbf{v}\|_p = \left(\sum_{k=1}^d |v_k|^p \right)^{\frac{1}{p}} .$$

We now show that $\|\cdot\|_p^\star = \|\cdot\|_q$, where

$$\frac{1}{p} + \frac{1}{q} = 1 .$$

To show this³, we denote by B_p the unit ball of $\|\cdot\|_p$, and similarly, B_q denotes the unit ball of $\|\cdot\|_q$. It is sufficient to show that $B_q = B_p^\star$.

1. (\subseteq) : Let $\mathbf{x} \in B_q$. Then, from Hölder inequality (see next subsection) we get that for every vector \mathbf{v} there holds $\langle \mathbf{x}, \mathbf{v} \rangle \leq \|\mathbf{x}\|_q \|\mathbf{v}\|_p \leq \|\mathbf{v}\|_p$. Thus, for each $\mathbf{v} \in B_p$ we get that, $\langle \mathbf{x}, \mathbf{v} \rangle \leq 1$ which implies that $\mathbf{x} \in B_p^\star$.
2. (\supseteq) : Let $\mathbf{x} \in B_p^\star$. Then, for each \mathbf{v} such that $\|\mathbf{v}\|_p \leq 1$ we have that $\langle \mathbf{x}, \mathbf{v} \rangle \leq 1$. Define $\mathbf{v} \in \mathbb{R}^d$ to be the vector for which

$$v_k = \frac{\text{sign}(x_k) (x_k)^{q-1}}{\|\mathbf{x}\|_q^{q/p}} .$$

Note that $p q = p + q$ and thus,

$$\|\mathbf{v}\|_p^p = \frac{1}{\|\mathbf{x}\|_q^q} \sum_{k=1}^d |x_k|^{(q-1)p} = \frac{1}{\|\mathbf{x}\|_q^q} \sum_{k=1}^d |x_k|^q = 1 .$$

Therefore,

$$1 \geq \langle \mathbf{x}, \mathbf{v} \rangle = \sum_{k=1}^d \text{sign}(x_k) x_k^{1+q-1} = \sum_{k=1}^d |x_k|^q = \|\mathbf{x}\|_q^q ,$$

which implies that $\|\mathbf{x}\|_q \leq 1$ and thus $\mathbf{x} \in B_q$.

³If $p = 1$ then $q = \infty$ where $\|\mathbf{x}\|_\infty = \max_i |x_i|$. In this case, one can directly show that $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$.

15.4 Young, Hölder and Minkowski inequalities

Lemma 32 (Young inequality) Let $p, q \in \mathbb{R}$ be two positive scalars such that $\frac{1}{p} + \frac{1}{q} = 1$. Then,

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q .$$

Proof: The log function is a concave function. That is, for each $\lambda \in [0, 1]$ and scalars a, b we have that, $\log(\lambda a + (1 - \lambda)b) \geq \lambda \log(a) + (1 - \lambda) \log(b)$. Therefore,

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

Taking exponent of both sides gives the desired inequality. ■

Lemma 33 (Hölder inequality) Let $p, q \in \mathbb{R}$ be two positive scalars such that $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$|\langle \mathbf{x}, \mathbf{v} \rangle| \leq \|\mathbf{x}\|_p \|\mathbf{v}\|_q .$$

Proof: The proof uses the triangle inequality $|a + b| \leq |a| + |b|$ and Young inequality as follows,

$$\begin{aligned} \frac{|\langle \mathbf{x}, \mathbf{v} \rangle|}{\|\mathbf{x}\|_p \|\mathbf{v}\|_q} &= \frac{|\sum_{k=1}^d x_k v_k|}{\|\mathbf{x}\|_p \|\mathbf{v}\|_q} \leq \frac{\sum_{k=1}^d |x_k| |v_k|}{\|\mathbf{x}\|_p \|\mathbf{v}\|_q} = \sum_{k=1}^d \left(\frac{|x_k|}{\|\mathbf{x}\|_p} \frac{|v_k|}{\|\mathbf{v}\|_q} \right) \\ &\leq \sum_{k=1}^d \left(\frac{1}{p} \left(\frac{|x_k|}{\|\mathbf{x}\|_p} \right)^p + \frac{1}{q} \left(\frac{|v_k|}{\|\mathbf{v}\|_q} \right)^q \right) = \frac{1}{p} \frac{\sum_{k=1}^d |x_k|^p}{\|\mathbf{x}\|_p^p} + \frac{1}{q} \frac{\sum_{k=1}^d |v_k|^q}{\|\mathbf{v}\|_q^q} \\ &= \frac{1}{p} + \frac{1}{q} = 1 . \end{aligned}$$

■

Finally, the **Minkowski** inequality states that if $p \geq 1$, then, $\|\mathbf{x} + \mathbf{v}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{v}\|_p$. This is the triangle inequality property of the p-norm.

Lecture 6

Lecturer: Nati Linial

Scribe: Vladimir Grin

Last Update: 19 Feb 2009 9:39 a.m.

The topic of this lecture is Singular Value Decomposition (SVD). We start by presenting the singular value decomposition of a given matrix A . We then demonstrate how to use it to approximate a given matrix by lower rank matrix.

16 Singular Value Decomposition - SVD

A real matrix N is called normal if $NN^T = N^T N$, e.g. every real symmetric matrix is normal. A fundamental theorem in matrix theory states that a real normal matrix has the form $N = U^T D U$ where U is a real orthogonal matrix, i.e. $U U^T = U^T U = I$ and D is real and diagonal. Equivalently $U N = D U^T$ therefore this theorem represent the eigenvectors of the normal matrix N in the rows of U and its eigenvalues in the entries of D . This theorem is proved in basic algebra course for symmetric matrices and will not be proved here, although it is a special case of the SVD theorem described below. A possible approach to the proof is this: It is not hard to see that for every real symmetric matrix A and for every $\epsilon > 0$ there is a real symmetric matrix B such that $\forall i, j \quad |a_{ij} - b_{ij}| < \epsilon$ and such that every eigenvalue is simple (there are no repeated eigenvalues). Using the fact that eigenvectors with different eigenvalues must be orthogonal the statement hold for B . The desired conclusion for A follows by a simple continuity argument.

The singular value decomposition may be considered as a substitute for spectral (eigenvalue) decomposition for matrices which are not normal (or not even square). The singular value decomposition theorem states that every real matrix $M \in \mathbb{M}_{m \times n}$ with $m \geq n$ can be expressed as $U^T D V$ where $U \in M_m$ and $V \in M_n$ are orthogonal matrices and $D \in \mathbb{M}_{m \times n}(\mathbb{R})$ is non-negative and "diagonal", i.e., all entries of D are zeros except for the singular values $D_{11} = \sigma_1, \dots, D_{nn} = \sigma_n$. By convention we will always assume that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$. There is a close connection between the singular value decomposition of a matrix A and the diagonalization of AA^T and $A^T A$.

$$\text{If } A = U^T D V \text{ then } AA^T = U^T D^2 U \text{ and } A^T A = V^T D^2 V.$$

We conclude that the singular values of A are the square roots of the eigenvalues of AA^T and $A^T A$, and the rows of the orthogonal matrices U, V are the eigenvectors of AA^T and $A^T A$ respectively (Note that these eigenvalues are all nonnegative, since every matrix of the form ZZ^T is positive semi definite). From the singular value decomposition we can see that the rank of A is $\max\{r : \sigma_r > 0\}$. We can also see that the basis for the image and the kernel of the matrix A are first r rows of U and last $m - r$ rows of U respectively.

Theorem 34 Every matrix $A \in \mathbb{M}_{m \times n}$ with $m \geq n$ can be expressed as $A = U D V^T$, where $U \in \mathbb{M}_{m \times m}$ and $V \in \mathbb{M}_{n \times n}$ are orthogonal and $D \in \mathbb{M}_{m \times n}(\mathbb{R})$ is "diagonal", i.e., the diagonal entries of

D_{11}, \dots, D_{mm} are $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ (they called singular values of A), and all other entries of D are zero.

Proof: We prove equivalently that there are two orthogonal matrices $U \in \mathbb{M}_{m \times m}$ and $V \in \mathbb{M}_{n \times n}$ such that

$$U^T AV = D, \quad \text{where } D_{ii} = \sigma_i \text{ are the singular values}$$

We prove the theorem by induction on the size of matrices $A \in \mathbb{M}_{m \times n}$. Specifically, we assume that for every matrix $B \in \mathbb{M}_{(m-1) \times (n-1)}$ there are orthogonal matrices $U_{m-1} \in \mathbb{M}_{(m-1) \times (m-1)}$ and $V_{n-1} \in \mathbb{M}_{(n-1) \times (n-1)}$ such that $U_{m-1}^T B V_{n-1} = D_{(m-1) \times (n-1)}$ is diagonal and non-negative. We want to conclude that $U_m^T A V_n = D_{m,n}$ for some orthogonal matrices U_m, V_n , that are related to U_{m-1}, V_{n-1} . We need to introduce some notation: Given a vector $v_1 \in \mathbb{C}^n$ and a matrix $V_2 \in \mathbb{M}_{n \times (n-1)}$ define $(v_1|V_2)$ to be the matrix in $\mathbb{M}_{n \times n}$ constructed by concatenating the column v_1 to the columns V_2 . Similarly, given a vector $u_1 \in \mathbb{C}^m$ and a matrix $U_2 \in \mathbb{M}_{m \times (m-1)}$ define $\left(\frac{u_1^T}{(U_2)^T}\right)$ to be $(u_1|U_2)^T$.

Our proof plan is as follows: Given a matrix $A \in \mathbb{M}_{m \times n}$ we will find two orthogonal matrices $U = (u_1|U_2) \in \mathbb{M}_{m \times m}$ and $V = (v_1|V_2) \in \mathbb{M}_{n \times n}$ such that

$$(u_1|U_2)^T A (v_1|V_2) = \left(\begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array}\right),$$

for some non-negative constant σ_1 , and a matrix $B \in \mathbb{M}_{(m-1) \times (n-1)}$. We finish the proof by applying the induction hypothesis to B .

We now turn to the actual proof: Define $\sigma_1 := \|A\|_{2 \rightarrow 2}$, and let v_1 and u_1 be unit vectors (in ℓ_2 norm), s.t. $Av_1 = \sigma_1 u_1$. Extend v_1 and u_1 to two orthogonal matrices: $V = (v_1|V_2)$ and $U = (u_1|U_2)$. Now,

$$\begin{aligned} U^T AV &= \left(\frac{u_1^T}{U_2^T}\right)(A)(v_1|V_2) = \left(\frac{u_1^T}{U_2^T}\right)(Av_1|AV_2) = \left(\frac{u_1^T Av_1}{U_2^T Av_1} \mid \frac{u_1^T AV_2}{U_2^T AV_2}\right) = \\ &= \left(\begin{array}{c|c} \sigma_1 & w \\ \hline 0 & B \end{array}\right) \end{aligned}$$

The matrix $U^T AV$ has the same operator norm as A , i.e. $\|U^T AV\|_{2 \rightarrow 2} = \|A\|_{2 \rightarrow 2}$ since the matrices U and V are orthogonal and hence define an isometry. We want to show that $w = 0$. Note that

$$\left(\frac{\sigma_1}{0} \mid \frac{w}{B}\right) \cdot \left(\frac{\sigma_1}{w^T}\right) = \left(\frac{\sigma_1^2 + \|w\|^2}{z}\right).$$

Denote $\phi = (\sigma_1 \mid w)^T$ and $\psi = (\sigma_1^2 + \|w\|^2 \mid z)^T$. Then $\|\phi\|_2 = \sqrt{\sigma_1^2 + \|w\|^2}$ and $\|\psi\|_2 \geq \sigma_1^2 + \|w\|^2$ which together gives $\sigma_1 \geq \|\phi\|_2 / \|\psi\|_2 \geq \sqrt{\sigma_1^2 + \|w\|^2} \geq \sigma_1$ and the equality holds only when $w = 0$.

We have shown that

$$U^T AV = \left(\begin{array}{c|c} \sigma & 0 \\ \hline 0 & B \end{array}\right) \text{ or equivalently } A = U \left(\begin{array}{c|c} \sigma & 0 \\ \hline 0 & B \end{array}\right) V^T$$

By the induction hypothesis $B = U_{m-1} D_{(m-1) \times (n-1)} V_{n-1}^T$, thus

$$A = (u_1|U_2) \cdot \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & U_{m-1} \end{array}\right) \cdot \left(\begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & D_{(m-1) \times (n-1)} \end{array}\right) \cdot \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & V_{n-1}^T \end{array}\right) \cdot \left(\begin{array}{c} v_1^T \\ \hline V_2^T \end{array}\right)$$

■

Following the SVD theorem it is straightforward to verify the following statements:

- The Frobenius norm of $A = U^T D V$ is the square root of the sum of squares of the singular values. In words, $\|A\|_F = \text{trace}(A^T A) = \text{trace}(D^2)$, which follows from the following property of the trace operator: $\text{trace}(AB) = \text{trace}(BA)$ (both are equal to $\sum_{ij} a_{ij} b_{ij}$)
- The matrix norm $\|A\|_{2 \rightarrow 2}$ is the largest singular value of A .

16.1 Approximating a matrix A by a matrix of rank k

We can use SVD in order to solve the following problem: Given a matrix A and $k \in \mathbb{N}$, find a matrix B with $\text{rank} \leq k$, s.t. A and B are as close as possible, i.e. $\min\{\|A - B\|\}$ using some matrix norm. Two natural choices for this norm are $2 \rightarrow 2$ and Frobenius norm $\|M\|_F = \sqrt{\sum M_{i,j}^2}$.

Theorem 35 *Let $A \in \mathbb{M}_{m \times n}$ with $m \geq n$ be a matrix. Let $A = U D V^T$ be its singular value decomposition. Then the minimum of $\|A - B\|$ over all matrices B of rank at most k is achieved with $B = U D^{(k)} V^T$, where $D_{i,j}^{(k)} = D_{i,j}$ if $i, j \leq k$ and 0 otherwise.*

Proof: We will prove the theorem for the norm $2 \rightarrow 2$. Note that for this choice of B we have $\|A - B\|_{2 \rightarrow 2} = \sigma_{k+1}$, and we will show that for any matrix C of rank $\leq k$ there exist a unit vector z , s.t. $\|(A - C)z\| \geq \sigma_{k+1}$.

Since $\text{rank}\{C\} \leq k$ it follows that $\ker(C)$ has dimension $\geq m - k$. Consequently, $\ker(C)$ has a nonempty intersection with any $(k + 1)$ -dimensional subspace. In particular, there is a unit vector z in the subspace $\text{span}(v_1, \dots, v_{k+1}) \cap \ker(C)$, i.e., $z = \sum_1^{k+1} \alpha_i v_i$, and $Cz = 0$. We will show that $\|(A - C)z\| \geq \sigma_{k+1}$ and this will establish our claim. The matrix A can be written as $A = \sum_{i=1}^m \sigma_i \cdot u_i \otimes \bar{v}_i$ and:

$$(A - C) \cdot z = Az = \left(\sum_{i=1}^m \sigma_i \cdot u_i \otimes \bar{v}_i \right) \cdot z = \sum_{i=1}^m \sigma_i u_i \cdot \langle v_i, z \rangle = \sum_1^{k+1} \sigma_i \cdot u_i \cdot \langle v_i, z \rangle,$$

where the last equality follows from the fact the v_{k+2}, \dots, v_m are orthogonal to z . Using the orthogonality of the u_i 's we compute the norm

$$\|(A - C)z\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 \cdot |\langle v_i, z \rangle|^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} |\langle v_i, z \rangle|^2 = \sigma_{k+1}^2 \|z\|^2 = \sigma_{k+1}^2,$$

since the unit vector z is spanned by the orthonormal vectors v_1, \dots, v_{k+1} . This shows that the lower bound on the matrix norm, i.e., $\|A - C\|_{2 \rightarrow 2} \geq \sigma_{k+1}$. ■

The SVD theorem as well as theorem 35 can be stated and proved for matrices with complex entries in the same manner, while substituting any orthogonal matrix with its complex counterpart unitary matrix. The singular values in this case remain real numbers.

Lecture 7

Lecturer: Nati Linial

Scribe: Noa Eidelstein

Last Update: 19 Feb 2009 9:39 a.m.

In this lecture we first remind you how to express real symmetric or Hermitian matrices in terms of their eigenvalues and eigenvectors. We describe the *cone*⁴ of *positive semi definite* matrices. We present analytical characterizations of the eigenvalues and eigenvectors called the Rayleigh-Ritz theorem and Courant-Fischer theorem. Then we turn to non-negative matrices and discuss some aspects of the Perron-Frobenius theorem.

17 Real symmetric matrices, eigenvalues and eigenvectors

A matrix N is called normal if $NN^* = N^*N$. Normal matrices are exactly those matrices which are unitarily diagonalizable, i.e., N is a normal matrix iff there exists a unitary matrix U such that UNU^* is a diagonal matrix. In particular, all real symmetric matrices are normal. An important additional feature of real symmetric matrices is that their eigenvalues are real:

Theorem 36 *Let $A \in M_n(\mathbb{R})$ be a symmetric real matrix, then*

$$A = U\Lambda U^T, \quad U \text{ is orthogonal and } \Lambda \text{ is real and diagonal.}$$

For each i , the i -th column vector of U and the i -th diagonal entry of Λ are an eigenvalue-eigenvector pair of A .

We define a special kind of symmetric matrices called *positive semi definite* matrices, abbreviated *PSD*. A matrix A is *PSD* if all its eigenvalues are non-negative. Below are some equivalent definitions:

Claim 37 *The following conditions are equivalent for real symmetric matrices:*

1. A is *PSD*.
2. $A = MM^T$ for some real matrix M .
3. For every vector \mathbf{x} there holds $\mathbf{x}^T A \mathbf{x} \geq 0$.

Proof:

- $1 \rightarrow 2$: By Theorem 36, we can write $A = U\Lambda U^T$. Let $M = U\sqrt{\Lambda}$ where $\sqrt{\Lambda}$ is a diagonal matrix, whose (i, i) -entry is $\sqrt{\Lambda_{ii}}$ (and $\Lambda_{ii} \geq 0$, since it is an eigenvalue of a PSD matrix).
- $2 \rightarrow 3$: By the assumption $\mathbf{x}^T A \mathbf{x} = \mathbf{x}^T M M^T \mathbf{x} = \langle M^T \mathbf{x}, M^T \mathbf{x} \rangle = \|M^T \mathbf{x}\|_2^2 \geq 0$.

⁴A cone is a subset $C \subseteq \mathbb{R}^n$ that is closed under vector addition and under multiplication by non negative scalars. Namely, if $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \geq 0$, then $\mathbf{x} + \mathbf{y} \in C$ and $\lambda \mathbf{x} \in C$.

- $3 \rightarrow 1$: If \mathbf{v} is an eigenvector with eigenvalue λ , then $A\mathbf{v} = \lambda\mathbf{v}$. By assumption $\mathbf{v}^\top A\mathbf{v} \geq 0$. But $\mathbf{v}^\top A\mathbf{v} = \lambda\langle \mathbf{v}, \mathbf{v} \rangle = \lambda\|\mathbf{v}\|^2$ and it follows that $\lambda \geq 0$.

■

The set of PSD matrices is a cone, namely for every pair of real numbers $c_1, c_2 \geq 0$ and every pair of matrices $A_1, A_2 \in PSD$ there holds $c_1A_1 + c_2A_2 \in PSD$. This is easily derived by considering the quadratic form $\mathbf{x}^\top(c_1A_1 + c_2A_2)\mathbf{x} = c_1\mathbf{x}^\top A_1\mathbf{x} + c_2\mathbf{x}^\top A_2\mathbf{x} \geq 0$.

18 The Variational Approach to Eigenvalues

Definitions:

1. The spectrum of a matrix is the collection of its eigenvalues (with multiplicities).
2. A spectral decomposition of a vector is the expression of this vector as a linear combination of eigenvectors.

Theorem 38 (Rayleigh-Ritz) *Let $A \in M_n(\mathbb{R})$ be a symmetric real matrix, and let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ be its eigenvalues (in descending order). Then*

$$\lambda_1 = \max_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^\top A\mathbf{x}}{\|\mathbf{x}\|^2} = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top A\mathbf{x},$$

and the maximum is attained by an eigenvector \mathbf{v}_1 corresponding to λ_1 .

Furthermore, let $\mathbf{v}_1, \dots, \mathbf{v}_k$ denote k eigenvectors corresponding to the eigenvalues $\lambda_1, \dots, \lambda_k$ respectively, and let V_k^\perp denote the linear subspace orthogonal to their span (i.e., $V_k^\perp = \{\mathbf{v} \mid \mathbf{v} \perp \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_k)\}$). Then

$$\lambda_{k+1} = \max_{\mathbf{x} \in V_k^\perp} \frac{\mathbf{x}^\top A\mathbf{x}}{\|\mathbf{x}\|^2} = \max_{\|\mathbf{x}\|=1, \mathbf{x} \in V_k^\perp} \mathbf{x}^\top A\mathbf{x}$$

Sketch of Proof based on an optimization-theoretic approach:

We first find the critical points of the function $\mathbf{x}^\top A\mathbf{x}/\|\mathbf{x}\|^2$, i.e. the points where the derivative of $\mathbf{x}^\top A\mathbf{x}/\|\mathbf{x}\|^2$ with respect to \mathbf{x} is zero. First,

$$\begin{aligned} \frac{\partial}{\partial x_i} (\mathbf{x}^\top A\mathbf{x}) &= \frac{\partial}{\partial x_i} \left(\sum_{r,s} a_{r,s} x_r x_s \right) = 2(A\mathbf{x})_i \\ \frac{\partial}{\partial x_i} (\|\mathbf{x}\|^2) &= \frac{\partial}{\partial x_i} \left(\sum_r x_r^2 \right) = 2x_i. \end{aligned}$$

Remember that $\left(\frac{f}{g}\right)' = 0 \iff f'g = fg'$. Thus, $\frac{\partial \mathbf{x}^\top A\mathbf{x}/\|\mathbf{x}\|^2}{\partial x_i} = 0$ for every i if and only if:

$$\forall i, (A\mathbf{x})_i \|\mathbf{x}\|^2 = (\mathbf{x}^\top A\mathbf{x}) x_i \iff \|\mathbf{x}\|^2 A\mathbf{x} = (\mathbf{x}^\top A\mathbf{x}) \mathbf{x} \iff A\mathbf{x} = \lambda \mathbf{x}, \text{ where } \lambda = \frac{\mathbf{x}^\top A\mathbf{x}}{\|\mathbf{x}\|^2}$$

This shows that the derivative vanishes at \mathbf{x} exactly when \mathbf{x} is one of the eigenvectors of A . The maximal value $\lambda_1 = \mathbf{x}^\top A \mathbf{x} / \mathbf{x}^\top \mathbf{x}$ is attained when we take the eigenvector \mathbf{v}_1 that correspond to the λ_1 .

A similar argument yields the result for other eigenvalues, but an additional tool is needed here, namely the method of Lagrange multipliers, which we encounter later on in this course. ■

Proof: (based on theorem 36)

Recall that a symmetric real matrix $A \in M_n(\mathbb{R})$ can be written as $A = V\Lambda V^\top$ where V is an orthogonal matrix (whose columns are eigenvectors) and Λ is a diagonal matrix (whose diagonal contains the eigenvalues).

Thus,

$$\mathbf{x}^\top A \mathbf{x} = \mathbf{x}^\top V \Lambda V^\top \mathbf{x} = (V^\top \mathbf{x})^\top \Lambda V^\top \mathbf{x} = \sum_{i=1}^n \lambda_i (V^\top \mathbf{x})_i^2 \leq \max_i \{\lambda_i\} \cdot \sum_{i=1}^n (V^\top \mathbf{x})_i^2 = \lambda_1 \sum_{i=1}^n (V^\top \mathbf{x})_i^2$$

Since V is orthogonal, we have $\sum_{i=1}^n (\mathbf{x}^\top V)_i^2 = \|\mathbf{x}^\top V\|^2 = \|\mathbf{x}\|^2$ and therefore, $\mathbf{x}^\top A \mathbf{x} \leq \lambda_1 \|\mathbf{x}\|^2$. We conclude that $\lambda_1 \geq \max_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{x}^\top A \mathbf{x}) / \|\mathbf{x}\|^2$ and since the quotient is actually λ_1 on the corresponding eigenvector we arrive at the desired result.

We proceed to the rest of the eigenvalues and eigenvectors. For $i \leq k$ it holds that $(V^\top \mathbf{x})_i = 0$ so

$$\mathbf{x}^\top A \mathbf{x} = \sum_{i=k+1}^n \lambda_i (V^\top \mathbf{x})_i^2 \leq \max_{i>k} \{\lambda_i\} \cdot \sum_{i=k+1}^n (V^\top \mathbf{x})_i^2 = \lambda_{k+1} \|\mathbf{x}\|^2$$

and the rest of the argument is as for λ_1 .

Note that since the eigenvectors are orthogonal, the constraint $\mathbf{x} \perp \{v_1, \dots, v_k\}$ is equivalent to the requirement that \mathbf{x} be in the span of the remaining eigenvectors $\mathbf{v}_{k+1}, \dots, \mathbf{v}_n$. The above argument yields the general case of the Rayleigh-Ritz theorem. ■

The expression $\max_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^\top A \mathbf{x}}{\|\mathbf{x}\|^2}$ is called *Rayleigh quotient*. Note that the same argumentation could be carried out using the minimum of the Rayleigh quotient to get λ_n , then $\lambda_{n-1}, \dots, \lambda_1$:

Theorem 39 (Rayleigh-Ritz) *Let $A \in M_n(\mathbb{R})$ be a symmetric real matrix, and let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ be its eigenvalues (in descending order). Then*

$$\lambda_n = \min_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^\top A \mathbf{x}}{\|\mathbf{x}\|^2} = \min_{\|\mathbf{x}\|=1} \mathbf{x}^\top A \mathbf{x},$$

and the minimum is attained by an eigenvector \mathbf{v}_n corresponding to λ_n .

Furthermore, let $\mathbf{v}_n, \dots, \mathbf{v}_{n-k+1}$ denote k eigenvectors corresponding to the eigenvalues $\lambda_n, \dots, \lambda_{n-k+1}$, and let W_k^\perp denote the linear subspace orthogonal to their span (i.e., $W_k^\perp = \{\mathbf{v} \mid \mathbf{v} \perp \text{sp}(\mathbf{v}_n, \dots, \mathbf{v}_{n-k+1})\}$). Then

$$\lambda_{n-k} = \min_{\mathbf{x} \in W_k^\perp} \frac{\mathbf{x}^\top A \mathbf{x}}{\|\mathbf{x}\|^2} = \min_{\|\mathbf{x}\|=1, \mathbf{x} \in W_k^\perp} \mathbf{x}^\top A \mathbf{x}$$

A closely related theorem is due to Courant-Fischer. It can be proved by the Rayleigh-Ritz theorem but we do not prove it here.

Theorem 40 (Courant-Fischer) Let $A \in M_n(\mathbb{R})$ be a symmetric real matrix, and let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ be its eigenvalues (in descending order), then for $i \in \{1, \dots, n\}$:

$$\lambda_{i+1} = \min_{F: \dim(F)=i} \max_{\mathbf{x} \perp F} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2},$$

and the eigenvector corresponding to λ_i is the vector on which the maximum is attained.

Similarly,

$$\lambda_i = \max_{G: \dim(G)=n-i} \min_{\mathbf{x} \perp G} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2}$$

and the eigenvector corresponding to λ_i is the vector on which the minimum is attained.

An easily corollary of Courant-Fischer theorem is the interlacing theorem:

Claim 41 (Interlacing Theorem) Let A be a $n \times n$ real symmetric matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$ and let B be a $n-1 \times n-1$ matrix attained by deleting the i^{th} row and i^{th} column from A for some index i . Let $\mu_1 \geq \dots \geq \mu_{n-1}$ be the eigenvalues of B . Then,

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

Proof: First, for a vector $y \in \mathbb{R}^{n-1}$ we can notate $\widehat{y} \in \mathbb{R}^n$ the 'extension' of y obtained by adding zero in the i^{th} entry. I.e. $\widehat{y}_k = \begin{cases} y_k & k < i \\ 0 & k = i \\ y_{k-1} & k > i \end{cases}$ (Similarly, we will define $\widehat{F} \subset \mathbb{R}^n$ for $F \subset \mathbb{R}^{n-1}$).

Notice that

- $\|\widehat{y}\| = \|y\|$
- $\widehat{y}^T A \widehat{y} = y^T B y$
- for $x \in \mathbb{R}^n$ [$x_k = 0 \iff \langle x, e_k \rangle = 0$]

and therefore we can write

$$\begin{aligned} \mu_{i+1} &= \min_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=i}} \max_{\substack{\mathbf{x} \in \mathbb{R}^{n-1} \\ \mathbf{x} \perp F}} \frac{\mathbf{x}^T B \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \min_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=i}} \max_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \widehat{\mathbf{x}} \perp \widehat{F}}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \min_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=i}} \max_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp \widehat{F} \\ \mathbf{x} \perp e_i}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \min_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=i \\ G \subset \mathbb{R}^n \\ G = \text{span}(\widehat{F} \cup \{e_i\})}} \max_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp G}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &\geq \min_{\substack{G \subset \mathbb{R}^n \\ \dim(G)=i+1}} \max_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp G}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \lambda_i + 2 \end{aligned}$$

■

$$\begin{aligned} \mu_i &= \max_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=(n-1)-i}} \min_{\substack{\mathbf{x} \in \mathbb{R}^{n-1} \\ \mathbf{x} \perp F}} \frac{\mathbf{x}^T B \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \max_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=(n-1)-i}} \min_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \widehat{\mathbf{x}} \perp \widehat{F}}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \max_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=(n-1)-i}} \min_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp \widehat{F} \\ \mathbf{x} \perp e_i}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \max_{\substack{F \subset \mathbb{R}^{n-1} \\ \dim(F)=(n-1)-i \\ G \subset \mathbb{R}^n \\ G = \text{span}(\widehat{F} \cup \{e_i\})}} \min_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp \widehat{F}}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &\leq \max_{\substack{G \subset \mathbb{R}^n \\ \dim(G)=n-i}} \min_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \perp \widehat{F}}} \frac{\mathbf{x}^T A \mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \lambda_i \end{aligned}$$

19 Perron-Frobenius Theorem

In this section we consider matrices whose entries are nonnegative reals. You should be aware, though, that sometimes people speak about "nonnegative matrices" when they mean "positive semi-definite". As mentioned, in the present section we simply mean that the entries in M are nonnegative. Similarly, we consider "positive matrices" whose entries are positive reals. We start with the following simple instance of Perron's theorem.

Lemma 42 *Let $M \in M_n(\mathbb{R}^+)$ be a real symmetric positive matrix and let \mathbf{v} be an eigenvectors corresponding to M 's largest eigenvalue. Then $\mathbf{v} \geq 0$*

Proof: By the SVD theorem the matrix $\lambda \mathbf{v} \mathbf{v}^\top$ is the best rank-1 approximation for M . But clearly,

$$\|M - \lambda \mathbf{v} \mathbf{v}^\top\|_F \leq \|M - |\lambda| \mathbf{u} \mathbf{u}^\top\|_F,$$

where \mathbf{u} is the vector defined by $u_i = |v_i|$ for every i . By the optimality of \mathbf{v} this inequality must hold with equality. Consequently, $\lambda \geq 0$ and $\mathbf{v} \geq 0$. ■

Perron-Frobenius theorem describes the leading eigenvalue-eigenvector pair of nonnegative matrix. It provides conditions under which the leading eigenvalue is unique and the leading eigenvector to be strictly positive. It is simpler to start with the strictly positive case (Perron theorem) and then reduce the nonnegative case to the Perron theorem.

Theorem 43 (The symmetric case of Perron's theorem) *Let $M \in M_n(\mathbb{R}^+)$ be a real symmetric positive matrix. Let λ be its largest eigenvalue and let \mathbf{v} be its corresponding eigenvector. Then, λ is simple (i.e. it has multiplicity 1), the vector \mathbf{v} is positive and*

$$\lim_{m \rightarrow \infty} \left(\frac{M}{\lambda} \right)^m = \mathbf{v} \mathbf{v}^\top$$

Proof: The SVD theorem implies that the matrix $V = \lambda \mathbf{v} \mathbf{v}^\top$ is the best rank-1 approximation of M . As we saw, $\mathbf{v} \geq 0$. If it has a zero coordinate, say $v_i = 0$ let $\mathbf{u} = \mathbf{v} + \epsilon \mathbf{e}_i$. Since M is strictly positive, it is easy to verify that there is a positive value of ϵ for which $\|M - \lambda \mathbf{v} \mathbf{v}^\top\|_F > \|M - \lambda \mathbf{u} \mathbf{u}^\top\|_F$.

The vector \mathbf{v} is the unique eigenvector corresponding to the largest eigenvalue λ . Otherwise we can find another eigenvector \mathbf{v}_2 corresponding to λ . We can choose it to be orthogonal to \mathbf{v} , i.e. $\langle \mathbf{v}, \mathbf{v}_2 \rangle = 0$. This is impossible, since $\mathbf{v}_2 \geq 0$ by Lemma 42 and $\mathbf{v} > 0$ as we just saw.

The eigenvalues of M/λ are $1 > \lambda_2/\lambda \geq \dots \geq \lambda_n/\lambda$.

We prove $\lambda > -\lambda_n$ using the Rayleigh-Ritz theorem as follows:

$$\lambda = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top M \mathbf{x} \quad \text{and} \quad \lambda_n = \min_{\|\mathbf{y}\|=1} \mathbf{y}^\top M \mathbf{y}$$

Let \mathbf{y} be the unit eigenvector corresponding to λ_n . I.e. Let \mathbf{x} be the vector defined by $\forall i x_i = |y_i|$.

$$\begin{aligned}
 \text{Then, } \sum_{i,j} m_{ij} y_i y_j &= \mathbf{y}^t M \mathbf{y} \\
 &= \lambda_n \\
 \sum_{i,j} m_{ij} x_i x_j &= \mathbf{x}^t M \mathbf{x} \\
 &\leq \lambda \\
 y_i y_j + x_i x_j &\in [0, 2y_i y_j] \\
 y_i y_j + x_i x_j &\geq 0 \\
 \lambda - \lambda_n &\geq \sum_{i,j} m_{ij} y_i y_j + \sum_{i,j} m_{ij} x_i x_j \\
 &= \sum_{i,j} m_{ij} [y_i y_j + x_i x_j] \\
 &\geq 0
 \end{aligned}$$

The matrix M/λ can be written as $U\Lambda U^T$ with $\Lambda_{11} = 1$ and $|\Lambda_{ii}| < 1$ for $i = 2, \dots, n$. Consequently $(M/\lambda)^k = U\Lambda^k U^T$, but the limit of Λ^k is the matrix with one in the (1, 1) entry and zero elsewhere. It follows that $(M/\lambda)^k$ tends to $\mathbf{v}\mathbf{v}^T$ as claimed. ■

To generalize Perron's theorem to the nonnegative case we take a combinatorial perspective of the problem and determine when there is a power M^k that is strictly positive, and find the relation between the eigenvalues-eigenvectors pairs of M and of M^k via the SVD theorem. We construct the undirected graph $G = (V, E)$ with an edge between v_i and v_j if and only if $M_{ij} > 0$. Let A be the adjacency matrix of G . Since M is nonnegative, clearly $A^k > 0$ if and only if $M^k > 0$. The (i, j) entry of A^k is known to be the number of paths of length k from i to j in the graph G . This observation is useful in establishing the nonnegative counterpart to Perron's theorem.

Theorem 44 (The symmetric case of Perron-Frobenius) *Let $M \in M_n(\mathbb{R}^+)$ be a real symmetric nonnegative matrix and let A be its indicator matrix, i.e. $A_{ij} = 1$ if $M_{ij} > 0$ and zero otherwise. Let $G = (V, E)$ be the undirected graph whose adjacency matrix is A , and assume that G is connected and non-bipartite. Then λ the leading eigenvalue of M is simple, the corresponding eigenvector \mathbf{v} is positive, and*

$$\lim_{m \rightarrow \infty} \left(\frac{M}{\lambda} \right)^m = \mathbf{v}\mathbf{v}^T$$

Proof: As mentioned above $M^k > 0$ iff $A^k > 0$. We prove that indeed there is some integer k such that $A^k > 0$, i.e. for every two vertices i, j in G there is a path of length k connecting i and j . We first consider the case $i = j$ and show that there is a number k such that $A^k_{jj} > 0$ for every j . We then use the connectivity of G to show a similar relation for all pairs i, j .

In a connected undirected graph there is always a closed walk of even length (just go back and forth on some edge), hence $A^k_{jj} > 0$ for $k = 2, 4, 6, \dots$. The graph G is not bipartite therefore it contains a cycle C of some odd length c . G is connected thus every vertex j is on a closed walk of odd length $l_j \leq 2n + c$ where n is the number of vertices in G . Let k be the least common multiplicity of all the integers l_j . Then $A^k_{jj} > 0$ for every j . Let B be the indicator matrix of A^k . Then we can write $B = I + C$ where C is the matrix of the graph G^k . Now $C^n > 0$, since $C^n = (I + B)^n = \sum \binom{n}{j} B^j$ and the claim is a simple consequence of G 's connectivity.

It is an easy observation that $M\mathbf{v} = \lambda\mathbf{v}$ implies $M^k\mathbf{v} = \lambda^k\mathbf{v}$. consequently, if (λ, \mathbf{v}) is spectral pair of M , then (λ^k, \mathbf{v}) is a pair of M^k . Since $M^k > 0$, we can apply Perron's theorem and conclude the proof. ■

The Perron-Frobenius theorem holds for general nonnegative matrices but the uniqueness and the existence of leading eigenvalue and left and right eigenvectors require more work which we not do here. We state the general theorem without a proof:

Theorem 45 (Perron-Frobenius) *Let $M \in M_n(\mathbb{R}^+)$ be a real nonnegative matrix and let*

$$\rho(M) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M\}$$

be its spectral radius. Let the directed graph $G = (V, E)$ corresponding to M satisfy the following two conditions:

1. *G is strongly connected. That is, for every two vertices $i, j \in V$, there is a directed path from i to j and one from j to i .*
2. *V can not be partitioned into subsets S_1, \dots, S_d such that $E \subseteq (S_1 \times S_2) \cup (S_2 \times S_3) \cup \dots \cup (S_{d-1} \times S_d) \cup (S_d \times S_1)$.*

Then, $\rho(M)$ is a simple eigenvalue of M and its corresponding eigenvectors are positive. In addition, if \mathbf{x}_r and \mathbf{x}_ℓ are the right and left eigenvalues that correspond to $\rho(M)$, namely, $M\mathbf{x}_r = \rho(M)\mathbf{x}_r$ and $\mathbf{x}_\ell^\top M = \rho(M)\mathbf{x}_\ell^\top$, then

$$\lim_{m \rightarrow \infty} \left(\frac{M}{\rho(M)} \right)^m = \frac{\mathbf{x}_r \mathbf{x}_\ell^\top}{\langle \mathbf{x}_r, \mathbf{x}_\ell \rangle}.$$

Lecture 8

Lecturer: Nati Linial

Scribe: Yoav Lustig

Last Update: 19 Feb 2009 9:39 a.m.

This lecture presents some applications of spectral properties of non-negative matrices. We first introduce the notion of Markov chain and its relation to non-negative matrices. We then associate between non-negative matrices in $M_n(\mathbb{R})$ and directed graphs on n vertices. Finally we discuss the relation between expander-graphs and rapidly mixing Markov chains.

20 Markov Chains

A Markov chain is a random (or stochastic) process. We first introduce the notion of Markov chain and discuss some of its basic properties. We then use Perron's theorem to investigate the limiting distribution of Markov chains.

We deal with a random variable X_t that takes values in $\{1, \dots, n\}$. We call elements in $\{1, \dots, n\}$ *states*. A Markov chain is a sequence of random variables X_1, \dots, X_T . Intuitively, we think of a Markov chain as a process moving from a state i in time t to a state j in time $t + 1$ with probability $P(X_{t+1} = j | X_t = i)$. Thus a Markov chain is completely determined by its initial distribution $f^{(1)} = P(X_1 = i)$ and its transition probability $P_{ij} = P(X_{t+1} = j | X_t = i)$, we described it by $(f^{(1)}, P)$ where $f^{(1)} \in \mathbb{R}^n$ is a *distribution vector* and $P \in \mathbb{R}^{n \times n}$ is called a *stochastic matrix*⁵.

We investigate the algebraic properties of the Markov process $(f^{(1)}, P)$. we state three simple properties:

1. Let $f^{(t)}$ be the distribution vector $P(X_t)$. Following the calculation $P(X_{t+1} = j) = \sum_{i=1}^n P(X_t = i)P(X_{t+1} = j | X_t = i)$ we induce that $f^{(t+1)} = f^{(1)}P^t$.
2. The transition matrix P has the positive right eigenvector $\bar{1}$ with eigenvalue 1, and $\rho(P) = 1$ is the spectral radius of the matrix P , since for every $i \in \{1, \dots, n\}$: $\sum_j P_{ij} = \sum_j P(X_t = j | X_{t-1} = i) = 1$
3. The non-negative left eigenvector π of the matrix P corresponding to the eigenvalue $\rho(P) = 1$ is the stationary distribution of the Markov process, i.e., if $f^{(t)} = \pi$ then $f^{(t+1)} = \pi P = \pi$.

We will use Perron's theorem to show that for every Markov process with initial distribution $f^{(1)}$ and transition matrix P , the distribution $f^{(1)}P^t$ approaches as $t \rightarrow \infty$ to the stationary distribution π . By Perron's theorem:

$$\lim_{t \rightarrow \infty} \left(\frac{P}{\rho(P)} \right)^t = \frac{\mathbf{x}_r \otimes \mathbf{x}_\ell}{\langle \mathbf{x}_r, \mathbf{x}_\ell \rangle} .$$

⁵Some define P^T as stochastic matrix. In this case the multiplication of stochastic matrix with distribution vector changes sides: $(f^{(1)}P)^T = P^T f^{(1)T}$

where x_r and x_ℓ are the right and left eigenvectors that correspond to $\rho(A)$. As we just mentioned, $x_r = \bar{1}$ and $\rho(P) = 1$. So,

$$\lim_{t \rightarrow \infty} P^t = \bar{1} \otimes x_\ell$$

It is easy to check that $\bar{1} \otimes x_\ell$ is a matrix each row of which is x_ℓ . Consequently, if u is any distribution vector, i.e. $\sum_i u_i = 1, u \geq 0$, we get $\lim_{t \rightarrow \infty} uP^t = x_\ell$. This means that the process tends to the same limit distribution regardless of its initial distribution.

For practical purposes, it is important to know the rate of convergence. The eigenvalues of a positive stochastic matrix are $1, \lambda_2, \dots, \lambda_n$, where $1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$. The rate of convergence is determined by $|\lambda_2|$. We define the spectral gap to be $1 - |\lambda_2|$. A large spectral gap translates to a fast convergence rate. We will say more about this below.

21 Graphs

A graph $G = (V, E)$ has vertex set V and edge set E .

Definition 46 (regular graphs) Let d be a natural number, a graph $G = (V, E)$ is d regular if every vertex has exactly d incident edges.

Definition 47 (adjacency matrix) The adjacency matrix associated with graph $G = (V, E)$ is a $A \in M_n(\mathbb{R})$ defined by

$$A_{ij} = \begin{cases} 1 & (i, j) \in E \\ 0 & (i, j) \notin E \end{cases}$$

Definition 48 (Periodic graphs) A directed graph $G = (V, E)$ is Periodic with period k if there exists a partition of V into k parts S_1, \dots, S_k such that:

1. The parts are disjoint (i.e. if $i \neq j$ then $S_i \cap S_j = \emptyset$).
2. The parts cover V (i.e. $V = \bigcup_{i=1}^k S_k$).
3. All the edges are between consecutive parts (mod k). That is, if $j \neq i + 1 \pmod{k}$ then $E \cap (S_i \times S_j) = \emptyset$.

Definition 49 (bipartite graphs) A graph $G = (V, E)$ is said to be bipartite if there is a decomposition $V = A \dot{\cup} B$ so that all edges in G connect a vertex in A and a vertex in B .

Note that an undirected graph is periodic iff it is bipartite (since if there is an edge from S_i to S_{i+1} there is also an edge from S_{i+1} to S_i).

A Stochastic matrix (a non-negative matrix which its rows sum up to 1) can be seen as *transition matrix* of some Markov chain, in which the states are $\{1, \dots, n\}$ and the probability of transition from state i to state j is given by A_{ij} .

For a graph $G = (V, E)$ and a set of vertices $S \subseteq V$ denote by \bar{S} the complement of S (i.e. $\bar{S} = V \setminus S$). Denote by $e(S, \bar{S})$ the number of edges between S and \bar{S} (i.e. $2|E \cap (S \times \bar{S})|$) and by $e(S)$ the number of edges within S (i.e. $|E \cap S \times S|$). Similarly, $e(\bar{S})$ denotes the number of edges within \bar{S} .

21.1 Random walks on graphs

Given a undirected graph $G = (V, E)$ with nonnegative weights w_{ij} to the edges, we construct a Markov chain with n states. The transition probability from state i to state j is defined to be $P_{i,j} = \frac{w_{ij}}{\sum_{k=1}^n w_{ik}}$. Let $w_i = \sum_{j=1}^n w_{ij}$ and $w = \sum_{i=1}^n w_i$ then $\pi = (w_1, \dots, w_n)/w$ is the left eigenvector of P :

$$(\pi P)_j = \sum_{i=1}^n \pi_i P_{i,j} = \sum_{i=1}^n \frac{w_i}{w} \frac{w_{ij}}{w_i} = \sum_{i=1}^n \frac{w_j}{w} \frac{w_{ji}}{w_j} = \frac{w_j}{w} \sum_{i=1}^n \frac{w_{ji}}{w_j} = \frac{w_j}{w} = \pi_j$$

Since we view a graph as Markov process we define a random walk on a graph as the running of the Markov process. Intuitively, from every node v the random walk continues to one of its neighbors according to the probability $P_{i,j}$.

we see that in a d -regular graph both the left and the right eigenvectors of ρ are $\vec{1}$. Therefore, by Perron-Frobenius theorem, the limit distribution is $\frac{1}{n} \vec{1}$ (namely, the uniform distribution):

Corollary 50 *For any $d > 0$, let G be a d -regular non-bipartite connected graph, and x an arbitrary probability distribution on its vertices. The probability to be in some vertex $v \in V$ after a random walk of length m with starting distribution x is $\frac{1}{n} + o(1)$. (Where $o(1)$ converges to 0 as m goes to ∞ .)*

The question about the rate of converges arises naturally. We have seen earlier, that the rate of convergence depends on the second eigenvalue $|\lambda_2|$ and its distance from the first eigenvalue λ_1 (which is 1 for a Markov chain). In the next section we will see a family of graphs in which the random walk converges exponentially fast. That is, after m steps the probability to be in a vertex $v \in V$ is $\frac{1}{n} + e^{-\Omega(m)}$. (Note that the transition matrix P is $\frac{1}{d}A$ where A is the adjacency matrix. Therefore the second eigenvalue $|\lambda_2|$ of P is $\frac{1}{d}$ times the second eigenvalue of A .)

The difference between the first and second eigenvalues is the *spectral gap*.

22 Expanders

Intuitively, expander graphs (a.k.a. expanders) are graphs without "bottlenecks". That is, for any subset of vertices S and its complement \bar{S} , the number of edges between S and \bar{S} is fairly large (when compared to the size of S).

Definition 51 (expander graphs) *Let $\delta \geq 0$, a graph $G = (V, E)$ is δ -edge expanding if for every set S of size at most $\frac{|V|}{2}$, it holds that $e(S, \bar{S}) \geq \delta|S|$.*

We will define the (edge) expansion of a graph $h(G)$ as the minimal expansion of a vertices set. I.e.

$$h(G) = \min_{\substack{S \subset V \\ |S| \leq \frac{|V|}{2}}} \frac{e(S, \bar{S})}{|S|}$$

It turns out that expanders are graphs with large spectral gaps. There are theorems for both directions. First, if the spectral gap is large then the graph is an expander (we will prove such

a theorem). Second, if the graph is an expander then the spectral gap is large. This is a harder theorem which we will not prove here. Formally, we will prove

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

Theorem 52 *Let $G = (V, E)$ be an undirected, d -regular non-bipartite graph with no loops and no parallel edges and let A be the adjacency matrix of G . Let $\lambda_1 \geq \dots \geq \lambda_n$ be the eigenvalues of A in descending order. Then G is δ -edge expanding for $\delta = \frac{d - \lambda_2}{2}$. I.e. $h(G) \geq \frac{d - \lambda_2}{2}$*

Proof: By Rayleigh-Ritz theorem $\lambda_2 = \max_{x \perp \vec{1}} \frac{x^t A x}{\|x\|^2}$.

For an arbitrary set $S \subseteq V$ with $|S| \leq \frac{|V|}{2}$ we define the following vector x : $x_i = \begin{cases} |\bar{S}| & i \in S \\ -|S| & i \in \bar{S} \end{cases}$

It is easy to see that $x \perp \vec{1}$ since the sum of its coordinates is 0. Therefore $\lambda_2 \geq \frac{x^t A x}{\|x\|^2}$. Next we compute the value of $\frac{x^t A x}{\|x\|^2}$.

$$\text{First, } \|x\|^2 = \sum_{i=1}^n x_i^2 = |S| |\bar{S}|^2 + |\bar{S}| |S|^2 = |S| |\bar{S}| (|S| + |\bar{S}|) = |S| |\bar{S}| n.$$

Next, $x^t A x = \sum_{i,j} x_i x_j A_{ij}$. However, A_{ij} is 0 unless there is an edge between i and j . Therefore, the sum can be split into three cases:

1. First, $(i, j) \in E$ and $i, j \in S$. In this case $x_i x_j = |\bar{S}|^2$ and these terms contribute to the sum a total of $2e(S) |\bar{S}|^2$. (Note that each edge is counted from both ends.)
2. Next, $(i, j) \in E$ and $i, j \in \bar{S}$. In this case $x_i x_j = |S|^2$ and this contributes to the sum a total of $2e(\bar{S}) |S|^2$.
3. Finally, $(i, j) \in E$ and one of the indices is in S while the other in \bar{S} . In this case $x_i x_j = -|S| |\bar{S}|$ and this contributes to the sum a total of $-2e(S, \bar{S}) |S| |\bar{S}|$.

$$\text{Thus, } x^t A x = \sum_{i,j} x_i x_j A_{ij} = 2(e(S) |\bar{S}|^2 + e(\bar{S}) |S|^2 - e(S, \bar{S}) |S| |\bar{S}|).$$

Since the graph is d -regular we also know that the number of edges touching a vertex in S is $d|S|$. These edges can be partitioned into edges within S and edges touching a vertex in \bar{S} . Therefore $d|S| = 2e(S) + e(S, \bar{S})$ (note that the edges within S are counted twice). Similarly $d|\bar{S}| = 2e(\bar{S}) + e(S, \bar{S})$. Denoting $e(|S|, |\bar{S}|)$ by n_e we do the bookkeeping and get:

$$\begin{aligned} x^t A x &= 2(e(S) |\bar{S}|^2 + e(\bar{S}) |S|^2 - n_e |\bar{S}| |S|) \\ &= 2\left(\frac{1}{2}(d|S| - n_e) |\bar{S}|^2 + \frac{1}{2}(d|\bar{S}| - n_e) |S|^2 - n_e |\bar{S}| |S|\right) \\ &= d|S| |\bar{S}| (|\bar{S}| + |S|) - n_e (|\bar{S}|^2 + 2|\bar{S}| |S| + |S|^2) \\ &= d|S| |\bar{S}| n - n_e n^2 \end{aligned}$$

Therefore $\lambda_2 \geq \frac{x^t A x}{\|x\|^2} = d - n_e \frac{n}{|S| |\bar{S}|}$. Since $|\bar{S}| \geq \frac{n}{2}$ we get $\lambda_2 \geq d - n_e \frac{2}{|S|}$ i.e. $\frac{n_e}{|S|} = \frac{e(S, \bar{S})}{|S|} \geq \frac{d - \lambda_2}{2}$ as claimed. ■

Lecture 9

Lecturer: Nati Linial

Scribe: Tamir Hazan

Last Update: 19 Feb 2009 9:39 a.m.

An Optimization problem is determined by a set \mathcal{D} , called the *domain*, and a real-valued function $f : \mathcal{D} \rightarrow \mathbb{R}$, called the *objective function*. We consider $f(\mathbf{x}) \in \mathbb{R}$ as the "worth" or the "cost" associated with $\mathbf{x} \in \mathcal{D}$ and we seek to maximize or minimize $f(\mathbf{x})$ over $\mathbf{x} \in \mathcal{D}$. In general, f may fail to have an optimum in \mathcal{D} . However, if \mathcal{D} is compact and f is continuous then it attains both its maximum and minimum in \mathcal{D} . This will be the case in most of the optimization problems that we consider here.

In the simplest case of this general problem, a linear objective function is to be optimized over a domain \mathcal{D} defined by finite list of linear equalities and inequalities. The linear function to be maximized (or minimized) is called the *objective function* and has the form $\mathbf{c}^\top \mathbf{x} = c_1 x_1 + \dots + c_n x_n$. Minimizing the objective function $\mathbf{c}^\top \mathbf{x}$ is equivalent to maximizing $-\mathbf{c}^\top \mathbf{x}$, and hence we can always pass to a maximization problem. Linear equality constraint $a_1 x_1 + \dots + a_n x_n = b$ can be represented with two inequality constraints: $a_1 x_1 + \dots + a_n x_n \leq b$ and $a_1 x_1 + \dots + a_n x_n \geq b$. Moreover, linear inequality constraint of the form $a_1 x_1 + \dots + a_n x_n \geq b$ is equivalent to the constraint $-a_1 x_1 - \dots - a_n x_n \leq -b$, thus we can assume all inequalities are of the same sign. After such modifications each linear program can be expressed in its *canonical* form:

$$\begin{array}{ll} \text{Maximize the value of} & \mathbf{c}^\top \mathbf{x} \\ \text{among all vectors } \mathbf{x} \in \mathbb{R}^n \text{ satisfying} & \mathbf{Ax} \leq \mathbf{b} \end{array}$$

The relation $\mathbf{Ax} \leq \mathbf{b}$ is a shorthand for the condition $(\mathbf{Ax})_i \leq b_i$ for every i . Any vector $\mathbf{x} \in \mathbb{R}^n$ satisfying all constraints of a given linear program is a *feasible solution*. Each $\mathbf{x}^* \in \mathbb{R}^n$ that attains the maximum possible value of $\mathbf{c}^\top \mathbf{x}$ is called an *optimal solution* or an *optimum*. A linear program may in general have a single optimal solution, or infinitely many solutions, or none at all. A linear program that has no feasible solutions is called *infeasible*. However, a linear program may have no optimal solution even when there are feasible solutions. This happens e.g. when the objective function can attain arbitrarily large values (such a program is called *unbounded*).

There are two important features of linear programming one should keep in mind: A linear program is efficiently solvable, both in theory and in practice. In practice a number of software packages are available to this end, e.g. *linprog* in Matlab. These packages can handle linear programs with thousands of variables and constraints. In theory, algorithms have been developed that provably solve each linear program in time bounded by a polynomial in the input size.

23 Linear Programming — Examples

Linear programming turns out to be a general framework that includes many seemingly unrelated problems. We present a few examples and demonstrate several tricks that allow us to deal with problems that do not appear like linear programs at first sight.

23.1 The diet problem

A farmer wants his cow to be as skinny as possible while still keeping her healthy. There are n different food types available, the j -th food type having $c_j \in \mathbb{R}$ calories per kilogram, $1 \leq j \leq n$, and $a_{ij} \in \mathbb{R}$ milligrams of vitamin i per kilogram, $1 \leq i \leq m$. In order to stay in good health the cow should consume at least $b_i \in \mathbb{R}$ milligrams of vitamin i a day. Given that the goal is to minimize the cow's daily caloric intake while keeping the supply of each vitamin above the threshold, how should she be fed? Letting x_j be the number of kilograms of food j the cow is fed in a day, we get the following linear program:

$$\begin{aligned} & \text{minimize} && \mathbf{c}^\top \mathbf{x} \\ & \text{subject to} && x_j \geq 0 && \text{for } j = 1, \dots, n \\ & && \sum_{j=1}^n a_{ij} x_j \geq b_j && \text{for } i = 1, \dots, m \end{aligned}$$

23.2 Separation with margin

Let $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be a set of white points and $\{(u_1, v_1), (u_2, v_2), \dots, (u_m, v_m)\}$ be a set of black points in \mathbb{R}^2 . We would like to find whether there exists a straight line having all the white points on one side and all the black points on the other side. Let $y = ax + b$ be the equation of the line. A point is above the line if $y_i > ax_i + b$ and below the line if $v_j < au_j + b$ so a suitable line exists if and only if we can find a and b for which

$$\begin{aligned} y_i &> ax_i + b && \text{for } i = 1, \dots, n \\ v_j &< au_j + b && \text{for } j = 1, \dots, m \end{aligned}$$

Among all the lines that separate the white and black points we wish choose the one with the largest margin with respect to the second axis of \mathbb{R}^2 :

$$\begin{aligned} & \text{maximize} && \epsilon \\ & \text{subject to} && y_i - \epsilon \geq ax_i + b && \text{for } i = 1, \dots, n \\ & && v_j + \epsilon \leq au_j + b && \text{for } j = 1, \dots, m \end{aligned}$$

A plane that separates two point sets in \mathbb{R}^3 can be computed by the same approach, and we can also likewise solve the analogous problem in higher dimensions.

Assume that we cannot separate the points by a straight line, then we could try to separate the points by a graph of quadratic function (a parabola) of the form $y = ax^2 + bx + c$. As before we reduce the separation problem to the linear program

$$\begin{aligned} & \text{maximize} && \epsilon \\ & \text{subject to} && y_i - \epsilon \geq ax_i^2 + bx_i + c && \text{for } i = 1, \dots, n \\ & && v_j + \epsilon \leq au_j^2 + bu_j + c && \text{for } j = 1, \dots, m \end{aligned}$$

Even though we consider here a quadratic polynomial, this is a linear program where the variables are the *coefficients* of the polynomial in question.

23.3 Largest disk in convex polygon

Let P be a given convex polygon with n sides. Our goal is to find the largest circular disk contained in P . Let l_i be the straight line that contains the i -th side of P and let its equation be $y = a_i x + b_i$ for $i = 1, \dots, n$. Let us choose a numbering of the sides in such a way that for $i = 1, \dots, k$ the lines l_i bound P from below and for $i = k + 1, \dots, n$ the lines l_i bound P from above. A fact that you may recall from plane analytic geometry is that the distance of a point (u, v) from the i 'th line is

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

Also, a disk of radius R around (u, v) lies completely within P if and only if the distance from (u, v) to each of the n lines is at least R . This allows us to formulate the problem as the following linear program with three variables (u, v, R) , where the values of a_i and b_i are the problem constants:

$$\begin{aligned} & \text{maximize} && R \\ & \text{subject to} && R \leq (v - a_i u - b_i) / \sqrt{1 + a_i^2} \quad \text{for } i = 1, \dots, k \\ & && -R \geq (v - a_i u - b_i) / \sqrt{1 + a_i^2} \quad \text{for } i = k + 1, \dots, n \end{aligned}$$

23.4 Fitting a Line

Consider n points in the plane $(x_1, y_1), \dots, (x_n, y_n)$. We seek a line $y = ax + b$ which "best fits" the points. The least squares method considers a line that is optimized according to the following criterion

$$\min_{a, b \in \mathbb{R}} \sum_{i=1}^n (ax_i + b - y_i)^2$$

This method is not always suitable. For example, a small number of points that are given to us with a very large error may badly influence the resulting line. An alternative method, more robust to such "outliers" minimizes the sum of absolute values

$$\min_{a, b \in \mathbb{R}} \sum_{i=1}^n |ax_i + b - y_i|$$

By a simple trick this apparently nonlinear problem can be formulated as a linear program as well:

$$\begin{aligned} & \min && e_1 + e_2 + \dots + e_n \quad \text{subject to} \\ & && e_i \geq ax_i + b - y_i \quad \text{for } i = 1, \dots, n \\ & && e_i \geq -(ax_i + b - y_i) \quad \text{for } i = 1, \dots, n \end{aligned}$$

The variables are a, b and e_1, \dots, e_n while x_i, y_i are given numbers. Each e_i is an auxiliary variable standing for the error at the i -th point. The constraints guarantee that

$$e_i \geq \max(ax_i + b - y_i, -(ax_i + b - y_i)) = |ax_i + b - y_i|.$$

In the optimal solution each of these inequalities has to be satisfied with equality, otherwise we could decrease the corresponding e_i . Thus an optimal solution yields a line minimizing the above expression.

24 Integer Linear Programming - ILP

An *integer linear program* is a linear program in which the variables have to take only integral values

$$\begin{aligned} & \text{maximize} && \mathbf{c}^\top \mathbf{x} \\ & \text{subject to} && \mathbf{Ax} \leq \mathbf{b} \\ & && \mathbf{x} \in \mathbb{Z}^n \end{aligned}$$

Here A is a $m \times n$ matrix, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$, \mathbb{Z} denotes the set of integers, and \mathbb{Z}^n is the set of n -dimensional vectors with integer entries.

Solving an integer linear program is NP-hard. This is in contrast with solving linear programs which can be done in polynomial time (in the input length). An easy way to see the hardness of ILP is to consider the 3-SAT problem. In this well-known NP-complete problem we seek a satisfying assignment to a formula $\bigwedge_{i=1}^m (l_{i1} \vee l_{i2} \vee l_{i3})$ where each literal l_{ij} is a variable from x_1, \dots, x_n or its negation. This problem can be formulated as the following ILP:

$$\begin{aligned} z_{i1} + z_{i2} + z_{i3} &\geq 1 && \text{for } i = 1, \dots, m \\ z_{ij} &= x_{ij} && \text{if the literal stands for the variable } x_{ij} \\ z_{ij} &= 1 - x_{ij} && \text{if the literal stands for the negation of } x_{ij} \\ x_i &\in \{0, 1\} \end{aligned}$$

24.1 Maximum Weight Matching

Many combinatorial optimization problems can be easily formulated as ILP's. For example, we present the maximum weight matching. Some company has n employees and n tasks to perform. Every employee should carry out exactly one task. The benefit of job j if performed by employee i is c_{ij} and we wish to maximize the benefit when we carry out all the jobs. Consider the indicator variable x_{ij} that takes the value 1 if the i 'th worker performs the j 'th job and zero otherwise. We arrive at the following integer program:

$$\begin{aligned} & \text{maximize} && \sum_{ij} c_{ij} x_{ij} \\ & \text{subject to} && \sum_j x_{ij} = 1 && \text{the } i\text{'th worker performs exactly one task} \\ & && \sum_i x_{ij} = 1 && \text{the } j\text{'th task is performed by exactly one employee} \\ & && x_{ij} \in \{0, 1\} \end{aligned}$$

If we replace the integrality constraints $x_{ij} \in \{0, 1\}$ with linear constraints $0 \leq x_{ij} \leq 1$ we obtain a linear program which we can solve. The optimum of this linear program is referred to as a fractional optimal solution. In order to (at least) approximate the integer optimum (which is what we seek), we usually try to round the fractional solution to recover a (typically non-optimal but hopefully near-optimal) integral solution.

Lecture 10

Lecturer: Nati Linial

Scribe: Tamir Hazan

Last Update: 19 Feb 2009 9:39 a.m.

25 A Standard Form for LP

A given linear program can be equivalently formulated in more than one way. For practical as well as for theoretical reasons, it is sometimes advantageous to have it stated in a certain format. We discuss here some of these different options.

Recall that in a linear program we seek $\max\{f(x) : x \in \mathcal{D}\}$, where the domain \mathcal{D} is defined by a finite list of linear equalities and inequalities, and f is linear. As seen in the previous lecture, such a program can be represented in its *canonical form*:

$$\begin{aligned} &\text{Maximize } \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } \mathbf{Ax} \leq \mathbf{b} \end{aligned}$$

A linear program of the type below is said to be in a *standard form*:

$$\begin{aligned} &\text{Maximize } \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } \mathbf{Ax} = \mathbf{b} \\ &\quad \mathbf{x} \geq 0 \end{aligned}$$

Every linear program can be (efficiently) transformed into an equivalent program in standard form. Given a linear program in canonical form, we transform it into one of the standard form as follows.

- We replace every inequality $\langle a_i, x \rangle \geq b_i$ with $\langle a_i, x \rangle - s_i = b_i$, where s_i is a new variable that is required to be nonnegative, i.e. we add the inequality $s_i \geq 0$.
- For each variable x_j , which is not already restricted to nonnegative values, we replace x_j by $y_j - z_j$, where y_j and z_j are new nonnegative variables.

26 Basic Feasible Solutions

In this section A is always a matrix with m rows and n columns ($m \leq n$), of rank m . For a subset $B \subseteq \{1, \dots, n\}$ we let A_B be the matrix consisting of columns of A whose indices belong to B . A point in the set $P = \{\mathbf{x} \mid \mathbf{x} \geq 0, \mathbf{Ax} = \mathbf{b}\}$ is called a feasible solution.

Definition 53 Basic feasible solution

A basic feasible solution of the linear program in standard form

$$\text{Maximize } \mathbf{c}^\top \mathbf{x} \text{ subject to } \mathbf{Ax} = \mathbf{b} \text{ and } \mathbf{x} \geq 0$$

is a feasible solution $\mathbf{x} \in \mathbb{R}^n$ for which there exists an m -element set $B \subseteq \{1, \dots, n\}$ such that:

- The square matrix A_B is non-singular, i.e. the columns indexed by B are linearly independent.
- $B \supseteq \text{supp}(\mathbf{x})$, i.e. $x_j = 0$ for all $j \notin B$.

Lemma 54 A feasible solution \mathbf{x} of a linear program in standard form is basic if and only if the columns of the matrix A_K are linearly independent, where $K = \text{supp}(\mathbf{x})$.

Proof: If \mathbf{x} is basic feasible solution then from the definition there is a set B with $K \subseteq B$ and the columns of A_B are linearly independent.

Conversely, $|K| \leq m$ since $\text{rank}(A) = m$ and we can extend K to a set B of m independent columns and this is the requested B . ■

In the following theorem we show that it suffices to look for optimal solutions among basic feasible solutions.

Theorem 55 Consider a linear program in standard form:

$$\text{Maximize } \mathbf{c}^\top \mathbf{x} \text{ subject to } \mathbf{Ax} = \mathbf{b} \text{ and } \mathbf{x} \geq 0$$

1. If there is at least one feasible solution and the objective function is bounded from above in the set of all feasible solutions, then there exists an optimal solution.
2. If an optimal solution exists, then there is a basic feasible solution that is optimal.

A proof is presented later. The theorem suggests to restrict the optimal solutions search for optimal solutions to the finite set of basic feasible solutions rather than to the infinite set of feasible solutions. In view of this theorem we can consider the following brute force algorithm: enumerate all the sets of m indices $B \subseteq \{1, \dots, n\}$ and check whether they corresponding set of columns in A is linearly independent. If A_B is non-singular recover the basic feasible solution \mathbf{x} which is all zero except for the indices in B and $\mathbf{x}_B = A_B^{-1} \mathbf{b}$. This algorithm considers $\binom{n}{m}$ sets B and may be exponential in the input size m, n , e.g. whenever $n = 2m$. In the Section 28 we describe the simplex method which goes over the basic feasible solutions in a clever way and is very useful in practice.

Proof: of Theorem 55:

We prove that under the assumptions of the theorem, for every feasible solution \mathbf{x}_0 there exists a basic feasible solution \mathbf{x}^* with the same or larger value of the objective function, i.e. $\mathbf{c}^\top \mathbf{x}^* \geq \mathbf{c}^\top \mathbf{x}_0$. Given a feasible solution \mathbf{x}_0 we choose among the feasible solutions \mathbf{x} with $\mathbf{c}^\top \mathbf{x} \geq \mathbf{c}^\top \mathbf{x}_0$ the one with the largest number of zero coordinates, and we call it \mathbf{x}^* . If the columns of $A_{\text{supp}(\mathbf{x}^*)}$ are linearly independent then \mathbf{x}^* is basic feasible solution. Suppose toward contradiction that these columns are linearly dependent. This mean that there is a vector $\mathbf{y} \neq 0$ with $\text{supp}(\mathbf{y}) \subseteq \text{supp}(\mathbf{x}^*)$ satisfying $A\mathbf{y} = 0$. The proof is carried out by constructing a feasible vector $\mathbf{x}^* + \epsilon\mathbf{y}$ for some $\epsilon > 0$ which does not decrease the objective function and has more zeros than \mathbf{x}^* , contrary to the minimality of \mathbf{x}^* .

Assume $\mathbf{c}^\top \mathbf{y} \geq 0$ and $y_j < 0$ for some index j . The vector $\mathbf{x}^* + \epsilon \mathbf{y}$ is feasible for small enough values of ϵ : It is non-negative since $\text{supp}(\mathbf{y}) \subseteq \text{supp}(\mathbf{x}^*)$, and it satisfies $A(\mathbf{x}^* + \epsilon \mathbf{y}) = \mathbf{b}$ since $A\mathbf{y} = 0$. Thus we can find $\epsilon > 0$ which cancels only one coordinate from \mathbf{x}^* while not reducing the objective, i.e. there holds $\mathbf{c}^\top(\mathbf{x}^* + \epsilon \mathbf{y}) \geq \mathbf{c}^\top \mathbf{x}^*$ by the assumption $\mathbf{c}^\top \mathbf{y} \geq 0$. This as noted before is a contradiction to the minimality of \mathbf{x}^* .

We need to explain why we can choose \mathbf{y} while $\mathbf{c}^\top \mathbf{y} \geq 0$ and $y_j < 0$ for some j . If $\mathbf{c}^\top \mathbf{y} < 0$ we change the sign of \mathbf{y} , so we assume that $\mathbf{c}^\top \mathbf{y} \geq 0$. If $\mathbf{y} \geq 0$ then the vector $\mathbf{x}^* + a\mathbf{y}$ is feasible for every $a > 0$ and the linear program is unbounded. ■

27 Basic terms in convexity

A set $C \subseteq \mathbb{R}^n$ is *convex* if for every two points $\mathbf{x}, \mathbf{y} \in C$ it also contains the interval connecting them $\{\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \mid \lambda \in [0, 1]\}$. If \mathbf{x}, \mathbf{y} are at distance $\|\mathbf{x} - \mathbf{y}\| = d$, then the point $\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}$ on the line connecting \mathbf{x} and \mathbf{y} is of distance λd from \mathbf{y} and distance $(1 - \lambda)d$ from \mathbf{x} . A real function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if its epigraph $\{(\mathbf{x}, y) \mid f(\mathbf{x}) \leq y\}$ is a convex set. The geometric interpretation of an epigraph of a function is the area above the graph of the function. Equivalently a function is convex if and only if $f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$ for every $0 \leq \lambda \leq 1$.

The intersection of any family of convex sets is convex. We define the *convex hull* of a set S to be the minimal convex set C containing S , i.e. the intersection of all convex sets that contain S . Equivalently, the convex hull of S is the set of all convex combinations of finitely many points in S :

$$C = \left\{ \sum_{i=1}^m p_i \mathbf{x}_i \mid \mathbf{x}_i \in S, p_i \geq 0, \sum_{i=1}^m p_i = 1 \right\}$$

27.1 Hyperplanes, halfspaces and polyhedra

A *hyperplane* is an affine subspace of dimension $n - 1$. in other words, it is the sets of all solutions to the linear equation

$$\sum_{i=1}^n a_i x_i = b.$$

Hyperplanes in \mathbb{R}^2 are lines and in \mathbb{R}^3 are ordinary planes. A hyperplane divides \mathbb{R}^n to two closed halfspaces

$$\left\{ \mathbf{x} \mid \sum_{i=1}^n a_i x_i \leq b \right\} \quad \text{and} \quad \left\{ \mathbf{x} \mid \sum_{i=1}^n a_i x_i \geq b \right\}$$

A *convex polyhedron* is an intersection of finitely many closed halfspaces. The solution space of a standard linear program $P = \{\mathbf{x} \mid \mathbf{x} \geq 0, A\mathbf{x} = \mathbf{b}\}$ is a convex polyhedron. A bounded convex polyhedron is called *convex polytope*.

The *dimension* of a convex polyhedron $P \subset \mathbb{R}^n$ is the dimension of the smallest affine space containing P . Equivalently, it is the largest d for which P contains $d + 1$ points $\mathbf{x}_0, \dots, \mathbf{x}_d$ such that the vectors $\mathbf{x}_i - \mathbf{x}_0$ for $i = 1, \dots, d$ are linearly independent. This notion corresponds to *affine dimension* which is different from the linear dimension often used in linear algebra.

27.2 Geometric aspects of basic feasible solutions

A vertex of a convex polyhedron is an extreme point. Mathematically, a vertex is defined as a point where some linear functional attains a unique maximum. Thus \mathbf{v} is a vertex of a convex polyhedron P if $\mathbf{v} \in P$ and there is a non-zero vector $\mathbf{c} \in \mathbb{R}^n$ such that $\mathbf{c}^\top \mathbf{v} > \mathbf{c}^\top \mathbf{x}$ for every $\mathbf{x} \in P \setminus \{\mathbf{v}\}$. Geometrically it means that the hyperplane $\{\mathbf{x} \mid \mathbf{c}^\top \mathbf{x} = \mathbf{c}^\top \mathbf{v}\}$ touches P exactly at \mathbf{v} . A cube in \mathbb{R}^3 has 8 vertices (zero dimensional faces) as well as 12 edges (one dimensional faces) and 6 two dimensional faces. A subset $F \subset P$ is a *k-dimensional face* of a convex polyhedron P if F has dimension k and there exist non-zero $\mathbf{c} \in \mathbb{R}^n$ and a real number b such that $\mathbf{c}^\top \mathbf{x} = b$ for all $\mathbf{x} \in F$ and $\mathbf{c}^\top \mathbf{x} < b$ for all $\mathbf{x} \in P \setminus F$. In other words there exists a hyperplane that touches P exactly in F .

Theorem 56 *Let P be the feasible set of a linear program in standard form, then the following two conditions for a point $\mathbf{v} \in P$ are equivalent:*

1. \mathbf{v} is a vertex of the polyhedron P .
2. \mathbf{v} is a basic feasible solution of the linear program.

Proof: The implication $1 \rightarrow 2$ follows from Theorem 55 with \mathbf{c} being hyperplane corresponds to the vertex \mathbf{v} .

$2 \rightarrow 1$: We prove that a basic feasible solution related to a index set B is a vertex in P . Set a vector \mathbf{c} with $c_j = 0$ for every $j \in B$ and $c_j = -1$ otherwise. $\mathbf{c}^\top \mathbf{v} = 0$ and $\mathbf{c}^\top \mathbf{x} \leq 0$ for every $\mathbf{x} \geq 0$. \mathbf{v} is the *only* feasible solution with $\text{supp}(\mathbf{v}) \subseteq B$ since $\mathbf{v}_B = A_B^{-1} \mathbf{b}$ is uniquely determined for non-singular A_B , therefore $\mathbf{c}^\top \mathbf{x} < 0$ for every $\mathbf{x} \in P \setminus \{\mathbf{v}\}$. ■

28 The Simplex Algorithm

The optimal solution of a linear program can be found in the set of basic feasible solution, as stated in Theorem 55. Moreover, in Theorem 56 we describe the basic feasible solutions as the vertices of the feasible set of the linear program. The simplex algorithm (Dantzig, 49) exploits these two observation and is most naturally described from a geometrical point of view. In this perspective we say that two vertices \mathbf{x}, \mathbf{y} of a polytope P are *adjacent* if the interval $[\mathbf{x}, \mathbf{y}]$ is an edge of P , i.e. the interval is one dimensional face of P .

Algorithm 1 (The simplex algorithm) *Find a vertex \mathbf{x} of the feasible set P of the linear program.*

1. *Iterate until convergence:*
 - (a) *Select a neighbor \mathbf{y} of \mathbf{x} , such that $\mathbf{c}^\top \mathbf{y} > \mathbf{c}^\top \mathbf{x}$*
 - (b) **if no such \mathbf{y} exists then terminate**
 - (c) **else $\mathbf{x} \leftarrow \mathbf{y}$**

The algorithm description above is conceptual and there is algebraic work needed to convert it to pseudo-code. Most of the work is to relate the geometrical concepts (such as vertex, edge) to

algebraic ones. The simplex is actually a family of algorithms, where two specific algorithms differ in two main aspects: (1) How to find the initial vertex \mathbf{x} , and (2) From all neighbors \mathbf{y} of \mathbf{x} for which $\mathbf{c}^\top \mathbf{y} > \mathbf{c}^\top \mathbf{x}$, which one to select (Pivoting).

28.1 Efficiency of the Simplex Method

Is the simplex method efficient? In 1972, Klee and Minty found a family of linear-programming problems for which the original version of the simplex method must perform an exponential number of pivot steps [1]. Later, it was shown that this is also the case for other known versions of the simplex. There are, however, versions of the algorithm for which we do not know of any example that leads to an exponential running time. A natural open question is, therefore, whether there are versions of the simplex method that generally run in polynomial time. An even more fundamental problem is this: Consider a polytope P and let G_P be the undirected graph that is defined as follows. The nodes of G_P are the vertices of P and $x \sim y$ is an edge in G_P if P contains the geometric edge $[x, y]$ (i.e. the interval between \mathbf{x} and \mathbf{y} is one-dimensional face in P). Note that the simplex algorithm starts with a node of G_P and reaches an optimal node by following a path in G_P . It is an open question whether the diameter of G_P is bounded by a polynomial function in n . That is, a polynomial that bounds the graph distance between every two nodes. The Hirsch conjecture states that the diameter is, in fact, linear in n . The best result so far gives a bound of $n^{O(\ln n)}$ due to [3]

In practice, the simplex method performs well. Spielman and Teng gave an interesting profound explanation to this phenomenon by means of *smoothed analysis* [2]. In principle, they showed that, given a linear program, the following holds. If one applies a small random perturbation to the problem, the result is, with a high probability, a linear program for which the simplex method terminates in polynomial time.

References

- [1] V. Klee and G. Minty. How good is the simplex algorithm? *Inequalities*, III:159–175, 1972.
- [2] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004.
- [3] G. Kalai and D. Kleitman. Quasi-polynomial bounds for the diameter of graphs of polyhedra *Bull. Amer. math. Soc.*, 26: 315-316, 1992.

Lecture 11

Lecturer: Nati Linial

Scribe: Tamir Hazan

Last Update: 19 Feb 2009 9:39 a.m.

29 Duality

We wish to derive a (tight) upper bound for the primal linear program:

$$\begin{aligned} & \text{Maximize } \mathbf{c}^\top \mathbf{x} \\ & \text{subject to } \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned} \tag{13}$$

Every vector \mathbf{y} which satisfies $\mathbf{y}^\top \mathbf{A} \geq \mathbf{c}$ provides the upper bound $\mathbf{b}^\top \mathbf{y}$ derived from the relations $\mathbf{c}^\top \mathbf{x} \leq \mathbf{y}^\top \mathbf{Ax} = \mathbf{y}^\top \mathbf{b}$ and referred as the weak duality theorem. What is the tightest upper bound we can achieve with this method? This is formulated as a linear program, called the *dual linear program*:

$$\begin{aligned} & \text{Minimize } \mathbf{b}^\top \mathbf{y} \\ & \text{subject to } \mathbf{y}^\top \mathbf{A} \geq \mathbf{c} \end{aligned} \tag{14}$$

The strong duality theorem states the dual linear program provides the tightest upper bound possible: There is a dual feasible solution \mathbf{y}^* satisfying $\mathbf{b}^\top \mathbf{y}^* = \mathbf{c}^\top \mathbf{x}^*$ for some primal feasible solution \mathbf{x}^* .

We take a geometrical approach to prove the strong duality theorem. We first prove a simplified version called the *Farkas lemma* and then substitute a composed matrix to into it thus deriving the duality theorem.

Lemma 57 (Farkas 1902) *Let A be a real matrix with m rows and n columns, and let $\mathbf{b} \in \mathbb{R}^m$ be a vector. There is no nonnegative solution $\mathbf{x} \geq 0$ satisfying $\mathbf{Ax} = \mathbf{b}$ if and only if there exists a vector $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{y}^\top \mathbf{A} \geq 0$ and $\langle \mathbf{y}, \mathbf{b} \rangle < 0$.*

From algebraic point of view Farkas lemma suggests a certificate \mathbf{y} to prove there is no nonnegative solution satisfying $\mathbf{Ax} = \mathbf{b}$. This certificate proves this inconsistency by linearly combining inequalities of the LP, namely $0 \leq \mathbf{y}^\top \mathbf{Ax} = \mathbf{y}^\top \mathbf{b} < 0$.

Farkas lemma has an appealing geometric interpretation. Let $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$ be the columns of the matrix A , then the convex cone generated by these vectors is the set

$$\text{cone}(\mathbf{a}_1, \dots, \mathbf{a}_n) = \{x_1 \mathbf{a}_1 + \dots + x_n \mathbf{a}_n : x_1, \dots, x_n \geq 0\}$$

From a geometric point of view whenever the point \mathbf{b} does not lie in the convex cone generated by $\mathbf{a}_1, \dots, \mathbf{a}_n$, there exists a hyperplane $H = \{\mathbf{z} \in \mathbb{R}^m : \mathbf{y}^\top \mathbf{z} = 0\}$ separating them, such that all the vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ (and thus the cone generated by them) lie on one side, namely $\mathbf{y}^\top \mathbf{a}_i \geq 0$ for $i = 1, \dots, n$, and the point \mathbf{b} lie on the other side, i.e. $\mathbf{y}^\top \mathbf{b} < 0$.

30 Separation theorems

The complete statement of the separation theorems require some basic topological terminology. A subset of \mathfrak{R}^n is called closed if it contains all its limit points. A set is called bounded if it is contained in a ball of some radius. A closed and bounded set is called compact. Weierstrass' theorem states that a continuous function defined on a compact set attains its minimal value in it (Compare to the cases of $[f : (0, 1] \rightarrow \mathfrak{R} f(x)=x]$, $[f : \mathfrak{R} \rightarrow \mathfrak{R} f(x)=-x]$ both does not attain a minimal value in their domain).

Given the vectors $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$, let C be the convex cone they generate. Proving Farkas lemma amounts to showing that for any vector $\mathbf{b} \notin C$ there exists a hyperplane separating it from C and passing through $\mathbf{0}$. The main technical challenge is to describe the properties of the nearest vector $\mathbf{z} \in C$ to the vector \mathbf{b} :

Lemma 58 (The projection lemma): *Let $C \subseteq \mathbb{R}^m$ be a closed convex set*

1. *For every $\mathbf{b} \in \mathbb{R}^m$ there exists a unique vector $\mathbf{z} \in \mathbb{R}^m$ that minimizes $\|\mathbf{x} - \mathbf{b}\|$ for all $\mathbf{x} \in C$. This vector is called the projection of \mathbf{b} on C .*
2. $\mathbf{z} = \underset{\mathbf{x} \in C}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{b}\| \quad \Leftrightarrow \quad \langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \leq 0 \quad \forall \mathbf{x} \in C$

Proof:

1. Let \mathbf{w} be some vector in C . Minimizing $\|\mathbf{x} - \mathbf{b}\|$ over all $\mathbf{x} \in C$ is equivalent to minimizing the same function over the bounded set of all $\mathbf{x} \in C$ satisfying $\|\mathbf{x} - \mathbf{b}\| \leq \|\mathbf{b} - \mathbf{w}\|$. Since the Euclidean norm is a continuous function this minimum is attained. Since the Euclidean norm is a strictly convex function, the minimum is attained at a unique point.

2. \Leftarrow : For every $\mathbf{x} \in C$ there holds

$$\begin{aligned} \|\mathbf{z} - \mathbf{b}\|^2 - \|\mathbf{x} - \mathbf{b}\|^2 &= \|\mathbf{z} - \mathbf{b}\|^2 - \|(\mathbf{x} - \mathbf{z}) - (\mathbf{b} - \mathbf{z})\|^2 \\ &= \|\mathbf{z} - \mathbf{b}\|^2 - \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{b} - \mathbf{z}\|^2 + 2\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \\ &= \|\mathbf{z} - \mathbf{b}\|^2 - \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{b} - \mathbf{z}\|^2 + 2\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \\ &= 2\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle - \|\mathbf{x} - \mathbf{z}\|^2 \\ &\leq 2\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \end{aligned}$$

\Rightarrow : We assume on the contrary there is a vector $\mathbf{x} \in C$ satisfying $\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle > 0$ and construct a vector $\mathbf{z} + t(\mathbf{x} - \mathbf{z})$ on the line between \mathbf{z} and \mathbf{x} which is closer to \mathbf{b} than \mathbf{z} . The set C is convex thus the line between \mathbf{z} and \mathbf{x} is also in C hence the contradiction follows. For every $0 < t < 1$ the vector $\mathbf{z} + t(\mathbf{x} - \mathbf{z})$ is on the line between \mathbf{z} and \mathbf{x} and its distance from \mathbf{b} is

$$\|(\mathbf{z} + t(\mathbf{x} - \mathbf{z})) - \mathbf{b}\|^2 = \|\mathbf{z} - \mathbf{b}\|^2 - t(2\langle \mathbf{b} - \mathbf{z}, \mathbf{x} - \mathbf{z} \rangle - t\|\mathbf{x} - \mathbf{z}\|^2)$$

We set $0 < t < \min\{1, 2\langle \mathbf{b} - \mathbf{z}, \mathbf{x} - \mathbf{z} \rangle / \|\mathbf{x} - \mathbf{z}\|^2\}$ to obtain a vector on the line between \mathbf{z} and \mathbf{x} closer to \mathbf{b} than \mathbf{z} .

■

The nearest point \mathbf{z} is said to be on the *boundary* of the set C . The hyperplane

$$H = \{\mathbf{x} : \langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle = 0\}$$

is called a *supporting hyperplane* since it contains the point $\mathbf{z} \in C$ and its corresponding halfspace $\{\mathbf{x} : \langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \leq 0\}$ contains the set C entirely.

Theorem 59 (Separating theorem) *Let C be a closed convex set and assume $\mathbf{b} \notin C$, then there exists a hyperplane separating \mathbf{b} from C . Moreover, if \mathbf{z} is the nearest vector to \mathbf{b} in the set C then the hyperplane $H = \{\mathbf{x} : \langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle = 0\}$ strongly separates C and \mathbf{b} .*

Proof: The projection lemma asserts that for every $\mathbf{x} \in C$ holds $\langle \mathbf{x} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle \leq 0$. The vector \mathbf{b} lies in the other side of the hyperplane, namely $\langle \mathbf{b} - \mathbf{z}, \mathbf{b} - \mathbf{z} \rangle > 0$, since the vector $\mathbf{b} \notin C$ or equivalently $\mathbf{b} - \mathbf{z} \neq \mathbf{0}$ ■

Farkas lemma is a particular separating theorem, where the closed convex set is a cone. In this case the vector $\mathbf{b} - \mathbf{z}$ is perpendicular to the nearest vector \mathbf{z} and also to the face of the cone containing the vector \mathbf{z} .

Proof: (of Farkas lemma)

\Leftarrow : If there exists a nonnegative vector \mathbf{x} satisfying $A\mathbf{x} = \mathbf{b}$ then for every \mathbf{y} satisfying $\mathbf{y}^T A \geq 0$ holds $\langle \mathbf{y}, \mathbf{b} \rangle = \langle \mathbf{y}, A\mathbf{x} \rangle \geq 0$

\Rightarrow : There is no nonnegative solution $\mathbf{x} \geq 0$ satisfying $A\mathbf{x} = \mathbf{b}$ therefore the vector \mathbf{b} is not contained in the cone generated by the columns of A denoted by $\mathbf{a}_1, \dots, \mathbf{a}_n$. The cone is a closed set ⁶ and Theorem 59 states that for the vector $\mathbf{y} = \mathbf{z} - \mathbf{b}$ the hyperplane $H = \{\mathbf{x} : \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{z}, \mathbf{y} \rangle\}$ strongly separates the cone and the vector \mathbf{b} , namely $\langle \mathbf{a}_i, \mathbf{y} \rangle \geq \langle \mathbf{z}, \mathbf{y} \rangle$ for the vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ in the cone and $\langle \mathbf{b}, \mathbf{y} \rangle < \langle \mathbf{z}, \mathbf{y} \rangle$.

To complete the proof we show that $\langle \mathbf{z}, \mathbf{y} \rangle = 0$. The projection theorem assures $\langle -\mathbf{z}, \mathbf{y} \rangle \leq 0$. We assume on the contrary that $\langle \mathbf{z}, \mathbf{y} \rangle > 0$ and show there is a vector in the ray $\{(1+t)\mathbf{z} : t > 0\}$ that satisfies $\langle t\mathbf{z}, \mathbf{b} - t\mathbf{z} \rangle = 0$. In this case the vector $(1+t)\mathbf{z}$ is closer to \mathbf{b} than \mathbf{z} since it is attained by a projection onto the ray of \mathbf{z} . The set C is a cone thus the ray of \mathbf{z} is also in C hence the contradiction follows. The distance of the vector $(1+t)\mathbf{z}$ and the vector \mathbf{b} is

$$\|(1+t)\mathbf{z} - \mathbf{b}\|^2 = \|\mathbf{z} - \mathbf{b}\|^2 - t(2\langle \mathbf{b} - \mathbf{z}, \mathbf{z} \rangle - t\|\mathbf{z}\|^2).$$

The assumption $\langle \mathbf{z}, \mathbf{y} \rangle > 0$ can be written as $\langle \mathbf{b} - \mathbf{z}, \mathbf{z} \rangle > 0$. We set $0 < t < 2\langle \mathbf{b} - \mathbf{z}, \mathbf{z} \rangle / \|\mathbf{z}\|^2$ to obtain a vector on the ray of \mathbf{z} which is closer to \mathbf{b} than \mathbf{z} . ■

31 Strong duality theorem

The Farkas lemma is a simplified version of the strong duality theorem between the primal linear program in Eqn. 13 and the dual linear program in Eqn. 14.

⁶It is a linear transformation of the nonnegative orthant

Theorem 60 (Strong duality theorem) (It is easy to show the following two forms are equivalent)

- If there is a solution to the primal problem:
$$\left[\begin{array}{ll} \text{Maximize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array} \right] \text{ then,}$$

$$\begin{array}{ll} \text{Maximize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{Ax} = \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array} = \begin{array}{ll} \text{Minimize} & \mathbf{b}^\top \mathbf{y} \\ \text{subject to} & \mathbf{y}^\top \mathbf{A} \geq \mathbf{c} \end{array}$$

- If there is a solution to the primal problem:
$$\left[\begin{array}{ll} \text{Maximize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{Ax} \\ \text{leqslant} & \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array} \right] \text{ then,}$$

$$\begin{array}{ll} \text{Maximize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{Ax} \\ \text{leqslant} & \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array} = \begin{array}{ll} \text{Minimize} & \mathbf{b}^\top \mathbf{y} \\ \text{subject to} & \mathbf{y}^\top \mathbf{A} \geq \mathbf{c} \\ & \mathbf{y} \geq 0 \end{array}$$

Example 61 (The diet problem and its dual problem)

We already saw the problem (with minor modification):

A farmer wants his cow to feed his cow while still keeping her healthy but not spending too much money. There are n different food types available, the j^{th} food type costs $c_j \in \mathfrak{R}$ per kilogram, $1 \leq j \leq n$, and includes $a_{ij} \in \mathfrak{R}$ milligrams of vitamin i per kilogram, $1 \leq i \leq m$. In order to stay in good health the cow should consume at least $b_i \in \mathfrak{R}$ milligrams of vitamin i a day. Given that the goal is to minimize the costs while keeping the supply of each vitamin above the threshold, how should she be fed?

Letting x_j be the number of kilograms of food j the cow is fed in a day, we get the following linear program:

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & x_j \geq 0 \quad \text{for } j = 1, \dots, n \\ & \sum_{j=1}^n a_{ij} x_j \geq b_j \quad \text{for } i = 1, \dots, m \end{array}$$

This problem can be written in a canonical form:

$$\begin{array}{ll} \text{Minimize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{Ax} \geq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{array}$$

- for n : Number of food types
- m : Number of vitamins
- $A \in \mathfrak{R}^{m \times n}$: A_{ij} = milligrams of vitamin i per kilogram in food j
- $c \in \mathfrak{R}^n$: Cost
- $b \in \mathfrak{R}^m$: Needed amount in milligrams of vitamin per day
- $x \in \mathfrak{R}^n$: Amount of food in kilograms

The DLP(Dual Linear Problem) is

$$\begin{aligned} & \text{Maximize } \mathbf{b}^\top \mathbf{y} \\ & \text{subject to } \mathbf{y}^\top A \leq \mathbf{c} \\ & \mathbf{y} \geq 0 \end{aligned}$$

This problem can be seen as:

One generates synthetic vitamins for cows and sell them to the farmer from the previous story. He wants to find a pricing system such that:

- Every synthetic vitamin has a non-negative price

$$\mathbf{y} \geq 0$$

- The price system should be compatible to any natural food. The farmer cannot replace the synthetic foods by any natural food and still get a cheaper diet.

$$\forall j \in [1, n] : \sum_{i=1}^m y_j a_{ij} \leq c_i$$

- From the possible pricing systems obeying the previous constraint, one wants to find the maximal profit per cow (Assuming the farmer will buy exactly the minimal amount of vitamins to the cow).

Three interesting corollaries of the Strong Duality Theorem are

- Min-Cut equals to Max-Flow in graph theory
- Von-Neumann Min-Max theorem in game theory
- Yao's principle in complexity theory

Theorem 62 (Min-Cut = Max-Flow (reminder))

Given a weighted directed graph $G = (V, E \subseteq V \times V, w : E \rightarrow \mathbb{R}_+)$ with a source $s \in V$ and target $t \in V$:

- A cut in the graph is a division of V to two set S, \bar{S} s.t. $s \in S$ and $t \notin S$.
- The weight of a cut is defined to be

$$\sum_{\substack{(u \rightarrow v) \in E \\ u \in S, v \notin S}} w(u, v)$$

- Min-Cut = The minimal weight of a cut in the graph

- A flow in the graph is defined to be a function $f : E \rightarrow \mathfrak{R}_+$ that satisfies:

$$\forall e \in E : f(e) \leq w(e) \forall v \in V \setminus \{s, t\} \sum_{w \rightarrow v} f(w, v) = \sum_{v \rightarrow w} f(v, w)$$

- The capacity of the flow is defined to be

$$\sum_{s \rightarrow v} f(s, v) - \sum_{v \rightarrow s} f(v, sx')$$

- Max-Flow = The maximal capacity of a flow in the graph

Then, for any such graph: Min-Cut=Max-Flow.

Proof: [Min-Cut = Max-Flow]

Here is the linear program LP_0 for the max flow problem:

$$\begin{aligned} & \text{Maximize} && \sum_{e=(s,u)} f_e - \sum_{e=(u,s)} f_e \\ & \text{subject to} && \\ LP_0: & \forall e & : & f_e \leq c_e \\ & \forall u \neq s, t: & \sum_{e=(v,u)} f_e = \sum_{e=(u,v)} f_e \\ & \forall e & : & f_e \geq 0 \end{aligned}$$

I claim that the value of LP_0 is equal to the value of the following linear program LP_1 :

$$\begin{aligned} & \text{Maximize} && \sum_{p \text{ path from } s \text{ to } t} x_p \\ LP_1: & \text{subject to} && \\ & \forall e: & \sum_{p:e \in p} x_p \leq c_e \\ & \forall p: & x_p \geq 0 \end{aligned}$$

To prove one direction of the claim, let $f \neq 0$ be a flow. Let H be a graph that has an edge e whenever $f_e \neq 0$. Then (by flow conservation properties) s is connected to t in H , so there is a path p such that $f_e > 0$ for every $e \in p$. Let x_p be the minimum of f_e for $e \in p$. Let $f'_e = \begin{cases} f_e & e \notin p \\ f_e - x_p & e \in p \end{cases}$. Then f' is a flow on the graph with capacities $c'_e \in \{c_e, c_e - x_p\}$ depending on whether $e \notin p$ or $e \in p$. At least one edge disappears: $c_e > 0, c'_e = 0$. By induction on the number of edges we can model flow f' as $\sum x'_q$ with

$$\begin{aligned} \forall e: & \sum_{q:e \in q} x'_q \leq c_e \\ \forall q: & x'_q \geq 0 \end{aligned}$$

So $f = \sum x'_q + x_p$ can also be written as a linear combination of flow path variables. Thus, given (f_e) feasible for LP_0 , we defined a (x_p) feasible for LP_1 , with same value, and so $\text{Value}(LP_0) \leq \text{Value}(LP_1)$.

To prove the other direction of the claim, given a linear combination of flow path variables, (x_p) , let $f_e = \sum_{p:e \in p} x_p$. It is easy to check that f satisfies the flow constraints. Thus $\text{Value}(LP_1) \leq \text{Value}(LP_0)$. The claim follows.

The dual LP: By the linear programming duality theorem, the value of LP_1 equals the value of its dual, the following linear program LP_2 :

$$\begin{array}{ll}
 & \text{Minimize} \quad \sum_e y_e c_e \\
 LP_2: & \text{subject to} \\
 & \forall p: \sum_{e \in p} y_e \geq 1 \\
 & \forall e: y_e \geq 0
 \end{array}$$

Equivalent integer program: I claim that LP_2 has the same value as the following integer program IP:

$$\begin{array}{ll}
 & \text{Minimize} \quad \sum_e y_e c_e \\
 IP: & \text{subject to} \\
 & \forall p: \sum_{e \in p} y_e \geq 1 \\
 & \forall e: y_e \in \{0, 1\}
 \end{array}$$

One direction is obvious: since LP_2 is a relaxation of IP, we have $\text{Value}(LP_2) \leq \text{Value}(IP)$. To prove the other direction (sketch), given an optimal solution (y_e) to LP_2 , let L denote the set of vertices reachable from s using edges of value $y_e = 0$ only, and $R = V \setminus L$. Note that $s \in L$ and $t \notin L$, so (L, R) is a cut. Let $y'_e = \begin{cases} 1 & e \in L \times R \\ 0 & \text{otherwise} \end{cases}$, and let $\alpha = \min\{y_e | e \in L \times R\}$. Let $y''_e = \frac{y_e - \alpha y'_e}{1 - \alpha}$. We have $y = \alpha y' + (1 - \alpha)y''$, with $\alpha \in (0, 1)$ (It is easy to see that $\forall e y_e \in [0, 1]$ and hence if y is not an integral solution $\alpha \in (0, 1)$).

Clearly, y' is feasible for LP_2 since it defines a cut. As to y'' , clearly $y''_e \geq 0$ for every e . Consider a path p which crosses the cut (L, R) two or more times: write $p = p_1(u, v)p_2$, where (u, v) is the last time that p crosses the cut. Consider the path $p' = P'_1(u, v)p_2$, where p'_1 is a path from s to u that stays entirely in L (it exists by definition of L). Since p' only crosses the cut once, the constraint for p' is satisfied (by an easy calculation); and that implies $\sum_{e \in p} y_e \geq \sum_{e \in p'} y_e \geq 1$, so the constraint for p is also satisfied. This means that y'' is also feasible for LP_2 . Thus y , which is optimal, is a convex combination of two feasible solutions: they must both be optimal as well. But y' is an integer solution. So $\text{Value}(LP_2) = \sum_e c_e y'_e \geq \text{Value}(IP)$. The claim follows.

Relating the integer program to the minimum cut: Finally, I claim that the value of IP is exactly the value of the minimum cut from s to t .

Indeed, given a feasible y , let L be the set of vertices reachable from s using edges such that $y_e = 0$ only and $R = V \setminus L$. Because of the constraints, (L, R) is a cut. Every edge of $L \times R$ has $y_e = 1$, and so the capacity of the cut is $\sum_{e \in L \times R} c_e = \sum_{e \in L \times R} c_e y_e \leq \sum_e c_e y_e$, so $\text{Value}(IP) \geq \text{Value}(\text{MinCut})$.

Conversely, given a cut (L, R) , let (y_e) be defined by $y_e = \begin{cases} 1 & e \in L \times R \\ 0 & \text{otherwise} \end{cases}$. Clearly, (y_e) is feasible for IP and its value equals the capacity of the cut, so $\text{Value}(\text{MinCut}) \geq \text{Value}(IP)$. The claim follows.

Concatenating everything:
 $\text{Value}(\text{MaxFlow}) = \text{Value}(LP_0) = \text{Value}(LP_1) = \text{Value}(LP_2) = \text{Value}(IP) = \text{Value}(\text{MinCut})$. ■

Theorem 63 (Von-Neumann Min-Max theorem) For every $A \in \mathbb{R}^{m \times n}$,

$$\max_{\mathbf{x} \in \Delta_m} \min_{\mathbf{y} \in \Delta_n} \mathbf{y}^\top \mathbf{A} \mathbf{x} = \min_{\mathbf{y} \in \Delta_n} \max_{\mathbf{x} \in \Delta_m} \mathbf{y}^\top \mathbf{A} \mathbf{x}$$

$$\Delta_k \stackrel{\Delta}{=} \{x \in \mathfrak{R}^k : \|x\|_1 = 1, \forall i x_i \geq 0\}$$

Theorem 64 (Yao's principle) Using the Min-Max theorem (or the weak duality theorem), prove Yao's principle:

The expected cost of any randomized algorithm for solving a given problem, on the worst case input for that algorithm, can be no better than the expected cost, for a worst-case random probability distribution on the inputs, of the deterministic algorithm that performs best against that distribution. Thus, to establish a lower bound on the performance of randomized algorithms, it suffices to find an appropriate distribution of difficult inputs, and to prove that no deterministic algorithm can perform well against that distribution.

Hints:

- A random algorithm can be refereed as a probability over the deterministic algorithms (Toss all random coins in advance).
- Consider a zero sum game in which A chooses a deterministic algorithm while an adversary (B) chooses an input. The payoff of A to B will be the run-time of the algorithm on the input.

31.1 Proving the Strong duality theorem

To prove the strong duality theorem we reduce the primal program in Eqn. 13 to the feasibility test in Farkas lemma: Whether there is nonnegative solution vector $\mathbf{x} \geq 0$ satisfying the linear equations $\mathbf{A} \mathbf{x} = \mathbf{b}$ and $\langle \mathbf{c}, \mathbf{x} \rangle = \lambda$. If λ is larger than the maximum of the linear program in Eqn. 13 these linear equations cannot be satisfied by a nonnegative assignment. One the other hand if λ is smaller or equal to the maximum of this linear program then there is nonnegative \mathbf{x} satisfying these constraints. Let $\bar{\mathbf{A}}$ be a matrix with $m + 1$ rows and n columns, whose rows correspond to the m row of A and to the vector \mathbf{c} , and the columns correspond to the nonnegative variables \mathbf{x} . Let $\bar{\mathbf{b}} \in \mathbb{R}^{m+1}$ be the vector (\mathbf{b}, λ) . With the introduced notation the feasibility test is stated as: Is there nonnegative solution $\mathbf{x} \geq 0$ satisfying $\bar{\mathbf{A}} \mathbf{x} = \bar{\mathbf{b}}$? If the answer is no then Farkas lemma tells us there exists a vector $\bar{\mathbf{y}} \in \mathbb{R}^{m+1}$ such that $\bar{\mathbf{y}}^\top \bar{\mathbf{A}} \geq 0$ and $\langle \bar{\mathbf{y}}, \bar{\mathbf{b}} \rangle < 0$.

- The last coordinate of $\bar{\mathbf{y}}$ is non-zero, since otherwise the primal linear program in Eqn. 13 is infeasible. Assume without loss of generality that the last coordinate of $\bar{\mathbf{y}}$ equals -1 , otherwise divide $\bar{\mathbf{y}}$ by $-\bar{y}_{m+1}$ and denote $\mathbf{y} = (\bar{y}_1, \dots, \bar{y}_m)$. In this case holds $\langle \bar{\mathbf{y}}, \bar{\mathbf{b}} \rangle = \langle \mathbf{y}, \mathbf{b} \rangle - \lambda$ and the inequality $\langle \bar{\mathbf{y}}, \bar{\mathbf{b}} \rangle < 0$ can be written as $\langle \mathbf{y}, \mathbf{b} \rangle < \lambda$.
- The j -th coordinate of the inequality $\bar{\mathbf{y}}^\top \bar{\mathbf{A}} \geq 0$ can be equivalently written as $(\mathbf{y}^\top \mathbf{A})_j - c_j \geq 0$ and in its matrix form $\mathbf{y}^\top \mathbf{A} \geq \mathbf{c}$.

λ is greater than the maximum of the primal linear program in Eqn. 13 if and only if there is a vector $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{y}^\top \mathbf{A} \geq \mathbf{c}$ and $\langle \mathbf{y}, \mathbf{b} \rangle < \lambda$. If we relax the constraint $\langle \mathbf{y}, \mathbf{b} \rangle < \lambda$ to be $\langle \mathbf{y}, \mathbf{b} \rangle \leq \lambda$

then there will be a single critical value of λ for which both systems have solutions. This critical λ is the maximum of the original linear program in Eqn. 13 as well as the minimum of the dual linear program in Eqn. 14.