Chapter 1

Intro, Quick Sort and BSP

By Sariel Har-Peled, December 30, 2015[①]

> Finally: It was stated at the outset, that this system would not be here, and at once, perfected. You cannot but plainly see that I have kept my word. But I now leave my cetological System standing thus unfinished, even as the great Cathedral of Cologne was left, with the crane still standing upon the top of the uncompleted tower. For small erections may be finished by their first architects; grand ones, true ones, ever leave the copestone to posterity. God keep me from ever completing anything. This whole book is but a draft - nay, but the draft of a draft. Oh, Time, Strength, Cash, and Patience!
>
> – Herman Melville, Moby Dick.

1.1. General Introduction

1.1.1. Administrivia

- prerequisites: algorithms course, ability to do proofs
- homework weekly (first next week)
- books.

1.1.2. What are randomized algorithms?

Randomized algorithms are algorithms that makes random decision during their execution. Specifically, they are allowed to use variables that their value is taken from some random distribution. It is not immediately clear why adding the ability to consult with randomness would help an algorithm. But it turns out that the benefits are quite substantial:

Best. There are cases were only randomized algorithm is known or possible, especially for games. For example, consider the 3 coins example.

Speed. In some cases randomized algorithms are considerably faster than any deterministic algorithm.

Simplicity. Even if a randomized algorithm is not faster, often it is considerably simpler than its deterministic counterpart.

Derandomization. Some deterministic algorithms arises from derandomizing the randomized algorithms, and this the only algorithm we know for these problems (i.e., discrepancy).

Adversary arguments and lower bounds. The standard worst case analysis relies on the idea that the adversary can select the input on which the algorithm performs worst. Inherently, the adversary is more powerful than the algorithm, since the algorithm is completely predictable. By using a randomized algorithm, we can make the algorithm unpredictable and break the adversary lower bound.

Namely, randomness makes the algorithm vs. adversary game a more balanced game, by giving the algorithm additional power against the adversary.

1.1.3. The benefits of unpredictability

Consider the following game. The adversary has a equilateral triangle, with three coins on the vertices of the triangle (which are, numbered by, I don't known, 1,2,3). At every step of the game, the play can ask the player can ask the adversary to flip certain coins (say, flip coins at vertex 1 and 3). If after the flips all three coins have the same side up, then the game stop. Otherwise, the adversary is allowed to rotate the board by $0, 120$ or -120 degrees, as she seems fit. And the game continues from this point on.

1.1.3.0.1. Randomized algorithm. The randomized algorithm in this case is easy – the play randomly chooses a number among $1, 2, 3$ at every stage. Since, at every point in time, there are two coins that have the same side up, and the other coin is the other side up, a random choice hits the lonely coin, and thus finishes the game, with probability $1/3$ at each step. In particular, the number of iterations of the game till it terminates is a geometric variable with geometric distribution with expectation 3. Clearly, the probability that the game continues for more than i rounds, when the player uses this random algorithm, is $(2/3)^i$. In particular, it vanishes to zero relatively quickly.

1.1.3.0.2. Deterministic algorithm. The surprise here is that there is no deterministic algorithm that can generate a winning sequence. Indeed, if the player uses a deterministic algorithm, then the adversary can simulate the algorithm herself, and know at every stage what coin the player would ask to flip (it is easy to verify that flipping two coins in a step is equivalent to flipping toe other coin – so we can restrict ourselves to a single coin flip at each step). In particular, the adversary can rotate the board such that every stage, the player flips one of the two coins that are in the same state. Namely, the player never wins.

1.1.3.0.3. The shocker. One can play the same game with a board of size 4 (i.e., a square), where at each stage the player can flip one or two coins, and the adversary can rotate the board by $0, 90, 180, 270$ degrees after each round. Surprisingly, there is a deterministic winning strategy for this case. The interested reader can think what it is (this is one of these brain teasers that are not immediate, and might take you 15 minutes to solve, or longer).

1.1.4. Randomized vs average-case analysis

Randomized algorithms are not the same as *average-case analysis*. In average case analysis, one assumes that is given some distribution on the input, and one tries to analyze an algorithm execution on such an input.

On the other hand, randomized algorithms do not assume random inputs – inputs can be arbitrary. As such, randomized algorithm analysis is more widely applicable, and more general.

while there is a lot of average case analysis in the literature, the problem that it is hard to find distribution on inputs that are meaningful in comparison to real world inputs. In particular, for numerous cases, the average case analysis exposes structure that does not exist in real world input.

1.2. Basic probability

Here we recall some definitions about probability. The reader already familiar with these definition can happily skip this section.

1.2.1. Formal basic definitions: Sample space, σ-algebra, and probability

Here we formally define some basic notions in probability. The reader familiar with these concepts can safely skip this part.

A *sample space* Ω is a set of all possible outcomes of an experiment. We also have a set of events $\mathcal{F}$, where every member of $\mathcal{F}$ is a subset of Ω. Formally, we will require that $\mathcal{F}$ is a σ-algebra.

Definition 1.2.1. A single element of Ω is an *elementary event* or an *atomic event*.

Definition 1.2.2. A set $\mathcal{F}$ of subsets of Ω is a *σ-algebra* if:

(i) $\mathcal{F}$ is not empty,
(ii) if $X \in \mathcal{F}$ then $\overline{X} = (\Omega \setminus X) \in \mathcal{F}$, and
(iii) if $X, Y \in \mathcal{F}$ then $X \cup Y \in \mathcal{F}$.

More generally, we will require that if $X_i \in \mathcal{F}$, for $i \in \mathbb{Z}$, then $\cup_i X_i \in \mathcal{F}$. A member of $\mathcal{F}$ an *event*.

As a concrete example, if we are rolling a dice, then $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{F}$ would be the power set of all possible subsets of Ω.

Definition 1.2.3. A *probability measure* is a mapping $\mathbf{Pr} : \mathcal{F} \to [0, 1]$ assigning *probabilities* to events. The function $\mathbf{Pr}$ needs to have the following properties:

(i) Additive: for $X, Y \in \mathcal{F}$ disjoint sets, we have that $\mathbf{Pr}\Big[X \cup Y\Big] = \mathbf{Pr}\Big[X\Big] + \mathbf{Pr}\Big[Y\Big]$, and
(ii) $\mathbf{Pr}[\Omega] = 1$.

Definition 1.2.4. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbf{Pr})$, where Ω is a sample space, $\mathcal{F}$ is a σ-algebra defined over Ω, and $\mathbf{Pr}$ is a probability measure.

Definition 1.2.5. A *random variable* f is a mapping from Ω into some set $\mathcal{G}$. We will require that the probability of the random variable to take on any value in a given subset of values is well defined. Formally, we will require that for any subset $U \subseteq \mathcal{G}$, we have that $f^{-1}(U) \in \mathcal{F}$. That is, $\mathbf{Pr}[f \in U] = \mathbf{Pr}\Big[f^{-1}(U)\Big]$ is defined.

Going back to the dice example, the number on the top of the dice when we roll it is a random variable. Similarly, let X be one if the number rolled is larger than 3, and zero otherwise. Clearly X is a random variable.

We denote the *probability* of a random variable X to get the value x, by $\mathbf{Pr}[X = x]$ (or sometime $\mathbf{Pr}[x]$, if we are lazy).

1.2.2. Expectation and conditional probability

Definition 1.2.6 (Expectation). The expectation of a random variable X, is its average. Formally, the *expectation* of X is

$$\mathbf{E}\Big[X\Big] = \sum_x x \, \mathbf{Pr}\Big[X = x\Big].$$

Definition 1.2.7 (Conditional Probability.). The *conditional probability* of X given Y, is the probability that $X = x$ given that $Y = y$. We denote this quantity by $\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big]$.

Conditional probability specially is mental chaos. One useful way to think about conditional probability $\mathbf{Pr}\Big[X \;\Big|\; Y\Big]$ is as a function, between the given value of Y (i.e., y), and the probability of X (to be equal to x) in this case. Since in many cases x and y are omitted in the notation, it is somewhat confusing.

The conditional probability can be computed using the formula

$$\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big] = \frac{\mathbf{Pr}[\ (X = x) \ \cap \ (Y = y)\ ]}{\mathbf{Pr}[Y = y]}.$$

For example, let us roll a dice and let X be the number we got. Let Y be the random variable that is true if the number we get is even. Then, we have that

$$\mathbf{Pr}\Big[X = 2 \;\Big|\; Y = \mathit{true}\Big] = \frac{1}{3}.$$

Definition 1.2.8. Two random variables X and Y are *independent* if $\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big] = \mathbf{Pr}[X = x]$, for all x and y.

Observation 1.2.9. *If X and Y are independent then $\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big] = \mathbf{Pr}[X = x]$ which is equivalent to $\dfrac{\mathbf{Pr}[X = x \cap Y = y]}{\mathbf{Pr}[Y = y]} = \mathbf{Pr}[X = x]$. That is, X and Y are* independent*, if for all x and y, we have that $\mathbf{Pr}\Big[X = x \cap Y = y\Big] = \mathbf{Pr}\Big[X = x\Big]\mathbf{Pr}\Big[Y = y\Big]$.*

Lemma 1.2.10 (Linearity of expectation). Linearity of expectation *is the property that for any two random variables X and Y, we have that $\mathbf{E}\Big[X + Y\Big] = \mathbf{E}\Big[X\Big] + \mathbf{E}\Big[Y\Big]$.*

Proof: $\mathbf{E}\Big[X + Y\Big] = \sum_{\omega\in\Omega} \mathbf{Pr}[\omega]\Big(X(\omega) + Y(\omega)\Big) = \sum_{\omega\in\Omega} \mathbf{Pr}[\omega]\, X(\omega) + \sum_{\omega\in\Omega} \mathbf{Pr}[\omega]\, Y(\omega) = \mathbf{E}\Big[X\Big] + \mathbf{E}\Big[Y\Big]$. ■

1.3. QuickSort

Let the input be a set $t_1, \ldots, t_n$ of n items to be sorted. We remind the reader, that the **QuickSort** algorithm randomly pick a pivot element (uniformly), splits the input into two subarrays of all the elements smaller than the pivot, and all the elements larger than the pivot, and then it recurses on these two subarrays (the pivot is not included in these two subproblems). Here we will show that the expected running time of **QuickSort** is $O(n \log n)$.

Definition 1.3.1. For an event $\mathcal{E}$, let X be a random variable which is 1 if $\mathcal{E}$ occurred and 0 otherwise. The random variable X is an *indicator variable*.

Observation 1.3.2. *For an indicator variable X of an event $\mathcal{E}$, we have*

$$\mathbf{E}\Big[X\Big] = 0 \cdot \mathbf{Pr}\Big[X = 0\Big] + 1 \cdot \mathbf{Pr}\Big[X = 1\Big] = \mathbf{Pr}\Big[X = 1\Big] = \mathbf{Pr}\Big[\mathcal{E}\Big].$$

Let $S_1, \ldots, S_n$ be the elements in their sorted order (i.e., the output order). Let $X_{ij} = 1$ be the indicator variable which is one iff **QuickSort** compares S_i to S_j, and let p_{ij} denote the probability that this happens. Clearly, the number of comparisons performed by the algorithm is $C = \sum_{i<j} X_{ij}$. By linearity of expectations, we have

$$\mathbf{E}\Big[C\Big] = \mathbf{E}\left[\sum_{i<j} X_{ij}\right] = \sum_{i<j} \mathbf{E}\Big[X_{ij}\Big] = \sum_{i<j} p_{ij}.$$

We want to bound p_{ij}, the probability that the S_i is compared to S_j. Consider the last recursive call involving both S_i and S_j. Clearly, the pivot at this step must be one of $S_i, \ldots, S_j$, all equally likely. Indeed, S_i and S_j were separated in the next recursive call.

Observe, that S_i and S_j get compared if and only if pivot is S_i or S_j. Thus, the probability for that is $2/(j-i+1)$. Indeed,

$$p_{ij} = \mathbf{Pr}\Big[S_i \text{ or } S_j \text{ picked} \;\Big|\; \text{picked pivot from } S_i, \ldots, S_j\Big] = \frac{2}{j-i+1}.$$

Thus,

$$\sum_{i=1}^{n}\sum_{j>i} p_{ij} = \sum_{i=1}^{n}\sum_{j>i} 2/(j-i+1) = \sum_{i=1}^{n}\sum_{k=1}^{n-i+1}\frac{2}{k} \leq 2\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{1}{k} \leq 2nH_n \leq n + 2n\ln n,$$

where H_n is the *harmonic number*[2] $H_n = \sum_{i=1}^{n}\frac{1}{i}$. We thus proved the following result.

Lemma 1.3.3. **QuickSort** *performs in expectation at most $n + 2n\ln n$ comparisons, when sorting n elements.*

Note, that this holds for all inputs. No assumption on the input is made. Similar bounds holds not only in expectation, but also with high probability.

This raises the question, of how does the algorithm pick a random element? We assume we have access to a random source that can get us number between 1 and n uniformly.

Note, that the algorithm always works, but it might take quadratic time in the worst case.

1.4. Binary space partition (BSP)

Let assume that we would like to render an image of a three dimensional scene on the computer screen. The input is in general a collection of polygons in three dimensions. The *painter* algorithm, render the scene by drawing things from back to front; and let front stuff overwrite what was painted before.

The problem is that it is not always possible to order the objects in three dimensions. This ordering might have cycles. So, one possible solution is to build a *binary space partition*. We build a binary tree. In the root, we place a polygon P. Let h be the plane containing P. Next, we partition the input polygons into two sets, depending on which side of h they fall into. We recursively construct a BSP for each set, and we hang it from the root node. If a polygon intersects h then we cut it into two polygons as split by h. We continue the construction recursively on the objects on one side of h, and the objects on the other side. What we get, is a binary tree that splits space into cells, and furthermore, one can use the painter algorithm on these objects. The natural question is how big is the resulting partition.

We will study the easiest case, of disjoint segments in the plane.

1.4.1. BSP for disjoint segments

Let $P = \{\mathsf{s}_1, \ldots, \mathsf{s}_n\}$ be n disjoint segments in the plane. We will build the BSP by using the lines defined by these segments. This kind of BSP is called *autopartition*.

To recap, the BSP is a binary tree, at every internal node we store a segment of P, where the line associated with it splits its region into its two children. Finally, each leaf of the BSP stores a single segment. A *fragment*

[2] Using integration to bound summation, we have $H_n \leq 1 + \int_{x=1}^{n}\frac{1}{x}dx \leq 1 + \ln n$. Similarly, $H_n \geq \int_{x=1}^{n}\frac{1}{x}dx = \ln n$.

is just going to be a subsegment formed by this splitting. Clearly, every internal node, stores a fragment that defines its split. As such, the size of the BSP is proportional to the number of fragments generated when building the BSP.

One application of such a BSP is ray shooting - given a ray you would like to determine what is the first segment it hits. Start from the root, figure out which child contains the apex of the ray, and first (recursively) compute the first segment stored in this child that the ray intersect. Contain into the second child only if the first subtree does not contain any segment that intersect the ray.

1.4.1.1. The algorithm

We pick a random permutation σ of $1, \ldots, n$, and in the ith step we insert $\mathsf{s}_{\sigma(i)}$ splitting all the cells that s_i intersects.

Observe, that if s_i crosses a cell completely, it just splits it into two and no new fragments are created. As such, the bad case is when a segment s is being inserted, and its line intersect some other segment t.

So, let $\mathcal{E}(\mathsf{s},\mathsf{t})$ denote the event that when inserted s it had split t. In particular, let $\mathrm{index}(\mathsf{s},\mathsf{t})$ denote the number of segments on the line of s between s (closer) endpoint and t (including t. If the line of s does not intersect t, then $\mathrm{index}(\mathsf{s},\mathsf{t}) = \infty$.

We have that

$$\mathbf{Pr}\Big[\mathcal{E}(\mathsf{s},\mathsf{t})\Big] = \frac{1}{1+\mathrm{index}(\mathsf{s},\mathsf{t})}.$$

Let $X_{\mathsf{s},\mathsf{t}}$ be the indicator variable that is 1 if $\mathcal{E}(\mathsf{s},\mathsf{t})$ happens. We have that

$$S = \text{ number of fragments } = \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} X_{\mathsf{s}_i,\mathsf{s}_j}.$$

As such, by linearity of expectations, we have

$$\begin{aligned}
\mathbf{E}[S] &= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} X_{\mathsf{s}_i,\mathsf{s}_j}\right] = \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathbf{E}\Big[X_{\mathsf{s}_i,\mathsf{s}_j}\Big] = \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathbf{Pr}\Big[\mathcal{E}(\mathsf{s}_i,\mathsf{s}_j)\Big] \\
&= \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\frac{1}{1+\mathrm{index}\Big(\mathsf{s}_i,\mathsf{s}_j\Big)} \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{2}{1+j} = 2nH_n.
\end{aligned}$$

Since the size of the BSP is proportional to the number of fragments created, we have the following result.

Theorem 1.4.1. *Given n disjoint segments in the plane, one can build a BSP for them of size $O(n\log n)$.*

Csaba Tóth [Tót03] showed that BSP for segments in the plane, in the worst case, has complexity $\Omega\left(n\dfrac{\log n}{\log\log n}\right)$.

1.5. Extra: QuickSelect running time

We remind the reader that **QuickSelect** receives an array $t[1\ldots n]$ of n real numbers, and a number k, and returns the element of rank k in the sorted order of the elements of t. We can of course, use **QuickSort**, and

just return the kth element in the sorted array, but a more efficient algorithm, would be to modify **QuickSelect**, so that it recurses on the subproblem that contains the element we are interested in. Formally, **QuickSelect** chooses a random pivot, splits the array according to the pivot. This implies that we now know the rank of the pivot, and if its equal to $\overline{\mathsf{m}}$, we return it. Otherwise, we recurse on the subproblem containing the required element (modifying $\overline{\mathsf{m}}$ as we go down the recursion. Namely, **QuickSelect** is a modification of **QuickSort** performing only a single recursive call (instead of two).

As before, to bound the expected running time, we will bound the expected number of comparisons. As before, let $S_1, \ldots, S_n$ be the elements of t in their sorted order. Now, for $i < j$, let X_{ij} be the indicator variable that is one if S_i is being compared to S_j during the execution of **QuickSelect**. There are several possibilities to consider:

(i) If $i < j < \overline{\mathsf{m}}$: Here, S_i is being compared to S_j, if and only if the first pivot in the range $S_i, \ldots, S_k$ is either S_i or S_j. The probability for that is $2/(k-i+1)$. As such, we have that

$$\alpha_1 = \mathbf{E}\left[\sum_{i<j<\overline{\mathsf{m}}} X_{ij}\right] = \mathbf{E}\left[\sum_{i=1}^{\overline{\mathsf{m}}-2}\sum_{j=i+1}^{\overline{\mathsf{m}}-1} X_{ij}\right] = \sum_{i=1}^{\overline{\mathsf{m}}-2}\sum_{j=i+1}^{\overline{\mathsf{m}}-1} \frac{2}{\overline{\mathsf{m}}-i+1} = \sum_{i=1}^{med-2} \frac{2(\overline{\mathsf{m}}-i-1)}{\overline{\mathsf{m}}-i+1} \leq 2\Big(\overline{\mathsf{m}}-2\Big).$$

(ii) If $\overline{\mathsf{m}} < i < j$: Using the same analysis as above, we have that $\mathbf{Pr}\Big[X_{ij}=1\Big] = 2/(j-\overline{\mathsf{m}}+1)$. As such,

$$\alpha_2 = \mathbf{E}\left[\sum_{j=\overline{\mathsf{m}}+1}^{n}\sum_{i=\overline{\mathsf{m}}+1}^{j-1} X_{ij}\right] = \sum_{j=\overline{\mathsf{m}}+1}^{n}\sum_{i=\overline{\mathsf{m}}+1}^{j-1} \frac{2}{j-\overline{\mathsf{m}}+1} = \sum_{j=\overline{\mathsf{m}}+1}^{n} \frac{2(j-\overline{\mathsf{m}}-1)}{j-\overline{\mathsf{m}}+1} \leq 2\Big(n-\overline{\mathsf{m}}\Big).$$

(iii) $i < \overline{\mathsf{m}} < j$: Here, we compare S_i to S_j if and only if the first indicator in the range $S_i, \ldots, S_j$ is either S_i or S_j. As such, $\mathbf{E}\Big[X_{ij}\Big] = \mathbf{Pr}\Big[X_{ij}=1\Big] = 2/(j-i+1)$. As such, we have

$$\alpha_3 = \mathbf{E}\left[\sum_{i=1}^{\overline{\mathsf{m}}-1}\sum_{j=\overline{\mathsf{m}}+1}^{n} X_{ij}\right] = \sum_{i=1}^{\overline{\mathsf{m}}-1}\sum_{j=\overline{\mathsf{m}}+1}^{n} \frac{2}{j-i+1}.$$

Observe, that for a fixed $\Delta = j-i+1$, we are going to handle the gap Δ in the above summation, at most $\Delta-2$ times. As such, $\alpha_3 \leq \sum_{\Delta=3}^{n} 2(\Delta-2)/\Delta \leq 2n$.

(iv) $i = \overline{\mathsf{m}}$. We have $\alpha_4 = \displaystyle\sum_{j=\overline{\mathsf{m}}+1}^{n} \mathbf{E}\Big[X_{ij}\Big] = \sum_{j=\overline{\mathsf{m}}+1}^{n} \frac{2}{j-\overline{\mathsf{m}}+1} = \ln n + 1.$

(v) $j = \overline{\mathsf{m}}$. We have $\alpha_5 = \displaystyle\sum_{i=1}^{\overline{\mathsf{m}}-1} \mathbf{E}\Big[X_{ij}\Big] = \sum_{i=1}^{\overline{\mathsf{m}}-1} \frac{2}{\overline{\mathsf{m}}-i+1} \leq \ln \overline{\mathsf{m}} + 1.$

Thus, the expected number of comparisons performed by **QuickSelect** is bounded by

$$\sum_i \alpha_i \leq 2(\overline{\mathsf{m}}-2) + 2(n-\overline{\mathsf{m}}) + 2n + \ln n + 1 + \ln \overline{\mathsf{m}} = 4n - 2 + \ln n + \ln \overline{\mathsf{m}}.$$

Theorem 1.5.1. *In expectation,* **QuickSelect** *performs at most* $4n - 2 + \ln n + \ln \overline{\mathsf{m}}$ *comparisons, when selecting the* $\overline{\mathsf{m}}$*th element out of* n *elements.*

A different approach can reduce the number of comparisons (in expectation) to $1.5n + o(n)$. More on that later in the course.

Bibliography

[Tót03] C. D. Tóth. A note on binary plane partitions. *Discrete Comput. Geom.*, 30(1):3–16, 2003.

Chapter 2

Verifying Identities, Changing Minimum, Closest Pair and Some Complexity

By Sariel Har-Peled, December 30, 2015[①]

> The events of 8 September prompted Foch to draft the later legendary signal: "My centre is giving way, my right is in retreat, situation excellent. I attack." It was probably never sent.
> – John Keegan, The first world war.

2.1. Verifying equality

2.1.1. Vectors

You are given two binary vectors $\mathbf{v} = (v_1, \ldots, v_n), \mathbf{u} = (u_1, \ldots, u_n) \in \left\{0,1\right\}^n$ and you would like to decide if they are equal or not. Unfortunately, the only access you have to the two vectors is via a black-box that enables you to compute the dot-product of two binary vectors over $\mathbb{Z}_2$. Formally, given two binary vectors as above, their dot-product is $\langle \mathbf{v}, \mathbf{u} \rangle = \sum_{i=1}^{n} v_i u_i$ (which is a non-negative integer number). Their dot product modulo 2, is $\langle \mathbf{v}, \mathbf{u} \rangle \mod 2$ (i.e., it is 1 if $\langle \mathbf{v}, \mathbf{u} \rangle$ is odd and 0 otherwise).

Naturally, we could the use the black-box to read the vectors (using $2n$ calls), but since we interested only in deciding if they are equal or not, this should require less calls to the black-box (which is expensive).

Lemma 2.1.1. *Given two binary vectors $\mathbf{v}, \mathbf{u} \in \left\{0,1\right\}^n$, a randomized algorithm can, using two computations of dot-product modulo 2, decide if $\mathbf{v}$ is equal to vecB or not. The algorithm may return the following.*

≠: *Then $\mathbf{v} \neq \mathbf{u}$.*

=: *Then the probability that the algorithm made a mistake (i.e., the vectors are different) is at most $1/2$.*

The running time of the algorithm is $O(n + B(n))$, where $B(n)$ is the time to compute a single dot-product of vectors of length n.

Proof: Pick a random vector $\mathbf{r} = (r_1, \ldots, r_n) \in \left\{0,1\right\}^n$ by picking each coordinate independently with probability $1/2$. Compute $\langle \mathbf{v}, \mathbf{r} \rangle$ and $\langle \mathbf{u}, \mathbf{r} \rangle$.

- If $\langle \mathbf{v}, \mathbf{r} \rangle \equiv \langle \mathbf{v}, \mathbf{r} \rangle \bmod 2 \quad \Rightarrow \quad$ the algorithm returns '$=$'.
- If $\langle \mathbf{v}, \mathbf{r} \rangle \not\equiv \langle \mathbf{v}, \mathbf{r} \rangle \bmod 2 \quad \Rightarrow \quad$ the algorithm returns '$\neq$'.

Clearly, if the '$\neq$' then it is correct.

So, assume that the algorithm returns '$=$' but $\mathbf{v} \neq \mathbf{u}$. For the sake of simplicity of exposition, assume that $u_n \neq v_n$. We then have that

$$\alpha = \langle \mathbf{v}, \mathbf{r} \rangle = \overbrace{\sum_{i=1}^{n-1} v_i r_i}^{=\alpha'} + v_n r_n \qquad \text{and} \qquad \beta = \langle \mathbf{u}, \mathbf{r} \rangle = \overbrace{\sum_{i=1}^{n-1} u_i r_i}^{=\beta'} + u_n r_n.$$

Now, there are two possibilities:

- If $\alpha' \not\equiv \beta' \bmod 2$, then, with probability half, we have $r_i = 0$, and as such $\alpha \not\equiv \beta \bmod 2$.
- If $\alpha' \equiv \beta' \bmod 2$, then, with probability half, we have $r_i = 1$, and as such $\alpha \not\equiv \beta \bmod 2$.

As such, with probability at most half, the algorithm would fail to discover that the two vectors are different. ■

2.1.1.1. Amplification

Of course, this is not a satisfying algorithm – it returns the correct answer only with probability half if the vectors are different. So, let us run the algorithm t times. Let $T_1, \ldots, T_t$ be the returned values from all these executions. If any of the t executions returns that the vectors are different, then we know that they are different.

$$\begin{aligned}
\mathbf{Pr}\Big[\text{Algorithm fails}\Big] &= \mathbf{Pr}\Big[\mathbf{v} \neq \mathbf{u}, \text{ but all } t \text{ executions return '}=\text{'}\Big] \\
&= \mathbf{Pr}\Big[\Big(T_1 = \text{'}=\text{'}\Big) \cap \Big(T_2 = \text{'}=\text{'}\Big) \cap \cdots \cap \Big(T_t = \text{'}=\text{'}\Big)\Big] \\
&= \mathbf{Pr}\Big[T_1 = \text{'}=\text{'}\Big] \mathbf{Pr}\Big[T_2 = \text{'}=\text{'}\Big] \cdots \mathbf{Pr}\Big[T_t = \text{'}=\text{'}\Big] \leq \prod_{i=1}^{t} \frac{1}{2} = \frac{1}{2^t}.
\end{aligned}$$

We thus get the following result.

Lemma 2.1.2. *Given two binary vectors $\mathbf{v}, \mathbf{u} \in \{0,1\}^n$ and a* confidence *parameter $\delta > 0$, a randomized algorithm can decide if $\mathbf{v}$ is equal to vecB or not. More precisely:*

$\neq$: *If the returned result is that the two vectors are not equal then indeed they are.*

$=$: *If the returned is that the two vectors are equal then the probability it made a mistake is at most δ.*

The running time of the algorithm is $O\Big((n + B(n)) \ln(1/\delta)\Big)$, where $B(n)$ is the time to compute a single dot-product of two vectors of length n.

Proof: Follows from the above by setting $t = \lceil \lg(1/\delta) \rceil$. ■

2.1.2. Matrices

Given three binary matrices B, C, D of size $n \times n$, we are interested in the question of deciding if $\mathsf{BC} = \mathsf{D}$. Computing BC is expensive – the fastest known (theoretical!) algorithm has running time (roughly) $O\Big(n^{2.37}\Big)$. On the other hand, multiplying such a matrix with a vector $\mathbf{r}$ (modulo 2, as usual) takes only $O(n^2)$ and is a much simpler algorithm.

Lemma 2.1.3. *Given three binary matrices $\mathsf{B}, \mathsf{C}, \mathsf{D} \in \{0,1\}^{n \times n}$ and aconfidence parameter $\delta > 0$, a randomized algorithm can decide if $\mathsf{BC} = \mathsf{D}$ or not. More precisely the algorithm can return one of the following two results:*

$\neq$: *Then* $\mathsf{BC} \neq \mathsf{D}$.

$=$: *Then the probability that* $\mathsf{BC} \neq \mathsf{D}$ *is at most* δ.

The running time of the algorithm is $O\left(n^2 \ln \frac{1}{\delta}\right)$.

Proof: Compute a random vector $\mathbf{r} = \{r_1, \ldots, r_n\}$, and compute the quantity $\mathbf{x} = \mathsf{BC}\mathbf{r} = \mathsf{B}(\mathsf{C}\mathbf{r})$ in $O(n^2)$ time, using the associative property of matrix multiplication. Similarly, compute $\mathbf{y} = \mathsf{D}\mathbf{r}$. Now, if $\mathbf{x} \neq \mathbf{y}$ then return '$=$'.

Now, we execute this algorithm $t = \lceil \lg 1/\delta \rceil$ times. If all of these independent runs return that the matrices are equal then return '$=$'.

The algorithm fails only if $\mathsf{BC} \neq \mathsf{D}$, but then, assume the ith row in two matrices BC and D are different. But then, the probability that the algorithm would not detect that these rows are different is at most $1/2$, by Lemma 2.1.1. As such, the probability that all t runs failed is at most $1/2^t \leq \delta$, as desired. ■

2.2. How many times can a minimum change?

Let $a_1, \ldots, a_n$ be a set of n numbers, and let us randomly permute them into the sequence $b_1, \ldots, b_n$. Next, let $c_i = \min_{k=1}^{i} b_i$, and let X be the random variable which is the number of distinct values that appears in the sequence $c_1, \ldots, c_n$. What is the expectation of X?

Lemma 2.2.1. *In expectation, the number of times the minimum of a prefix of n randomly permuted numbers change, is* $O(\log n)$. *That is* $\mathbf{E}[X] = O(\log n)$.

Proof: Consider the indicator variable X_i, such that $X_i = 1$ if $c_i \neq c_{i-1}$. The probability for that is $\leq q1/i$, since this is the probability that the smallest number if $b_1, \ldots, b_i$ is b_i. As such, we have $X = \sum_i X_i$, and

$$\mathbf{E}[X] = \sum_i \mathbf{E}[X_i] = \sum_{i=1}^{n} \frac{1}{i} = O(\log n).$$

■

2.3. Closest Pair

Assumption 2.3.1. Throughout the discourse, we are going to assume that every hashing operation takes (worst case) constant time. This is quite a reasonable assumption when true randomness is available (using for example perfect hashing [CLRS01]). We probably will revisit this issue later in the course.

For a real positive number r and a point $\mathsf{p} = (x, y)$ in $\mathbb{R}^2$, define

$$\mathrm{G}_r(\mathsf{p}) := \left(\left\lfloor \frac{x}{r} \right\rfloor r \;\;,\;\; \left\lfloor \frac{y}{r} \right\rfloor r \right) \;\; \in \mathbb{R}^2.$$

We call r the *width* of the *grid* G_r. Observe that G_r partitions the plane into square regions, which we call grid *cells*. Formally, for any $i, j \in \mathbb{Z}$, the intersection of the half-planes $x \geq ri$, $x < r(i+1)$, $y \geq rj$ and $y < r(j+1)$ is said to be a *grid cell*. Further we define a *grid cluster* as a block of 3×3 contiguous grid cells.

For a point set P, and a parameter r, the partition of P into subsets by the grid G_r, is denoted by $\mathrm{G}_r(\mathsf{P})$. More formally, two points $\mathsf{p}, \mathsf{q} \in \mathsf{P}$ belong to the same set in the partition $\mathrm{G}_r(\mathsf{P})$, if both points are being mapped to the same grid point or equivalently belong to the same grid cell.

Note, that every grid cell C of G_r, has a unique ID; indeed, let $\mathsf{p} = (x, y)$ be any point in C, and consider the pair of integer numbers $\mathrm{id}_C = \mathrm{id}(\mathsf{p}) = (\lfloor x/r \rfloor, \lfloor y/r \rfloor)$. Clearly, only points inside C are going to be mapped to id_C. This is very useful, since we can store a set P of points inside a grid efficiently. Indeed, given a point

p, compute its id(p). We associate with each unique id a data-structure that stores all the points falling into this grid cell (of course, we do not maintain such data-structures for grid cells which are empty). So, once we computed id(p), we fetch the data structure for this cell, by using hashing. Namely, we store pointers to all those data-structures in a hash table, where each such data-structure is indexed by its unique id. Since the ids are integer numbers, we can do the hashing in constant time.

We are interested in solving the following problem.

Problem 2.3.2. Given a set P of n points in the plane, find the pair of points closest to each other. Formally, return the pair of points realizing $\mathcal{CP}(\mathsf{P}) = \min_{\mathsf{p},\mathsf{q}\in P} \|\mathsf{pq}\|$.

Lemma 2.3.3. *Given a set P of n points in the plane, and a distance r, one can verify in linear time, whether or not $\mathcal{CP}(\mathsf{P}) < r$ or $\mathcal{CP}(\mathsf{P}) \geq r$.*

Proof: Indeed, store the points of P in the grid G_r. For every non-empty grid cell, we maintain a linked list of the points inside it. Thus, adding a new point p takes constant time. Indeed, compute id(p), check if id(p) already appears in the hash table, if not, create a new linked list for the cell with this ID number, and store p in it. If a data-structure already exist for id(p), just add p to it.

This takes $O(n)$ time. Now, if any grid cell in $\mathsf{G}_r(\mathsf{P})$ contains more than, say, 9 points of P, then it must be that the $\mathcal{CP}(\mathsf{P}) < r$. Indeed, consider a cell C containing more than four points of P, and partition C into 3×3 equal squares. Clearly, one of those squares must contain two points of P, and let C' be this square. Clearly, the diameter of $C' = \operatorname{diam}(C)/3 = \sqrt{r^2 + r^2}/3 < r$. Thus, the (at least) two points of P in C' are in distance smaller than r from each other.

Thus, when we insert a point p, we can fetch all the points of P that were already inserted, for the cell of P, and the 8 adjacent cells. All those cells, must contain at most 9 points of P (otherwise, we would already have stopped since the $\mathcal{CP}(\cdot)$ of inserted points, is smaller than r). Let S be the set of all those points, and observe that $|S| \leq 9 \cdot 9 = O(1)$. Thus, we can compute by brute force the closest point to p in S. This takes $O(1)$ time. If $\mathsf{d}(p, S) < r$, we stop, otherwise, we continue to the next point, where $\mathsf{d}(p, S) = \min_{s\in S} \|ps\|$.

Overall, this takes $O(n)$ time. As for correctness, first observe that if $\mathcal{CP}(\mathsf{P}) > r$ then the algorithm would never make a mistake, since it returns '$\mathcal{CP}(\mathsf{P}) < r$' only after finding a pair of points of P with distance smaller than r. Thus, assume that p, q are the pair of points of P realizing the closest pair, and $\|pq\| = \mathcal{CP}(\mathsf{P}) < r$. Clearly, when the later of them, say p, is being inserted, the set S would contain q, and as such the algorithm would stop and return "$\mathcal{CP}(\mathsf{P}) < r$". ■

Lemma 2.3.3 hints on a natural way to compute $\mathcal{CP}(\mathsf{P})$. Indeed, permute the points of P in arbitrary fashion, and let $P = \langle p_1, \ldots, p_n \rangle$. Next, let $r_i = \mathcal{CP}(\{p_1, \ldots, p_i\})$. We can check if $r_{i+1} < r_i$, by just calling the algorithm for Lemma 2.3.3 on P_{i+1} and r_i. In fact, if $r_{i+1} < r_i$, the algorithm of Lemma 2.3.3, would give us back the distance r_{i+1} (with the other point realizing this distance).

In fact, consider the "good" case, where $r_{i+1} = r_i = r_{i-1}$. Namely, the length of the shortest pair does not change. In this case, we do not need to rebuild the data structure of Lemma 2.3.3, for each point. We can just reuse it from the previous iteration. Thus, inserting a single point takes constant time, as long as the closest pair does not change.

Things become bad, when $r_i < r_{i-1}$. Because then, we need to rebuild the grid, and reinsert all the points of $\mathsf{P}_i = \langle p_1, \ldots, p_i \rangle$ into the new grid $\mathsf{G}_{r_i}(\mathsf{P}_i)$. This takes $O(i)$ time.

So, if the closest pair radius, in the sequence $r_1, \ldots, r_n$ changes only k times, then the running time of our algorithm would be $O(nk)$. In fact, we can do even better.

Theorem 2.3.4. *Let P be a set of n points in the plane, one can compute the closest pair of points of P in expected linear time.*

Proof: Pick a random permutation of the points of P, let $\langle p_1, \ldots, p_n \rangle$ be this permutation. Let $r_2 = \|p_1 p_2\|$, and start inserting the points into the data structure of Lemma 2.3.3. In the ith iteration, if $r_i = r_{i-1}$, then this insertion takes constant time. If $r_i < r_{i-1}$, then we rebuild the grid and reinsert the points. Namely, we recompute $\mathsf{G}_{r_i}(\mathsf{P}_i)$.

To analyze the running time of this algorithm, let X_i be the indicator variable which is 1 if $r_i \neq r_{i-1}$, and 0 otherwise. Clearly, the running time is proportional to

$$R = 1 + \sum_{i=2}^{n} (1 + X_i \cdot i).$$

Thus, the expected running time is

$$\mathbf{E}[R] \quad = \quad 1 + \mathbf{E}\left[1 + \sum_{i=2}^{n} (1 + X_i \cdot i)\right] = n + \sum_{i=2}^{n} (\mathbf{E}[X_i] \cdot i) = n + \sum_{i=2}^{n} i \cdot \mathbf{Pr}[X_1 = 1],$$

by linearity of expectation and since for an indicator variable X_i, we have that $\mathbf{E}[X_i] = \mathbf{Pr}[X_i = 1]$.

Thus, we need to bound $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[r_i < r_{i-1}]$. To bound this quantity, fix the points of P_i, and randomly permute them. A point $\mathsf{q} \in \mathsf{P}_i$ is called *critical*, if $\mathcal{CP}(\mathsf{P}_i \setminus \{\mathsf{q}\}) > \mathcal{CP}(\mathsf{P}_i)$. If there are no critical points, then $r_{i-1} = r_i$ and then $\mathbf{Pr}[X_i = 1] = 0$. If there is one critical point, than $\mathbf{Pr}[X_i = 1] = 1/i$, as this is the probability that this critical point, would be the last point in the random permutation of P_i.

If there are two critical points, and let p, q be this unique pair of points of P_i realizing $\mathcal{CP}(\mathsf{P}_i)$. The quantity r_i is smaller than r_{i-1}, if either p or q are p_i. But the probability for that is $2/i$ (i.e., the probability in a random permutation of i objects, that one of two marked objects would be the last element in the permutation).

Observe, that there can not be more than two critical points. Indeed, if p and q are two points that realize the closest distance, than if there is a third critical point r, then $\mathcal{CP}(\mathsf{P}_i \setminus \{\mathsf{r}\}) = \|\mathsf{pq}\|$, and r is not critical.

We conclude that

$$\mathbf{E}[R] = n + \sum_{i=2}^{n} i \cdot \mathbf{Pr}[X_1 = 1] \leq n + \sum_{i=2}^{n} i \cdot \frac{2}{i} \leq 3n.$$

As such, the expected running time of this algorithm is $O(\mathbf{E}[R]) = O(n)$. ■

Theorem 2.3.4 is a surprising result, since it implies that *uniqueness* (i.e., deciding if n real numbers are all distinct) can be solved in linear time. However, there is a lower bound of $\Omega(n \log n)$ on uniqueness, using the comparison tree model. This reality dysfunction, can be easily explained, once one realizes that the model of computation of Theorem 2.3.4 is considerably stronger, using hashing, randomization, and the floor function.

2.4. Las Vegas and Monte Carlo algorithms

Definition 2.4.1. A ***Las Vegas algorithm*** is a randomized algorithms that *always* return the correct result. The only variant is that it's running time might change between executions.

An example for a Las Vegas algorithm is the **QuickSort** algorithm.

Definition 2.4.2. A ***Monte Carlo algorithm*** is a randomized algorithm that might output an incorrect result. However, the probability of error can be diminished by repeated executions of the algorithm.

The **MinCut** algorithm was an example of a Monte Carlo algorithm.

2.4.1. Complexity Classes

I assume people know what are Turing machines, **NP**, **NPC**, RAM machines, uniform model, logarithmic model. **PSPACE**, and **EXP**. If you do now know what are those things, you should read about them. Some of that is covered in the randomized algorithms book, and some other stuff is covered in any basic text on complexity theory.

Definition 2.4.3. The class **P** consists of all languages L that have a polynomial time algorithm **Alg**, such that for any input Σ^*, we have

- $x \in L \Rightarrow$ **Alg**(x) accepts,
- $x \notin L \Rightarrow$ **Alg**(x) rejects.

Definition 2.4.4. The class **NP** consists of all languages L that have a polynomial time algorithm **Alg**, such that for any input Σ^*, we have:

(i) If $x \in L \Rightarrow$ then $\exists y \in \Sigma^*$, **Alg**(x, y) accepts, where $|y|$ (i.e. the length of y) is bounded by a polynomial in $|x|$.

(ii) If $x \notin L \Rightarrow$ then $\forall y \in \Sigma^*$**Alg**(x, y) rejects.

Definition 2.4.5. For a complexity class $\mathcal{C}$, we define the complementary class co-$\mathcal{C}$ as the set of languages whose complement is in the class $\mathcal{C}$. That is

$$co - \mathcal{C} = \left\{ L \;\middle|\; \overline{L} \in \mathcal{C} \right\},$$

where $\overline{L} = \Sigma^* \setminus L$.

It is obvious that **P** = co−**P** and **P** $\subseteq$ **NP** $\cap$ co−**NP**. (It is currently unknown if **P** = **NP** $\cap$ co−**NP** or whether **NP** = co−**NP**, although both statements are believed to be false.)

Definition 2.4.6. The class **RP** (for Randomized Polynomial time) consists of all languages L that have a randomized algorithm **Alg** with worst case polynomial running time such that for any input $x \in \Sigma^*$, we have

(i) If $x \in L$ then **Pr**[**Alg**(x) accepts] $\geq 1/2$.

(ii) $x \notin L$ then **Pr**[**Alg**(x) accepts] $= 0$.

An **RP** algorithm is a Monte Carlo algorithm, but this algorithm can make a mistake only if $x \in L$. As such, co−**RP** is all the languages that have a Monte Carlo algorithm that make a mistake only if $x \notin L$. A problem which is in **RP** $\cap$ co−**RP** has an algorithm that does not make a mistake, namely a Las Vegas algorithm.

Definition 2.4.7. The class **ZPP** (for Zero-error Probabilistic Polynomial time) is the class of languages that have Las Vegas algorithms in expected polynomial time.

Definition 2.4.8. The class **PP** (for Probabilistic Polynomial time) is the class of languages that have a randomized algorithm **Alg** with worst case polynomial running time such that for any input $x \in \Sigma^*$, we have

(i) If $x \in L$ then **Pr**[**Alg**(x) accepts] $> 1/2$.

(ii) If $x \notin L$ then **Pr**[**Alg**(x) accepts] $< 1/2$.

The class **PP** is not very useful. Why?

Well, lets think about it. A randomized algorithm that just return yes/no with probability half is almost in **PP**, as it return the correct answer with probability half. An algorithm is in **PP** needs to be slightly better, and be correct with probability better than half, but how much better can be made to be arbitrarily close to $1/2$. In particular, there is no way to do effective amplification with such an algorithm.

Definition 2.4.9. The class **BPP** (for Bounded-error Probabilistic Polynomial time) is the class of languages that have a randomized algorithm **Alg** with worst case polynomial running time such that for any input $x \in \Sigma^*$, we have

(i) If $x \in L$ then $\mathbf{Pr}[\mathbf{Alg}(x) \text{ accepts}] \geq 3/4$.

(ii) If $x \notin L$ then $\mathbf{Pr}[\mathbf{Alg}(x) \text{ accepts}] \leq 1/4$.

2.5. Bibliographical notes

Section 2.4 follows [MR95, Section 1.5]. The closest-pair algorithm follows Golin *et al.* [GRSS95]. This is in turn a simplification of a result of Rabin [Rab76]. Smid provides a survey of such algorithms [Smi00].

Bibliography

[CLRS01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press / McGraw-Hill, 2001.

[GRSS95] M. Golin, R. Raman, C. Schwarz, and M. Smid. Simple randomized algorithms for closest pair problems. *Nordic J. Comput.*, 2:3–27, 1995.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[Rab76] M. O. Rabin. Probabilistic algorithms. In J. F. Traub, editor, *Algorithms and Complexity: New Directions and Recent Results*, pages 21–39. Academic Press, Orlando, FL, USA, 1976.

[Smi00] M. Smid. Closest-point problems in computational geometry. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 877–935. Elsevier, Amsterdam, The Netherlands, 2000.

Chapter 3

Min Cut

By Sariel Har-Peled, December 30, 2015[1]

> **To acknowledge the corn** - This purely American expression means to admit the losing of an argument, especially in regard to a detail; to retract; to admit defeat. It is over a hundred years old. Andrew Stewart, a member of Congress, is said to have mentioned it in a speech in 1828. He said that haystacks and cornfields were sent by Indiana, Ohio and Kentucky to Philadelphia and New York. Charles A. Wickliffe, a member from Kentucky questioned the statement by commenting that haystacks and cornfields could not walk. Stewart then pointed out that he did not mean literal haystacks and cornfields, but the horses, mules, and hogs for which the hay and corn were raised. Wickliffe then rose to his feet, and said, "Mr. Speaker, I acknowledge the corn".
>
> – Funk, Earle, A Hog on Ice and Other Curious Expressions.

3.1. Branching processes – Galton-Watson Process

3.1.1. The problem

In the 19th century, Victorians were worried that aristocratic surnames were disappearing, as family names passed on only through the male children. As such, a family with no male children had its family name disappear. So, imagine the number of male children of a person is an independent random variable $X \in \{0, 1, 2, \ldots\}$. Starting with a single person, its family (as far as male children are concerned) is a random tree with the degree of a node being distributed according to X. We continue recursively in constructing this tree, again, sampling the number of children for each current leaf according to the distribution of X. It is not hard to see that a family disappears if $\mathbf{E}[X] \leq 1$, and it has a constant probability of surviving if $\mathbf{E}[X] > 1$.

Francis Galton asked the question of what is the probability of such a blue-blood family name to survive, and this question was answered by Henry William Watson [WG75]. The Victorians were worried about strange things, see [Gre69] for a provocatively titled article from the period, and [Ste12] for a more recent take on this issue.

Of course, since infant mortality is dramatically down (as is the number of aristocrat males dying to maintain the British empire), the probability of family names to disappear is now much lower than it was in the 19th century. Interestingly, countries with family names that were introduced long time ago have very few surnames

(i.e., Korean have 250 surnames, and three surnames form 45% of the population). On the other hand, countries that introduced surnames more recently have dramatically more surnames (for example, the Dutch have surnames only for the last 200 years, and there are 68,000 different family names).

Here we are going to look on a very specific variant of this problem. Imagine that starting with a single male. A male has exactly two children, and one of them is a male with probability half (i.e., the Y-chromosome is being passed only to its male children). As such, the natural question is what is the probability that h generations down, there is a male decedent that all his ancestors are male (i.e., it caries the original family name, and the original Y-chromosome).

3.1.2. On coloring trees

Let T_h be a complete binary tree of height h. We randomly color its edges by black and white. Namely, for each edge we independently choose its color to be either black or white, with equal probability (say, black indicates the child is male). We are interested in the event that there exists a path from the root of T_h to one of its leafs, that is all black. Let $\mathcal{E}_h$ denote this event, and let $\rho_h = \mathbf{Pr}[\mathcal{E}_h]$. Observe that $\rho_0 = 1$ and $\rho_1 = 3/4$ (see below).

To bound this probability, consider the root u of T_h and its two children u_l and u_r. The probability that there is a black path from u_l to one of its children is ρ_{h-1}, and as such, the probability that there is a black path from u through u_l to a leaf of the subtree of u_l is $\mathbf{Pr}[\text{the edge } uu_l \text{ is colored black}] \cdot \rho_{h-1} = \rho_{h-1}/2$. As such, the probability that there is no black path through u_l is $1 - \rho_{h-1}/2$. As such, the probability of not having a black path from u to a leaf (through either children) is $(1 - \rho_{h-1}/2)^2$. In particular, there desired probability, is the complement; that is

$$\rho_h = 1 - \left(1 - \frac{\rho_{h-1}}{2}\right)^2 = \frac{\rho_{h-1}}{2}\left(2 - \frac{\rho_{h-1}}{2}\right) = \rho_{h-1} - \frac{\rho_{h-1}^2}{4}.$$

In particular, $\rho_0 = 1$, and $\rho_1 = 3/4$.

Lemma 3.1.1. *We have that $\rho_h \geq 1/(h+1)$.*

Proof: The proof is by induction. For $h = 1$, we have $\rho_1 = 3/4 \geq 1/(1+1)$.

Observe that $\rho_h = f(\rho_{h-1})$ for $f(x) = x - x^2/4$, and $f'(x) = 1 - x/2$. As such, $f'(x) > 0$ for $x \in [0,1]$ and $f(x)$ is increasing in the range $[0,1]$. As such, by induction, we have that $\rho_h = f(\rho_{h-1}) \geq f\left(\frac{1}{(h-1)+1}\right) = \frac{1}{h} - \frac{1}{4h^2}$. We need to prove that $\rho_h \geq 1/(h+1)$, which is implied by the above if

$$\frac{1}{h} - \frac{1}{4h^2} \geq \frac{1}{h+1} \quad \Leftrightarrow \quad 4h(h+1) - (h+1) \geq 4h^2 \quad \Leftrightarrow \quad 4h^2 + 4h - h - 1 \geq 4h^2 \quad \Leftrightarrow \quad 3h \geq 1,$$

which trivially holds. ■

Lemma 3.1.2. *We have that $\rho_h = O(1/h)$.*

Proof: The claim trivially holds for small values of h. Let h_j be the minimal index such that $\rho_{h_j} \leq 1/2^j$. It is easy to verify that $\rho_{h_j} \geq 1/2^{j+1}$. As such,

$$h_{j+1} - h_j \leq \frac{\rho_{h_j} - \rho_{h_{j+1}}}{\left(\rho_{h_{j+1}}\right)^2/4} \leq \frac{1/2^j - 1/2^{j+2}}{1/2^{2(j+2)+2}} = 2^{j+6} + 2^{j+4} = O\left(2^j\right).$$

Arguing similarly, we have

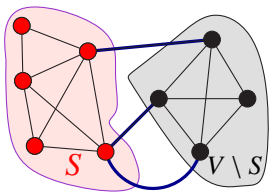
$$h_{j+2} - h_j \geq \frac{\rho_{h_j} - \rho_{h_{j+2}}}{\left(\rho_{h_j}\right)^2/4} \geq \frac{1/2^{j+1} - 1/2^{j+2}}{1/2^{2j+2}} = 2^{j+1} + 2^j = \Omega\left(2^j\right).$$

We conclude that $h_j = (h_j - h_{j-2}) + (h_{j-2} - h_{j-4}) + \cdots = \Omega(2^j)$, implying the claim. ■

3.2. Min Cut

3.2.1. Problem Definition

Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an undirected graph with n vertices and m edges. We are interested in *cuts* in G.



Definition 3.2.1. A ***cut*** in G is a partition of the vertices of V into two sets S and $\mathsf{V} \setminus S$, where the edges of the cut are

$$(S, \mathsf{V} \setminus S) = \left\{ uv \;\middle|\; u \in S, v \in \mathsf{V} \setminus S, \text{ and } uv \in E \right\},$$

where $S \neq \emptyset$ and $\mathsf{V} \setminus S \neq \emptyset$. We will refer to the number of edges in the cut $(S, \mathsf{V} \setminus S)$ as the *size of the cut*. For an example of a cut, see figure on the right.

We are interested in the problem of computing the ***minimum cut*** (i.e., *mincut*), that is, the cut in the graph with minimum cardinality. Specifically, we would like to find the set $S \subseteq \mathsf{V}$ such that $(S, \mathsf{V} \setminus S)$ is as small as possible, and S is neither empty nor $\mathsf{V} \setminus S$ is empty.

3.2.2. Some Definitions

We remind the reader of the following concepts. The *conditional probability* of X given Y is $\mathbf{Pr}\left[X = x \;\middle|\; Y = y\right] = \mathbf{Pr}[(X = x) \cap (Y = y)]/\mathbf{Pr}[Y = y]$. An equivalent, useful restatement of this is that

$$\mathbf{Pr}\left[(X = x) \cap (Y = y)\right] = \mathbf{Pr}\left[X = x \;\middle|\; Y = y\right] \cdot \mathbf{Pr}[Y = y]. \tag{3.1}$$

The following is easy to prove by induction using Eq. (3.1).

Lemma 3.2.2. *Let $\mathcal{E}_1, \ldots, \mathcal{E}_n$ be n events which are not necessarily independent. Then,*

$$\mathbf{Pr}\left[\cap_{i=1}^{n} \mathcal{E}_i\right] = \mathbf{Pr}\left[\mathcal{E}_1\right] * \mathbf{Pr}\left[\mathcal{E}_2 \;\middle|\; \mathcal{E}_1\right] * \mathbf{Pr}\left[\mathcal{E}_3 \;\middle|\; \mathcal{E}_1 \cap \mathcal{E}_2\right] * \ldots * \mathbf{Pr}\left[\mathcal{E}_n \;\middle|\; \mathcal{E}_1 \cap \ldots \cap \mathcal{E}_{n-1}\right].$$

3.3. The Algorithm

The basic operation used by the algorithm is ***edge contraction***, depicted in Figure 3.1. We take an edge $e = xy$ in G and merge the two vertices into a single vertex. The new resulting graph is denoted by G/xy. Note, that we remove self loops created by the contraction. However, since the resulting graph is no longer a regular graph, it has parallel edges – namely, it is a multi-graph. We represent a multi-graph, as a regular graph with multiplicities on the edges. See Figure 3.2.

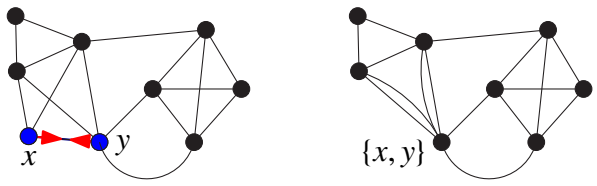


Figure 3.1: (a) A contraction of the edge xy. (b) The resulting graph.

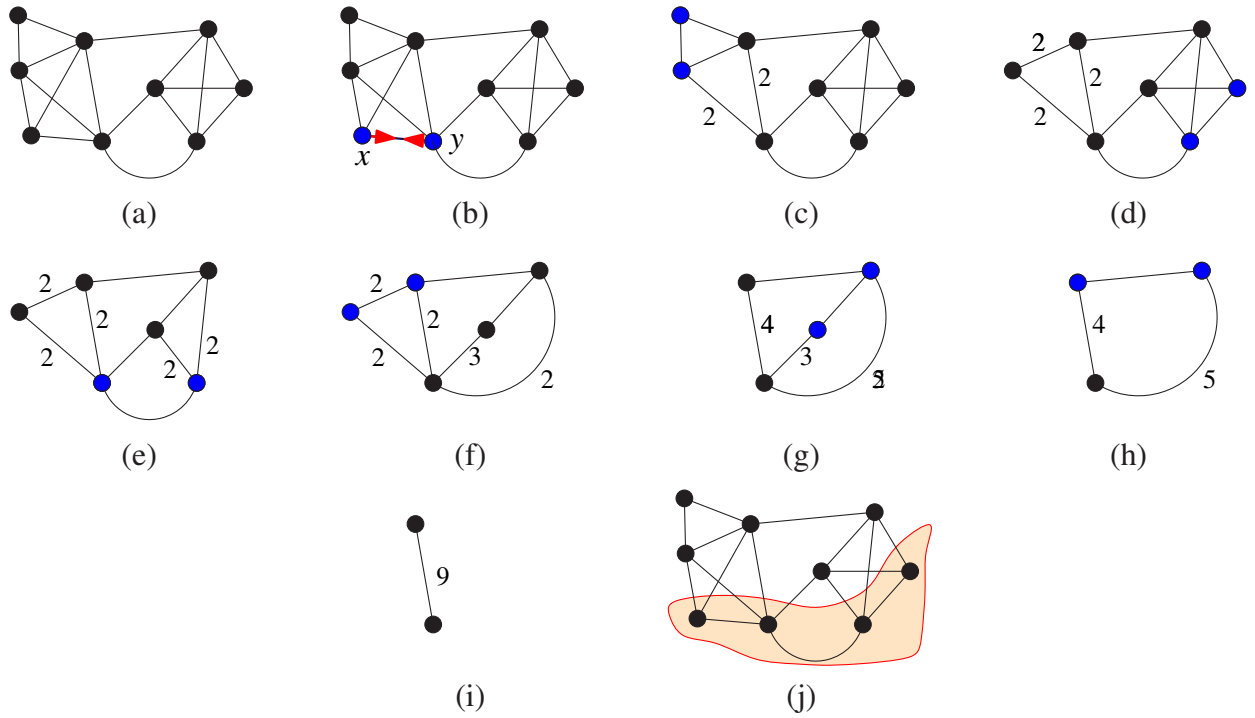


Figure 3.3: (a) Original graph. (b)–(j) a sequence of contractions in the graph, and (h) the cut in the original graph, corresponding to the single edge in (h). Note that the cut of (h) is not a mincut in the original graph.

The edge contraction operation can be implemented in $O(n)$ time for a graph with n vertices. This is done by merging the adjacency lists of the two vertices being contracted, and then using hashing to do the fix-ups (i.e., we need to fix the adjacency list of the vertices that are connected to the two vertices).

Note, that the cut is now computed counting multiplicities (i.e., if e is in the cut and it has weight w, then the contribution of e to the cut weight is w).

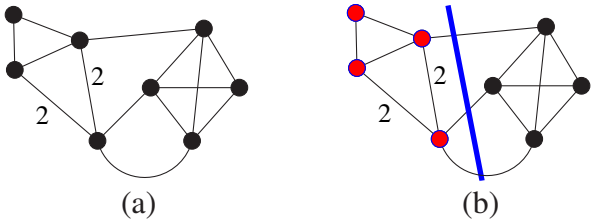


Figure 3.2: (a) A multi-graph. (b) A minimum cut in the resulting multi-graph.

Observation 3.3.1. *A set of vertices in* G/xy *corresponds to a set of vertices in the graph* G*. Thus a cut in* G/xy *always corresponds to a valid cut in* G*. However, there are cuts in* G *that do not exist in* G/xy*. For example, the cut* $S = \{x\}$*, does not exist in* G/xy*. As such, the size of the minimum cut in* G/xy *is at least as large as the minimum cut in* G *(as long as* G/xy *has at least one edge). Since any cut in* G/xy *has a corresponding cut of the same cardinality in* G*.*

Our algorithm works by repeatedly performing edge contractions. This is beneficial as this shrinks the underlying graph, and we would compute the cut in the resulting (smaller) graph. An "extreme" example of this, is shown in Figure 3.3, where we contract the graph into a single edge, which (in turn) corresponds to a cut in the original graph. (It might help the reader to think about each vertex in the contracted graph, as corresponding to a connected component in the original graph.)

Figure 3.3 also demonstrates the problem with taking this approach. Indeed, the resulting cut is not the minimum cut in the graph.

So, why did the algorithm fail to find the minimum cut in this case?[2] The failure occurs because of the contraction at Figure 3.3 (e), as we had contracted an edge in the minimum cut. In the new graph, depicted in Figure 3.3 (f), there is no longer a cut of size 3, and all cuts are of size 4 or more. Specifically, the algorithm succeeds only if it does not contract an edge in the minimum cut.

Observation 3.3.2. *Let $e_1, \ldots, e_{n-2}$ be a sequence of edges in G, such that none of them is in the minimum cut, and such that $\mathsf{G}' = G/\{e_1, \ldots, e_{n-2}\}$ is a single multi-edge. Then, this multi-edge corresponds to a minimum cut in G.*

Note, that the claim in the above observation is only in one direction. We might be able to still compute a minimum cut, even if we contract an edge in a minimum cut, the reason being that a minimum cut is not unique. In particular, another minimum cut might survived the sequence of contractions that destroyed other minimum cuts.

Using Observation 3.3.2 in an algorithm is problematic, since the argumentation is circular, how can we find a sequence of edges that are not in the cut without knowing what the cut is? The way to slice the Gordian knot here, is to randomly select an edge at each stage, and contract this random edge.

See Figure 3.4 for the resulting algorithm **MinCut**.

Algorithm **MinCut**(G)
  $\mathsf{G}_0 \leftarrow G$
  $i = 0$
  while G_i has more than two vertices **do**
    Pick randomly an edge e_i from the edges of G_i
    $\mathsf{G}_{i+1} \leftarrow G_i/e_i$
    $i \leftarrow i + 1$
  Let $(S, \mathsf{V} \setminus S)$ be the cut in the original graph corresponding to the single edge in G_i
  return $(S, \mathsf{V} \setminus S)$.

Figure 3.4: The minimum cut algorithm.

3.3.1. Analysis

3.3.1.1. The probability of success.

Naturally, if we are extremely lucky, the algorithm would never pick an edge in the mincut, and the algorithm would succeed. The ultimate question here is what is the probability of success. If it is relatively "large" then this algorithm is useful since we can run it several times, and return the best result computed. If on the other hand, this probability is tiny, then we are working in vain since this approach would not work.

Lemma 3.3.3. *If a graph G has a minimum cut of size k and G has n vertices, then $|E(G)| \geq \frac{kn}{2}$.*

Proof: Each vertex degree is at least k, otherwise the vertex itself would form a minimum cut of size smaller than k. As such, there are at least $\sum_{v \in \mathsf{V}} \text{degree}(v)/2 \geq nk/2$ edges in the graph. ■

Lemma 3.3.4. *If we pick in random an edge e from a graph G, then with probability at most $2/n$ it belong to the minimum cut.*

Proof: There are at least $nk/2$ edges in the graph and exactly k edges in the minimum cut. Thus, the probability of picking an edge from the minimum cut is smaller then $k/(nk/2) = 2/n$. ■

The following lemma shows (surprisingly) that **MinCut** succeeds with reasonable probability.

Lemma 3.3.5. **MinCut** *outputs the mincut with probability* $\geq \dfrac{2}{n(n-1)}$.

[2] Naturally, if the algorithm had succeeded in finding the minimum cut, this would have been **our** success.

Proof: Let $\mathcal{E}_i$ be the event that e_i is not in the minimum cut of G_i. By Observation 3.3.2, **MinCut** outputs the minimum cut if the events $\mathcal{E}_0, \ldots, \mathcal{E}_{n-3}$ all happen (namely, all edges picked are outside the minimum cut).

By Lemma 3.3.4, it holds $\mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \mathcal{E}_0 \cap \mathcal{E}_1 \cap \ldots \cap \mathcal{E}_{i-1}\Big] \geq 1 - \dfrac{2}{|\mathsf{V}(G_i)|} = 1 - \dfrac{2}{n-i}$. Implying that

$$\Delta = \mathbf{Pr}[\mathcal{E}_0 \cap \ldots \cap \mathcal{E}_{n-3}] = \mathbf{Pr}[\mathcal{E}_0] \cdot \mathbf{Pr}\Big[\mathcal{E}_1 \;\Big|\; \mathcal{E}_0\Big] \cdot \mathbf{Pr}\Big[\mathcal{E}_2 \;\Big|\; \mathcal{E}_0 \cap \mathcal{E}_1\Big] \cdot \ldots \cdot \mathbf{Pr}\Big[\mathcal{E}_{n-3} \;\Big|\; \mathcal{E}_0 \cap \ldots \cap \mathcal{E}_{n-4}\Big].$$

As such, we have

$$\Delta \geq \prod_{i=0}^{n-3}\left(1 - \frac{2}{n-i}\right) = \prod_{i=0}^{n-3} \frac{n-i-2}{n-i} = \frac{n-2}{n} * \frac{n-3}{n-1} * \frac{n-4}{n-2} \cdots \frac{2}{4} \cdot \frac{1}{3} = \frac{2}{n \cdot (n-1)}. \qquad \blacksquare$$

3.3.1.2. Running time analysis.

Observation 3.3.6. **MinCut** *runs in* $O(n^2)$ *time.*

Observation 3.3.7. *The algorithm always outputs a cut, and the cut is not smaller than the minimum cut.*

Definition 3.3.8. (informal) Amplification is the process of running an experiment again and again till the things we want to happen, with good probability, do happen.

Let **MinCutRep** be the algorithm that runs **MinCut** $n(n-1)$ times and return the minimum cut computed in all those independent executions of **MinCut**.

Lemma 3.3.9. *The probability that* **MinCutRep** *fails to return the minimum cut is* < 0.14.

Proof: The probability of failure of **MinCut** to output the mincut in each execution is at most $1 - \frac{2}{n(n-1)}$, by Lemma 3.3.5. Now, **MinCutRep** fails, only if all the $n(n-1)$ executions of **MinCut** fail. But these executions are independent, as such, the probability to this happen is at most

$$\left(1 - \frac{2}{n(n-1)}\right)^{n(n-1)} \leq \exp\left(-\frac{2}{n(n-1)} \cdot n(n-1)\right) = \exp(-2) < 0.14,$$

since $1 - x \leq e^{-x}$ for $0 \leq x \leq 1$. $\blacksquare$

Theorem 3.3.10. *One can compute the minimum cut in* $O(n^4)$ *time with constant probability to get a correct result. In* $O\left(n^4 \log n\right)$ *time the minimum cut is returned with high probability.*

3.4. A faster algorithm

The algorithm presented in the previous section is extremely simple. Which raises the question of whether we can get a faster algorithm[③]?

So, why **MinCutRep** needs so many executions? Well, the probability of success in the first ν iterations is

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}_0 \cap \ldots \cap \mathcal{E}_{\nu-1}] &\geq \prod_{i=0}^{\nu-1}\left(1 - \frac{2}{n-i}\right) = \prod_{i=0}^{\nu-1} \frac{n-i-2}{n-i} \\ &= \frac{n-2}{n} * \frac{n-3}{n-1} * \frac{n-4}{n-2} \cdots = \frac{(n-\nu)(n-\nu-1)}{n \cdot (n-1)}. \end{aligned} \tag{3.2}$$

[③] This would require a more involved algorithm, thats life.

```
Contract ( G, t )
begin
   while |(G)| > t do
         Pick a random edge e in G.
         G ← G/e
   return G
end
```

```
FastCut(G = (V, E))
   G – multi-graph
begin
   n ← |V(G)|
   if n ≤ 6 then
         Compute (via brute force) minimum cut
         of G and return cut.
   t ← ⌈1 + n/√2⌉
   H_1 ← Contract(G, t)
   H_2 ← Contract(G, t)
   /* Contract is randomized!!! */
   X_1 ← FastCut(H_1),
   X_2 ← FastCut(H_2)
   return minimum cut out of X_1 and X_2.
end
```

Figure 3.5: **Contract**(G,t) shrinks G till it has only t vertices. **FastCut** computes the minimum cut using **Contract**.

Namely, this probability deteriorates very quickly toward the end of the execution, when the graph becomes small enough. (To see this, observe that for $\nu = n/2$, the probability of success is roughly $1/4$, but for $\nu = n - \sqrt{n}$ the probability of success is roughly $1/n$.)

So, the key observation is that as the graph get smaller the probability to make a bad choice increases. So, instead of doing the amplification from the outside of the algorithm, we will run the new algorithm more times when the graph is smaller. Namely, we put the amplification directly into the algorithm.

The basic new operation we use is **Contract**, depicted in Figure 3.5, which also depict the new algorithm **FastCut**.

Lemma 3.4.1. *The running time of* **FastCut**(G) *is* $O\left(n^2 \log n\right)$, *where* $n = |V(G)|$.

Proof: Well, we perform two calls to **Contract**(G,t) which takes $O(n^2)$ time. And then we perform two recursive calls on the resulting graphs. We have:

$$T(n) = O\left(n^2\right) + 2T\left(\frac{n}{\sqrt{2}}\right).$$

The solution to this recurrence is $O\left(n^2 \log n\right)$ as one can easily (and should) verify. ■

Exercise 3.4.2. Show that one can modify **FastCut** so that it uses only $O(n^2)$ space.

Lemma 3.4.3. *The probability that* **Contract**$\left(\mathsf{G}, n/\sqrt{2}\right)$ *had not contracted the minimum cut is at least* $1/2$.

Namely, the probability that the minimum cut in the contracted graph is still a minimum cut in the original graph is at least $1/2$.

Proof: Just plug in $\nu = n - t = n - \left\lceil 1 + n/\sqrt{2} \right\rceil$ into Eq. (3.2). We have

$$\mathbf{Pr}\left[\mathcal{E}_0 \cap \ldots \cap \mathcal{E}_{n-t}\right] \geq \frac{t(t-1)}{n \cdot (n-1)} = \frac{\left\lceil 1 + n/\sqrt{2} \right\rceil \left(\left\lceil 1 + n/\sqrt{2} \right\rceil - 1\right)}{n(n-1)} \geq \frac{1}{2}.$$ ■

The following lemma bounds the probability of success.

Lemma 3.4.4. **FastCut** *finds the minimum cut with probability larger than* $\Omega(1/\log n)$.

Proof: Let T_h be the recursion tree of the algorithm of depth $h = \Theta(\log n)$. Color an edge of recursion tree by black if the contraction succeeded. Clearly, the algorithm succeeds if there is a path from the root to a leaf that is all black. This is exactly the settings of Lemma 3.1.1, and we conclude that the probability of success is at least $1/(h+1) = \Theta(1/\log n)$, as desired. ■

Exercise 3.4.5. Prove, that running **FastCut** repeatedly $c \cdot \log^2 n$ times, guarantee that the algorithm outputs the minimum cut with probability $\geq 1 - 1/n^2$, say, for c a constant large enough.

Theorem 3.4.6. *One can compute the minimum cut in a graph* G *with* n *vertices in* $O(n^2 \log^3 n)$ *time. The algorithm succeeds with probability* $\geq 1 - 1/n^2$.

Proof: We do amplification on **FastCut** by running it $O(\log^2 n)$ times. The running time bound follows from Lemma 3.4.1. The bound on the probability follows from Lemma 3.4.4, and using the amplification analysis as done in Lemma 3.3.9 for **MinCutRep**. ■

3.5. Bibliographical Notes

The **MinCut** algorithm was developed by David Karger during his PhD thesis in Stanford. The fast algorithm is a joint work with Clifford Stein. The basic algorithm of the mincut is described in [MR95, pages 7–9], the faster algorithm is described in [MR95, pages 289–295].

3.5.0.0.1. Galton-Watson process. The idea of using coloring of the edges of a tree to analyze **FastCut** might be new (i.e., Section 3.1.2).

Bibliography

[Gre69] W.R. Greg. *Why are Women Redundant?* Trübner, 1869.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[Ste12] E. Steinlight. Why novels are redundant: Sensation fiction and the overpopulation of literature. *ELH*, 79(2):501–535, 2012.

[WG75] H. W. Watson and F. Galton. On the probability of the extinction of families. *J. Anthrop. Inst. Great Britain*, 4:138–144, 1875.

Chapter 4

The Occupancy and Coupon Collector problems

By Sariel Har-Peled, December 30, 2015[①]

4.1. Preliminaries

Definition 4.1.1 (Variance and Standard Deviation). For a random variable X, let $\mathbf{V}\left[X\right] = \mathbf{E}\left[(X-\mu_X)^2\right] = \mathbf{E}\left[X^2\right] - \mu_X^2$ denote the *variance* of X, where $\mu_X = \mathbf{E}\left[X\right]$. Intuitively, this tells us how concentrated is the distribution of X.

The *standard deviation* of X, denoted by σ_X is the quantity $\sqrt{\mathbf{V}[X]}$.

Observation 4.1.2. *(i) For any constant $c \geq 0$, we have $\mathbf{V}\left[cX\right] = c^2\,\mathbf{V}\left[X\right]$.*

(ii) For X and Y independent variables, we have $\mathbf{V}\left[X+Y\right] = \mathbf{V}\left[X\right] + \mathbf{V}\left[Y\right]$.

Definition 4.1.3 (Bernoulli distribution). Assume, that one flips a coin and get 1 (heads) with probability p, and 0 (i.e., tail) with probability $q = 1-p$. Let X be this random variable. The variable X is has *Bernoulli distribution* with parameter p.

We have that $\mathbf{E}\left[X\right] = 1\cdot p + 0\cdot(1-p) = p$, and

$$\mathbf{V}\left[X\right] = \mathbf{E}\left[X^2\right] - \mu_X^2 = \mathbf{E}\left[X^2\right] - p^2 = p - p^2 = p(1-p) = pq.$$

Definition 4.1.4 (Binomial distribution). Assume that we repeat a Bernoulli experiment n times (independently!). Let $X_1,\ldots,X_n$ be the resulting random variables, and let $X = X_1 + \cdots + X_n$. The variable X has the *binomial distribution* with parameters n and p. We denote this fact by $X \sim \mathrm{Bin}(n,p)$. We have

$$b(k;n,p) = \mathbf{Pr}\left[X=k\right] = \binom{n}{k} p^k q^{n-k}.$$

Also, $\mathbf{E}\left[X\right] = np$, and $\mathbf{V}\left[X\right] = \mathbf{V}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{V}\left[X_i\right] = npq$.

Observation 4.1.5. *Let $C_1, \ldots, C_n$ be random events (not necessarily independent). Than*

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

(This is usually referred to as the union bound*.) If $C_1, \ldots, C_n$ are* disjoint *events then*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} C_i\right] = \sum_{i=1}^{n} \mathbf{Pr}[C_i]\,.$$

4.1.1. Geometric distribution

Definition 4.1.6. Consider a sequence $X_1, X_2, \ldots$ of independent Bernoulli trials with probability p for success. Let X be the number of trials one has to perform till encountering the first success. The distribution of X is *geometric distribution* with parameter p. We denote this by $X \sim \mathrm{Geom}(p)$.

Lemma 4.1.7. *For a variable $X \sim \mathrm{Geom}(p)$, we have, for all i, that $\mathbf{Pr}\Big[X = i\Big] = (1-p)^{i-1}p$. Furthermore, $\mathbf{E}\Big[X\Big] = 1/p$ and $\mathbf{V}\Big[X\Big] = (1-p)/p^2$.*

Proof: The proof of the expectation and variance is included for the sake of completeness, and the reader is of course encouraged to skip (reading) this proof. So, let $f(x) = \sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$, and observe that $f'(x) = \sum_{i=1}^{\infty} i x^{i-1} = (1-x)^{-2}$. As such, we have

$$\mathbf{E}\Big[X\Big] = \sum_{i=1}^{\infty} i\,(1-p)^{i-1}p = pf'(1-p) = \frac{p}{(1-(1-p))^2} = \frac{1}{p}.$$

$$\mathbf{V}\Big[X\Big] = \mathbf{E}\Big[X^2\Big] - \frac{1}{p^2} = \sum_{i=1}^{\infty} i^2\,(1-p)^{i-1}p - \frac{1}{p^2}. = p + p(1-p)\sum_{i=2}^{\infty} i^2\,(1-p)^{i-2} - \frac{1}{p^2}.$$

We need to do a similar trick to this end, we observe that

$$f''(x) = \sum_{i=2}^{\infty} i(i-1)x^{i-2} = \left((1-x)^{-1}\right)'' = \frac{2}{(1-x)^3}.$$

As such, we have that

$$\begin{aligned}
\Delta(x) &= \sum_{i=2}^{\infty} i^2 x^{i-2} = \sum_{i=2}^{\infty} i(i-1)x^{i-2} + \sum_{i=2}^{\infty} i x^{i-2} = f''(x) + \frac{1}{x}\sum_{i=2}^{\infty} i x^{i-1} = f''(x) + \frac{1}{x}(f'(x) - 1) \\
&= \frac{2}{(1-x)^3} + \frac{1}{x}\left(\frac{1}{(1-x)^2} - 1\right) = \frac{2}{(1-x)^3} + \frac{1}{x}\left(\frac{1-(1-x)^2}{(1-x)^2}\right) = \frac{2}{(1-x)^3} + \frac{1}{x}\cdot\frac{x(2-x)}{(1-x)^2} \\
&= \frac{2}{(1-x)^3} + \frac{2-x}{(1-x)^2}.
\end{aligned}$$

As such, we have that

$$\begin{aligned}
\mathbf{V}\Big[X\Big] &= p + p(1-p)\Delta(1-p) - \frac{1}{p^2} = p + p(1-p)\left(\frac{2}{p^3} + \frac{1+p}{p^2}\right) - \frac{1}{p^2} = p + \frac{2(1-p)}{p^2} + \frac{1-p^2}{p} - \frac{1}{p^2} \\
&= \frac{p^3 + 2(1-p) + p - p^3 - 1}{p^2} = \frac{1-p}{p^2}.
\end{aligned}$$

■

4.1.2. Some needed math

Lemma 4.1.8. *For any positive integer n, we have:*

(i) $(1 + 1/n)^n \leq e$.

(ii) $(1 - 1/n)^{n-1} \geq e^{-1}$.

(iii) $n! \geq (n/e)^n$.

(iv) For any $k \leq n$, *we have:* $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

Proof: (i) Indeed, $1 + 1/n \leq \exp(1/n)$, since $1 + x \leq e^x$, for $x \geq 0$. As such $(1 + 1/n)^n \leq \exp(n(1/n)) = e$.

(ii) Rewriting the inequality, we have that we need to prove $\left(\frac{n-1}{n}\right)^{n-1} \geq \frac{1}{e}$. This is equivalence to proving $e \geq \left(\frac{n}{n-1}\right)^{n-1} = \left(1 + \frac{1}{n-1}\right)^{n-1}$, which is our friend from (i).

(iii) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

(iv) Indeed, for any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$ since $kn - n = n(k-1) \leq k(n-1) = kn - k$. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k-1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-i}{k-i} \cdots \frac{n-k+1}{1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

As for the other direction, we have

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{\left(\frac{k}{e}\right)^k} = \left(\frac{ne}{k}\right)^k,$$

by (iii). ∎

4.2. Occupancy Problems

Problem 4.2.1. We are throwing m balls into n bins randomly (i.e., for every ball we randomly and uniformly pick a bin from the n available bins, and place the ball in the bin picked). There are many natural questions one can ask here:

(A) What is the maximum number of balls in any bin?

(B) What is the number of bins which are empty?

(C) How many balls do we have to throw, such that all the bins are non-empty, with reasonable probability?

Let X_i be the number of balls in the ith bins, when we throw n balls into n bins (i.e., $m = n$). Clearly,

$$\mathbf{E}\left[X_i\right] = \sum_{j=1}^{n} \mathbf{Pr}[\text{The } j\text{th ball fall in } i\text{th bin}] = n \cdot \frac{1}{n} = 1,$$

by linearity of expectation. The probability that the first bin has exactly i balls is

$$\binom{n}{i}\left(\frac{1}{n}\right)^i\left(1 - \frac{1}{n}\right)^{n-i} \leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$$

This follows by Lemma 4.1.8 (iv).

Let $C_j(k)$ be the event that the jth bin has k or more balls in it. Then,

$$\mathbf{Pr}\Big[C_1(k)\Big] \leq \sum_{i=k}^{n} \left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \frac{e^2}{k^2} + \ldots\right) = \left(\frac{e}{k}\right)^k \frac{1}{1 - e/k}.$$

Let $k^* = \lceil (3 \ln n)/\ln\ln n \rceil$. Then,

$$\begin{aligned}
\mathbf{Pr}\Big[C_1(k^*)\Big] &\leq \left(\frac{e}{k^*}\right)^{k^*} \frac{1}{1 - e/k^*} \leq 2\left(\frac{e}{(3\ln n)/\ln\ln n}\right)^{k^*} = 2\Big(\exp(1 - \ln 3 - \ln\ln n + \ln\ln\ln n)\Big)^{k^*} \\
&\leq 2(\exp(-\ln\ln n + \ln\ln\ln n))^{k^*} \\
&\leq 2\exp\left(-3\ln n + 6\ln n \frac{\ln\ln\ln n}{\ln\ln n}\right) \leq 2\exp(-2.5\ln n) \leq \frac{1}{n^2},
\end{aligned}$$

for n large enough. We conclude, that since there are n bins and they have identical distributions that

$$\mathbf{Pr}[\text{any bin contains more than } k^* \text{ balls}] \leq \sum_{i=1}^{n} C_i(k^*) \leq \frac{1}{n}.$$

Theorem 4.2.2. *With probability at least $1 - 1/n$, no bin has more than $k^* = \left\lceil \frac{3\ln n}{\ln\ln n} \right\rceil$ balls in it.*

Exercise 4.2.3. Show that for $m = n \ln n$, with probability $1 - o(1)$, every bin has $O(\log n)$ balls.

It is interesting to note, that if at each iteration we randomly pick d bins, and throw the ball into the bin with the smallest number of balls, then one can do much better. We currently do not have the machinery to prove the following theorem, but hopefully we would prove it later in the course.

Theorem 4.2.4. *Suppose that n balls are sequentially places into n bins in the following manner. For each ball, $d \geq 2$ bins are chosen independently and uniformly at random (with replacement). Each ball is placed in the least full of the d bins at the time of placement, with ties broken randomly. After all the balls are places, the maximum load of any bin is at most $\ln\ln n/(\ln d) + O(1)$, with probability at least $1 - o(1/n)$.*

Note, even by setting $d = 2$, we get considerable improvement. A proof of this theorem can be found in the work by Azar *et al.* [ABKU00].

4.2.1. The Probability of all bins to have exactly one ball

Next, we are interested in the probability that all m balls fall in distinct bins. Let X_i be the event that the ith ball fell in a distinct bin from the first $i - 1$ balls. We have:

$$\begin{aligned}
\mathbf{Pr}\left[\bigcap_{i=2}^{m} X_i\right] &= \mathbf{Pr}[X_2] \prod_{i=3}^{m} \mathbf{Pr}\left[X_i \,\middle|\, \bigcap_{j=2}^{i-1} X_j\right] \leq \prod_{i=2}^{m}\left(\frac{n - i + 1}{n}\right) \leq \prod_{i=2}^{m}\left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=2}^{m} e^{-(i-1)/n} \leq \exp\left(-\frac{m(m-1)}{2n}\right),
\end{aligned}$$

thus for $m = \left\lceil \sqrt{2n} + 1 \right\rceil$, the probability that all the m balls fall in different bins is smaller than $1/e$.

This is sometime referred to as the *birthday paradox*. You have $m = 30$ people in the room, and you ask them for the date (day and month) of their birthday (i.e., $n = 365$). The above shows that the probability of all birthdays to be distinct is $\exp(-30 \cdot 29/730) \leq 1/e$. Namely, there is more than 50% chance for a birthday collision, a simple but counterintuitive phenomena.

4.3. The Markov and Chebyshev's inequalities

We remind the reader that for a random variable X assuming real values, its *expectation* is $\mathbf{E}[Y] = \sum_y y \cdot \mathbf{Pr}[Y = y]$. Similarly, for a function $f(\cdot)$, we have $\mathbf{E}[f(Y)] = \sum_y f(y) \cdot \mathbf{Pr}[Y = y]$.

Theorem 4.3.1 (Markov's Inequality). *Let Y be a random variable assuming only non-negative values. Then for all $t > 0$, we have*

$$\mathbf{Pr}\Big[Y \geq t\Big] \leq \frac{\mathbf{E}[Y]}{t}$$

Proof: Indeed,

$$\begin{aligned}\mathbf{E}\Big[Y\Big] &= \sum_{y \geq t} y \,\mathbf{Pr}\Big[Y = y\Big] + \sum_{y<t} y \,\mathbf{Pr}[Y = y] \geq \sum_{y \geq t} y \,\mathbf{Pr}\Big[Y = y\Big] \\ &\geq \sum_{y \geq t} t \,\mathbf{Pr}\Big[Y = y\Big] = t \,\mathbf{Pr}\Big[Y \geq t\Big].\end{aligned}$$

■

Markov inequality is tight, as the following exercise testifies.

Exercise 4.3.2. For any (integer) $k > 1$, define a random positive variable X_k such that $\mathbf{Pr}[X_k \geq k\,\mathbf{E}[X_k]] = \frac{1}{k}$.

Theorem 4.3.3 (Chebyshev's inequality). $\mathbf{Pr}\Big[|X - \mu_X| \geq t\sigma_X\Big] \leq \frac{1}{t^2}$, *where* $\mu_X = \mathbf{E}\Big[X\Big]$ *and* $\sigma_X = \sqrt{\mathbf{V}[X]}$.

Proof: Note that

$$\mathbf{Pr}\Big[|X - \mu_X| \geq t\sigma_X\Big] = \mathbf{Pr}\Big[\,(X - \mu_X)^2 \geq t^2\sigma_X^2\Big].$$

Set $Y = (X - \mu_X)^2$. Clearly, $\mathbf{E}\Big[Y\Big] = \sigma_X^2$. Now, apply Markov's inequality to Y. ■

4.4. The Coupon Collector's Problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

Let $C_i \in \Big\{1, \ldots, n\Big\}$ be the coupon picked in the ith trial. The jth trial is a success, if C_j was not picked before in the first $j - 1$ trials. Let X_i denote the number of trials from the ith success, till after the $(i + 1)$th success. Clearly, the number of trials performed is

$$X = \sum_{i=0}^{n-1} X_i.$$

Now, the probability of X_i to succeed in a trial is $p_i = (n - i)/n$, and X_i has the geometric distribution with probability p_i. As such $\mathbf{E}\Big[X_i\Big] = 1/p_i$, and $\mathbf{V}\Big[X_i\Big] = q/p^2 = (1 - p_i)/p_i^2$.

Thus,

$$\mathbf{E}\Big[X\Big] = \sum_{i=0}^{n-1} \mathbf{E}\Big[X_i\Big] = \sum_{i=0}^{n-1} \frac{n}{n-i} = nH_n = n(\ln n + \Theta(1)) = n \ln n + O(n),$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the nth Harmonic number.

As for variance, using the independence of $X_0, \ldots, X_{n-1}$, we have

$$\mathbf{V}\Big[X\Big] = \sum_{i=0}^{n-1} \mathbf{V}\Big[X_i\Big] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{1-(n-i)/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i}{n}\left(\frac{n}{n-i}\right)^2$$
$$= n \sum_{i=0}^{n-1} \frac{i}{(n-i)^2} = n \sum_{i=1}^{n} \frac{n-i}{i^2} = n\left(\sum_{i=1}^{n} \frac{n}{i^2} - \sum_{i=1}^{n} \frac{1}{i}\right) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n.$$

Since, $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \pi^2/6$, we have $\lim_{n\to\infty} \dfrac{\mathbf{V}[X]}{n^2} = \dfrac{\pi^2}{6}$.

Corollary 4.4.1. *Let X be the number of rounds till we collection all n coupons. Then, $\mathbf{V}\Big[X\Big] \approx \left(\frac{\pi^2}{6}\right)n^2$ and its standard deviation is $\sigma_X \approx \frac{\pi}{\sqrt{6}}n$.*

This implies a weak bound on the concentration of X, using Chebyshev inequality, but this is going to be quite weaker than what we implied we can do. Indeed, we have

$$\mathbf{Pr}\left[X \geq n \log n + n + t \cdot n\frac{\pi}{\sqrt{6}}\right] \leq \mathbf{Pr}\left[\Big|X - \mathbf{E}[X]\Big| \geq t\sigma_X\right] \leq \frac{1}{t^2},$$

Note, that this is somewhat approximate, and hold for n sufficiently large.

4.5. Notes

The material in this note covers parts of [MR95, sections 3.1,3.2,3.6]

Bibliography

[ABKU00] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, 2000.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

Chapter 5

Sampling, Estimation, and More on the Coupon's Collector Problems II

By Sariel Har-Peled, December 30, 2015[①]

> There is not much talking now. A silence falls upon them all. This is no time to talk of hedges and fields, or the beauties of any country. Sadness and fear and hate, how they well up in the heart and mind, whenever one opens the pages of these messengers of doom. Cry for the broken tribe, for the law and custom that is gone. Aye, and cry aloud for the man who is dead, for the woman and children bereaved. Cry, the beloved country, these things are not yet at an end. The sun pours down on the earth, on the lovely land that man cannot enjoy. He knows only the fear of his heart.
>
> – Alan Paton, Cry, the beloved country.

5.1. Randomized selection – Using sampling to learn the world

5.1.1. Sampling

One of the big advantages of randomized algorithms, is that they sample the world; that is, learn how the input looks like without reading all the input. For example, consider the following problem: We are given a set of U of n objects $u_1, \ldots, u_n$. and we want to compute the number of elements of U that have some property. Assume, that one can check if this property holds, in constant time, for a single object, and let $\psi(u)$ be the function that returns 1 if the property holds for the element u. and zero otherwise. Now, let α be the number of objects in U that have this property. We want to reliably estimate α without computing the property for all the elements of U.

A natural approach, would be to pick a random sample R of m objects, $r_1, \ldots, r_m$ from U (with repetition), and compute $Y = \sum_{i=1}^{m} \psi(r_1)$, and our estimate for α is $\beta = (n/m)Y$. It is natural to ask how far is β from the true estimate.

Lemma 5.1.1. *Let U be a set of n elements, with α of them having a certain property ψ. Let R be a uniform random sample from U (with repetition), and let Y be the number of elements in R that have the property ψ, and let $Z = (n/m)Y$ be the estimate for α. Then, for any $t \geq 1$, we have that*

$$\mathbf{Pr}\left[\alpha - t\frac{n}{2\sqrt{m}} \leq Z \leq \alpha + t\frac{n}{2\sqrt{m}}\right] \geq 1 - \frac{1}{t^2}.$$

Similarly, we have that $\mathbf{Pr}\Big[\mathbf{E}[Y] - t\sqrt{m}/2 \leq Y \leq \mathbf{E}[Y] + t\sqrt{m}/2\Big] \geq 1 - \frac{1}{t^2}$.

Proof: Let $Y_i = \psi(r_i)$ be an indicator variable that is 1 if the ith sample r_i has the property ψ. Now, $Y = \sum_i Y_i$ is a binomial distribution with probability $p = \alpha/n$, and m samples; that is, $Y \sim \mathrm{Bin}(m, p)$. We saw in the previous lecture that, $\mathbf{E}[Y] = mp$, $\mathbf{V}[Y] = mp(1-p)$, and its standard deviation is as such $\sigma_Y = \sqrt{mp(1-p)} \leq \sqrt{m}/2$, as $\sqrt{p(1-p)}$ is maximized for $p = 1/2$. We have $\Delta = \frac{t\sigma_Y n}{m} \leq t\frac{\sqrt{m}n}{2m} = t\frac{n}{2\sqrt{m}}$, since $\sqrt{(\alpha/n)(1-(\alpha/n))}$ is maximized for $\alpha = n/2$. As such,

$$\begin{aligned}\mathbf{Pr}\left[\left|Z - \alpha\right| \geq t\frac{n}{2\sqrt{m}}\right] &\leq \mathbf{Pr}\Big[\left|Z - \alpha\right| \geq \Delta\Big] = \mathbf{Pr}\left[\left|\frac{n}{m}Y - \alpha\right| \geq \Delta\right] = \mathbf{Pr}\left[\left|Y - \frac{m}{n}\alpha\right| \geq \frac{m}{n}\Delta\right] \\ &= \mathbf{Pr}\Big[\left|Y - \mathbf{E}[Y]\right| \geq t\sigma_Y\Big] \leq \frac{1}{t^2},\end{aligned}$$

by Chebychev's inequality. ∎

5.1.1.1. Inverse estimation

We are given a set $U = \{u_1, \ldots, u_n\}$ of n distinct numbers. Let s_i denote the ith smallest number in U – that is s_i is the number of *rank* i in U. We are interested in estimating s_k quickly. So, let us take a sample R of size m. Let $\mathsf{R}_{\leq s_k}$ be the set of all the numbers in R that are $\leq s_k$. For $Y = \left|\mathsf{R}_{\leq s_k}\right|$, we have that $\mu = \mathbf{E}[Y] = km/n$. Furthermore, for any $t \geq 1$, Lemma 5.1.1 implies that $\mathbf{Pr}\Big[\mu - t\sqrt{m}/2 \leq Y \leq \mu + t\sqrt{m}/2\Big] \geq 1 - \frac{1}{t^2}$. In particular, with probability $\geq 1 - 1/t^2$ the number r_- of rank $\ell_- = \Big\lfloor \mu - t\sqrt{m}/2 \Big\rfloor - 1$ in R is smaller than s_k, and similarly, the number r_+ of rank $\ell_+ = \Big\lceil \mu + t\sqrt{m}/2 \Big\rceil + 1$ in R is larger than s_k.

One can conceptually think about the interval $\mathcal{I}(k) = [r_-, r_+]$ as confidence interval – we know that $s_k \in \mathcal{I}(k)$ with probability $\geq 1 - 1/t^2$. But how big is this interval? Namely, how many elements are there in $\mathcal{I}(k) \cap Sample$?

To this end, consider the interval of ranks in the sample that might contain the kth element. By the above, this is

$$\mathcal{I}(k,t) = k\frac{n}{m} + \Big[-t\sqrt{m}/2 - 1, t\sqrt{m}/2 + 1\Big].$$

In particular, consider the maximum $\nu \leq k$, such that $\mathcal{I}(\nu, t)$ and $\mathcal{I}(k,t)$ are disjoint. We have the condition that

$$\nu\frac{n}{m} + t\sqrt{m}/2 + 1 \leq k\frac{n}{m} - t\sqrt{m}/2 - 1 \implies \nu \leq k - t\frac{m^{3/2}}{n} - 1.$$

Setting $g = k - t\frac{m^{3/2}}{n} - 1$ and $h = k + t\frac{m^{3/2}}{n} + 1$, we have that $\mathcal{I}(g,t)$ and $\mathcal{I}(k,t)$ and $\mathcal{I}(h,t)$ are all disjoint with probability $\geq 1 - 3/t^2$.

To this end, let $g = k - \Big\lceil 2\Big(t\frac{n}{2\sqrt{m}}\Big)\Big\rceil$ and $h = k + \Big\lceil 2\Big(t\frac{n}{2\sqrt{m}}\Big)\Big\rceil$. It is easy to verify (using the same argumentation as above) that with probability at least $1 - 3/t^3$, the three confidence $\mathcal{I}(g), \mathcal{I}(k)$ and $\mathcal{I}(h)$ do not intersect. As such, we have

$$\Big|\mathcal{I}(k) \cap \mathsf{R}\Big| \leq h - g \leq 4\left(t\frac{n}{2\sqrt{m}}\right).$$

We thus get the following.

Func **LazySelect**(S , k)
  `Input:` S - set of n elements, k - index of element to be output.
begin
  repeat
    $R \leftarrow \left\{\text{Sample with replacement of } n^{3/4} \text{ elements from } S\right\}$
      $\cup \{-\infty, +\infty\}$.
    Sort R.
    $l \leftarrow \max\left(1, \left\lfloor kn^{-1/4} - \sqrt{n} \right\rfloor\right)$, $h \leftarrow \min\left(n^{3/4}, \left\lfloor kn^{-1/4} + \sqrt{n} \right\rfloor\right)$
    $a \leftarrow R_{(l)}$, $b \leftarrow R_{(h)}$.
    Compute the ranks $r_S(a)$ and $r_S(b)$ of b in S
      `/* using` $2n$ `comparisons */`
    $P \leftarrow \left\{y \in S \;\middle|\; a \le y \le b\right\}$
    `/* done when computing the rank of` a `and` b `*/`
  Until $(r_S(a) \le k \le r_S(b))$ **and** $\left(|P| \le 8n^{3/4} + 2\right)$
  Sort P in $O(n^{3/4} \log n)$ time.
  return $P_{k - r_S(a) + 1}$
end **LazySelect**

Figure 5.1: The **LazySelect** algorithm.

Lemma 5.1.2. *Given a set U of n numbers, a number k, and parameters t and m, one can compute, in $O(m \log m)$ time, two numbers $r_-, r_+ \in U$, such that:*

(A) The number of rank k in U is in the interval $\mathcal{I}[r_-, r_+]$.

(B) There are at most $O\left(tn/\sqrt{m}\right)$ numbers of U in $\mathcal{I}$.

The algorithm succeeds with probability $\ge 1 - 3/t^3$.

Proof: Compute the sample in $O(m)$ time (assuming the input numbers are in an array, say. Next sort the numbers of R in $O(n \log n)$ time, and return the two elements of rank ℓ_- and ℓ_+ in the sorted set, as the boundaries of the interval. The correctness follows from the above discussion. ■

We next use the above observation to get a fast algorithm for selection.

5.1.2. Randomized selection

We are given a set S of n distinct elements, with an associated ordering. For $t \in S$, let $r_S(t)$ denote the rank of t (the smallest element in S has rank 1). Let $S_{(i)}$ denote the ith element in the sorted list of S.

Given k, we would like to compute S_k (i.e., select the kth element). The code of **LazySelect** is depicted in Figure 5.1.

Exercise 5.1.3. Show how to compute the ranks of $r_S(a)$ and $r_S(b)$, such that the expected number of comparisons performed is $1.5n$.

Consider the element $S_{(k)}$ and where it is mapped to in the random sample R. Consider the interval of values

$$\mathrm{I}(j) = \left[R_{(\alpha(j))}, R_{(\beta(j))}\right] = \left\{R_{(k)} \;\middle|\; \alpha(j) \le k \le \beta(j)\right\},$$

where $\alpha(j) = j \cdot n^{-1/4} - \sqrt{n}$ and $\beta(j) = j \cdot n^{-1/4} + \sqrt{n}$.

Lemma 5.1.4. *For a fixed j, we have that* $\mathbf{Pr}\left[S_{(j)} \in \mathrm{I}(j)\right] \geq 1 - 1/(4n^{1/4})$.

Proof: There are two possible bad events: (i) $S_{(j)} < R_{\alpha(j)}$ and (ii) $R_{\beta(j)} < S_{(j)}$. Let X_i be an indicator variable which is 1 if the ith sample is smaller equal to $S_{(j)}$, otherwise 0. We have $p = \mathbf{Pr}[X_i]] = j/n$ and $q = 1 - j/n$. The random variable $X = \sum_{i=1}^{n^{3/4}} X_i$ is the rank of $S_{(j)}$ in the random sample. Clearly, $X \sim B\left(3/4, j/n\right)$ (i.e., X has a binomial distribution with $p = j/n$, and $n^{3/4}$ trials). As such, we have $\mathbf{E}[X] = pn^{3/4}$ and $\mathbf{V}[X] = n^{3/4}pq$.

Now, by Chebyshev inequality

$$\mathbf{Pr}\left[\left|X - pn^{3/4}\right| \geq t\sqrt{n^{3/4}pq}\right] \leq \frac{1}{t^2}.$$

Since $pn^{3/4} = jn^{-1/4}$ and $\sqrt{n^{3/4}(j/n)(1-j/n)} \leq n^{3/8}/2$, we have that the probability of $a > S_{(j)}$ or $b > S_{(j)}$ is

$$\begin{aligned}
\mathbf{Pr}\left[S_{(j)} < R_{\alpha(j)} \text{ or } R_{\beta(j)} < S_{(j)}\right] &= \mathbf{Pr}\left[X < (jn^{-1/4} - \sqrt{n}) \quad \text{or} \quad X > (jn^{-1/4} + \sqrt{n})\right] \\
&= \mathbf{Pr}\left[|X - jn^{-1/4}| \geq 2n^{1/8} \cdot \frac{n^{3/8}}{2}\right] \\
&\leq \frac{1}{(2n^{1/8})^2} = \frac{1}{4n^{1/4}}.
\end{aligned}$$

■

Lemma 5.1.5. **LazySelect** *succeeds with probability* $\geq 1 - O(n^{-1/4})$ *in the first iteration. And it performs only* $2n + o(n)$ *comparisons.*

Proof: By Lemma 5.1.4, we know that $S_{(k)} \in \mathrm{I}(k)$ with probability $\geq 1 - 1/(4n^{1/4})$. This in turn implies that $S_{(k)} \in P$. Thus, the only possible bad event is that the set P is too large. To this end, set $k^- = k - 3n^{3/4}$ and $k^+ = k + 3n^{3/4}$, and observe that, by definition, it holds $\mathrm{I}(k^-) \cap \mathrm{I}(k) = \emptyset$ and $\mathrm{I}(k) \cap \mathrm{I}(k^+) = \emptyset$. As such, we know by Lemma 5.1.4, that $S_{(k^-)} \in \mathrm{I}(k^-)$ and $S_{(k^+)} \in \mathrm{I}(k^+)$, and this holds with probability $\geq 1 - \frac{2}{4n^{1/4}}$. As such, the set P, which is by definition contained in the range $\mathrm{I}(k)$, has only elements that are larger than $S_{(k^-)}$ and smaller than $S_{(k^+)}$. As such, the size of P is bounded by $k^+ - k^- = 6n^{3/4}$. Thus, the algorithm succeeds in the first iteration, with probability $\geq 1 - \frac{3}{4n^{1/4}}$.

As for the number of comparisons, an iteration requires

$$O(n^{3/4}\log n) + 2n + O(n^{3/4}\log n) = 2n + o(n)$$

comparisons ■

Any deterministic selection algorithm requires $2n$ comparisons, and **LazySelect** can be changed to require only $1.5n + o(n)$ comparisons (expected).

5.2. The Coupon Collector's Problem Revisited

5.2.1. Some technical lemmas

Unfortunately, in Randomized Algorithms, many of the calculations are awful[②]. As such, one has to be dexterous in approximating such calculations. We present quickly a few of these estimates.

[②] "In space travel," repeated Slartibartfast, "all the numbers are awful." – Life, the Universe, and Everything Else, Douglas Adams.

Lemma 5.2.1. *For $x \geq 0$, we have $1 - x \leq \exp(-x)$ and $1 + x \leq e^x$. Namely, for all x, we have $1 + x \leq e^x$.*

Proof: For $x = 0$ we have equality. Next, computing the derivative on both sides, we have that we need to prove that $-1 \leq -\exp(-x) \iff 1 \geq \exp(-x) \iff e^x \geq 1$, which clearly holds for $x \geq 0$.

A similar argument works for the second inequality. ■

Lemma 5.2.2. *For any $y \geq 1$, and $|x| \leq 1$, we have $\left(1 - x^2\right)^y \geq 1 - yx^2$.*

Proof: Observe that the inequality holds with equality for $x = 0$. So compute the derivative of x of both sides of the inequality. We need to prove that

$$y(-2x)\left(1 - x^2\right)^{y-1} \geq -2yx \iff \left(1 - x^2\right)^{y-1} \leq 1,$$

which holds since $1 - x^2 \leq 1$, and $y - 1 \geq 0$. ■

Lemma 5.2.3. *For any $y \geq 1$, and $|x| \leq 1$, we have $\left(1 - x^2 y\right)e^{xy} \leq (1 + x)^y \leq e^{xy}$.*

Proof: The right side of the inequality is standard by now. As for the left side. Observe that

$$(1 - x^2)e^x \leq 1 + x,$$

since dividing both sides by $(1 + x)e^x$, we get $1 - x \leq e^{-x}$, which we know holds for any x. By Lemma 5.2.2, we have

$$\left(1 - x^2 y\right)e^{xy} \leq \left(1 - x^2\right)^y e^{xy} = \left(\left(1 - x^2\right)e^x\right)^y \leq \left(1 + x\right)^y \leq e^{xy}.$$ ■

5.2.2. Back to the coupon collector's problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

In the previous lecture, we showed that

$$\mathbf{Pr}\left[\#\text{ of trials} \geq n \log n + n + t \cdot n\frac{\pi}{\sqrt{6}}\right] \leq \frac{1}{t^2},$$

for any t.

A stronger bound, follows from the following observation. Let Z_i^r denote the event that the ith coupon was not picked in the first r trials. Clearly,

$$\mathbf{Pr}\left[Z_i^r\right] = \left(1 - \frac{1}{n}\right)^r \leq \exp\left(-\frac{r}{n}\right).$$

Thus, for $r = \beta n \log n$, we have $\mathbf{Pr}\left[Z_i^r\right] \leq \exp\left(-\frac{\beta n \log n}{n}\right) = n^{-\beta}$. Thus,

$$\mathbf{Pr}\left[X > \beta n \log n\right] \leq \mathbf{Pr}\left[\bigcup_i Z_i^{\beta n \log n}\right] \leq n \cdot \mathbf{Pr}\left[Z_1\right] \leq n^{-\beta+1}.$$

Lemma 5.2.4. *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then, we have* $\mathbf{Pr}\Big[X > n\ln n + cn\Big] \leq e^{-c}$.

Proof: The probability we fail to pick the first type of coupon is $\alpha = (1-1/n)^m \leq \exp\Big(-\frac{n\ln n+cn}{n}\Big) = \exp(-c)/n$. As such, using the union bound, the probability we fail to pick all n types of coupons is bounded by $n\alpha = \exp(-c)$, as claimed. ■

In the following, we show a slightly stronger bound on the probability, which is $1-\exp(-e^{-c})$. To see that it is indeed stronger, observe that $e^{-c} \geq 1-\exp(-e^{-c})$.

5.2.3. An asymptotically tight bound

Lemma 5.2.5. *Let $c > 0$ be a constant, $m = n\ln n + cn$ for a positive integer n. Then for any constant k, we have* $\lim_{n\to\infty}\binom{n}{k}\Big(1-\frac{k}{n}\Big)^m = \frac{\exp(-ck)}{k!}$.

Proof: By Lemma 5.2.3, we have

$$\left(1-\frac{k^2m}{n^2}\right)\exp\left(-\frac{km}{n}\right) \leq \left(1-\frac{k}{n}\right)^m \leq \exp\left(-\frac{km}{n}\right).$$

Observe also that $\lim_{n\to\infty}\left(1-\frac{k^2m}{n^2}\right) = 1$, and $\exp\left(-\frac{km}{n}\right) = n^{-k}\exp(-ck)$. Also,

$$\lim_{n\to\infty}\binom{n}{k}\frac{k!}{n^k} = \lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{n^k} = 1.$$

Thus, $\lim_{n\to\infty}\binom{n}{k}\Big(1-\frac{k}{n}\Big)^m = \lim_{n\to\infty}\frac{n^k}{k!}\exp\left(-\frac{km}{n}\right) = \lim_{n\to\infty}\frac{n^k}{k!}n^{-k}\exp(-ck) = \frac{\exp(-ck)}{k!}$. ■

Theorem 5.2.6. *Let the random variable X denote the number of trials for collecting each of the n types of coupons. Then, for any constant $c \in \mathbb{R}$, and $m = n\ln n + cn$, we have* $\lim_{n\to\infty}\mathbf{Pr}\Big[X > m\Big] = 1-\exp\Big(-e^{-c}\Big)$.

Before dwelling into the proof, observe that $1-\exp(-e^{-c}) \approx 1-(1-e^{-c}) = e^{-c}$. Namely, in the limit, the upper bound of Lemma 5.2.4 is tight.

Proof: We have $\mathbf{Pr}\Big[X > m\Big] = \mathbf{Pr}\Big[\cup_i Z_i^m\Big]$. By inclusion-exclusion, we have

$$\mathbf{Pr}\left[\bigcup_i Z_i^m\right] = \sum_{i=1}^{n}(-1)^{i+1}P_i^n,$$

where $P_j^n = \sum_{1\leq i_1<i_2<\ldots<i_j\leq n}\mathbf{Pr}\left[\bigcap_{v=1}^{j}Z_{i_v}^m\right]$. Let $S_k^n = \sum_{i=1}^{k}(-1)^{i+1}P_i^n$. We know that $S_{2k}^n \leq \mathbf{Pr}\Big[\bigcup_i Z_i^m\Big] \leq S_{2k+1}^n$.

By symmetry,

$$P_k^n = \binom{n}{k}\mathbf{Pr}\left[\bigcap_{v=1}^{k}Z_v^m\right] = \binom{n}{k}\Big(1-\frac{k}{n}\Big)^m,$$

Thus, $P_k = \lim_{n\to\infty} P_k^n = \exp(-ck)/k!$, by Lemma 5.2.5. Thus, we have

$$S_k = \sum_{j=1}^{k} (-1)^{j+1} P_j = \sum_{j=1}^{k} (-1)^{j+1} \cdot \frac{\exp(-cj)}{j!}.$$

Observe that $\lim_{k\to\infty} S_k = 1 - \exp(-e^{-c})$ by the Taylor expansion of $\exp(x)$ (for $x = -e^{-c}$). Indeed,

$$\exp\Big(x\Big) = \sum_{j=0}^{\infty} \frac{x^j}{j!} = \sum_{j=0}^{\infty} \frac{(-e^{-c})^j}{j!} = 1 + \sum_{j=1}^{\infty} \frac{(-1)^j \exp(-cj)}{j!}.$$

Clearly, $\lim_{n\to\infty} S_k^n = S_k$ and $\lim_{k\to\infty} S_k = 1 - \exp(-e^{-c})$. Thus, (using fluffy math), we have

$$\lim_{n\to\infty} \mathbf{Pr}\Big[X > m\Big] = \lim_{n\to\infty} \mathbf{Pr}\Big[\cup_{i=1}^{n} Z_i^m\Big] = \lim_{n\to\infty} \lim_{k\to\infty} S_k^n = \lim_{k\to\infty} S_k = 1 - \exp(-e^{-c}).$$ ■

Chapter 6

Sampling and other Stuff

By Sariel Har-Peled, December 30, 2015[①]

6.1. Two-Point Sampling

Definition 6.1.1. A collection of random variables $X_1, \ldots, X_n$ is ***pairwise-independent***, if for any pair of variables X_i and X_j, and any pair of values α and β we have that $\mathbf{Pr}\Big[X_i = \alpha \cap X_j = \beta\Big] = \mathbf{Pr}\Big[X_i = \alpha\Big]\mathbf{Pr}\Big[X_j = \beta\Big]$.

Similarly, this collection is ***k-wise independent***, if for any $t \leq k$ variables $X_{i_1}, \ldots, X_{i_t}$ in this collection, and any set of t values, $\alpha_1, \ldots, \alpha'_t$ we have that

$$\mathbf{Pr}[(X_{i_1} = \alpha_1) \cap \ldots \cap (X_{i_t} = \alpha_t)] = \prod_{j=1}^{t} \mathbf{Pr}\Big[X_{i_j} = \alpha_j\Big].$$

Namely, pairwise independent variables behaves like independent random variables as long as you look only in pairs.

Example 6.1.2. Consider the probability space show on the right, where the triple of variables X, Y, Z can be assigned any of the rows with equal probability (i.e., 1/4).

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	0

Clearly, for any $\alpha, \beta \in \{0, 1\}$ we have $\mathbf{Pr}[(X = \alpha) \cap (Y = \beta)] = \mathbf{Pr}[(X = \alpha)]\,\mathbf{Pr}[(Y = \beta)] = 1/4$ (this also holds for X, Z and Y, Z). Namely, X, Y, Z are all pairwise independent. However, they are not 3-wise independent (or just independent). Indeed, we have $\mathbf{Pr}[(X = 1) \cap (Y = 1) \cap (Z = 1)] = 0$, while it should have been $1/8$ if they were truly independent, or even just 3-wise independent.

6.1.1. About Modulo Rings and Pairwise Independence

Let p be a prime number, and let $\mathbb{Z}_p = \Big\{0, 1, \ldots, p-1\Big\}$ denote the ring of integers modules p. Two integers x and y are ***equivalent modulo*** p, if $x \equiv y \mod p$; namely, the reminder of dividing x and y by p is the same.

Lemma 6.1.3. *Given $y, i \in \mathbb{Z}_p$, and choosing a and b randomly, independently and uniformly from $\mathbb{Z}_p$, the probability of $y \equiv ai + b \pmod p$ is $1/p$.*

Proof: Imagine that we first choose a, then the required probability, is that we choose b such that $y - ai \equiv b \pmod p$. And the probability for that is $1/p$, as we choose b uniformly. ■

Lemma 6.1.4. *Let p be a prime, and fix $a \in \{1, \ldots, p-1\}$. Then, $\left\{ai \pmod p \;\middle|\; i = 0, \ldots, p-1\right\} = \mathbb{Z}_p$.*

Putting it differently, for any non-zero $a \in \mathbb{Z}_p$, there is a unique inverse $b \in \mathbb{Z}_p$ such that $ab \pmod p = 1$.

Proof: Assume, for the sake of contradiction, that the claim is false. Then, by the pigeon hole principle, there must exists $1 \le j < i \le p-1$ such that $ai \pmod p = aj \pmod p$. Namely, there are k', k, u such that

$$ai = u + kp \quad \text{and} \quad aj = u + k'p.$$

(Here, we know that $0 \le k < p$, $0 \le k' < p$ and $0 \le u < p$.) Since $i > j$ it must be that $k > k'$. Subtracting the two equalities, we get that $a(i-j) = (k-k')p > 0$. Now, $i - j$ must be larger than one, since if $i - j = 1$ then $a = p$, which is impossible. Similarly, $i - j < p$. Also, $i - j$ can not divide p, since p is a prime. Thus, it must be that $i - j$ must divide $k - k'$. So, let us set $\beta = (k-k')/(i-j) \ge 1$. This implies that $a = \beta p \ge p$, which is impossible. Thus, our assumption is false. ■

Lemma 6.1.5. *Given $y, z, x, w \in \mathbb{Z}_p$, such that $x \neq w$, and choosing a and b randomly and uniformly from $\mathbb{Z}_p$, the probability that $y \equiv ax + b \pmod p$ and $z = aw + b \pmod p$ is $1/p^2$.*

Proof: This equivalent to claiming that the system of equalities $y \equiv ax + b \pmod p$ and $z = aw + b$ have a unique solution in a and b.

To see why this is true, subtract one equation from the other. We get $y - z \equiv a(x-w) \pmod p$. Since $x - w \not\equiv 0 \pmod p$, it must be that there is a unique value of a such that the equation holds. This in turns, imply a specific value for b. The probability that a and b get those two specific values is $1/p^2$. ■

Lemma 6.1.6. *Let i and j be two distinct elements of $\mathbb{Z}_p$. And choose a and b randomly and independently from $\mathbb{Z}_p$. Then, the two random variables $Y_i = ai + b \pmod p$ and $Y_j = aj + b \pmod p$ are uniformly distributed on $\mathbb{Z}_p$, and are pairwise independent.*

Proof: The claim about the uniform distribution follows from Lemma 6.1.3, as $\mathbf{Pr}[Y_i = \alpha] = 1/p$, for any $\alpha \in \mathbb{Z}_p$. As for being pairwise independent, observe that

$$\mathbf{Pr}\Big[Y_i = \alpha \;\Big|\; Y_j = \beta\Big] = \frac{\mathbf{Pr}\Big[Y_i = \alpha \cap Y_j = \beta\Big]}{\mathbf{Pr}\Big[Y_j = \beta\Big]} = \frac{1/n^2}{1/n} = \frac{1}{n} = \mathbf{Pr}\Big[Y_i = \alpha\Big],$$

by Lemma 6.1.3 and Lemma 6.1.5. Thus, Y_i and Y_j are pairwise independent. ■

Remark 6.1.7. It is important to understand what independence between random variables mean: having information about the value of X, gives you no information about Y. But this is only pairwise independence. Indeed, consider the variables Y_1, Y_2, Y_3, Y_4 defined above. Every pair of them are pairwise independent. But, given the values of Y_1 and Y_2, one can compute the value of Y_3 and Y_4 immediately. Indeed, giving the value of Y_1 and Y_2 is enough to figure out the value of a and b. Once we know a and b, we immediately can compute all the Y_is.

Thus, the notion of independence can be extended to k-pairwise independence of n random variables, where only if you know the value of k variables, you can compute the value of all the other variables. More on that later in the course.

Lemma 6.1.8. *If X and Y are pairwise independent then $\mathbf{E}\Big[XY\Big] = \mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big]$.*

Proof: By definition, $\mathbf{E}\Big[XY\Big] = \sum_{x,y} xy\,\mathbf{Pr}\Big[(X=x)\cap(Y=y)\Big] = \sum_{x,y} xy\,\mathbf{Pr}\Big[X=x\Big]\mathbf{Pr}\Big[Y=y\Big] = \sum_x x\,\mathbf{Pr}[X=x]$ $\sum_y y\,\mathbf{Pr}[Y=y] = (\sum_x x\,\mathbf{Pr}[X=x])\Big(\sum_y y\,\mathbf{Pr}[Y=y]\Big) = \mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big]$. ■

Lemma 6.1.9. *Let $X_1, X_2, \ldots, X_n$ be pairwise independent random variables, and $X = \sum_{i=1}^n X_i$. Then $\mathbf{V}\Big[X\Big] = \sum_{i=1}^n \mathbf{V}\Big[X_i\Big]$.*

Proof: Observe, that $\mathbf{V}\Big[X\Big] = \mathbf{E}\Big[(X-\mathbf{E}[X])^2\Big] = \mathbf{E}\Big[X^2\Big] - \Big(\mathbf{E}[X]\Big)^2$. Let X and Y be pairwise independent variables. Observe that $\mathbf{E}\Big[XY\Big] = \mathbf{E}[X]\,\mathbf{E}[Y]$, as can be easily verified. Thus,

$$\begin{aligned}
\mathbf{V}\Big[X+Y\Big] &= \mathbf{E}\Big[(X+Y-\mathbf{E}[X]-\mathbf{E}[Y])^2\Big] \\
&= \mathbf{E}\Big[\Big(X+Y\Big)^2 - 2\Big(X+Y\Big)\Big(\mathbf{E}[X]+\mathbf{E}[Y]\Big) + \Big(\mathbf{E}[X]+\mathbf{E}[Y]\Big)^2\Big] \\
&= \mathbf{E}\Big[(X+Y)^2\Big] - \Big(\mathbf{E}[X]+\mathbf{E}[Y]\Big)^2 \\
&= \mathbf{E}\Big[X^2+2XY+Y^2\Big] - \Big(\mathbf{E}[X]\Big)^2 - 2\,\mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big] - (\mathbf{E}[Y])^2 \\
&= \Big(\mathbf{E}\Big[X^2\Big] - \Big(\mathbf{E}[X]\Big)^2\Big) + \Big(\mathbf{E}\Big[Y^2\Big] - \Big(\mathbf{E}[Y]\Big)^2\Big) + 2\,\mathbf{E}\Big[XY\Big] - 2\,\mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big] \\
&= \mathbf{V}\Big[X\Big] + \mathbf{V}\Big[Y\Big] + 2\,\mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big] - 2\,\mathbf{E}\Big[X\Big]\,\mathbf{E}\Big[Y\Big] \\
&= \mathbf{V}\Big[X\Big] + \mathbf{V}\Big[Y\Big],
\end{aligned}$$

by Lemma 6.1.8. Using the above argumentation for several variables, instead of just two, implies the lemma. ■

6.1.1.1. Generating k-wise independent variable

Consider the polynomial $f(x) = \sum_{i=0}^{k-1} \alpha_i x^i$ evaluated modulo p, where the coefficients $\alpha_0, \ldots, \alpha_{k-1}$ are taken from $\mathbb{Z}_p$. We claim that $f(0), f(1), \ldots, f(p-1)$ are k-wise independent. Indeed, for any k indices $i_1, \ldots, i_k \in \mathbb{Z}_p$, and k values $v_1, \ldots, v_k \in \mathbb{Z}_p$, we have that $\beta = \mathbf{Pr}[f(i_1) = v_1 \text{ and } \ldots \text{ and } f(i_k) = v_k]$ happens only for one specific choice of the αs, which implies that this probability is $1/p^k$, which is what we need.

6.1.2. Application: Using less randomization for a randomized algorithm

We can consider a randomized algorithm, to be a deterministic algorithm $\mathbf{Alg}(x, r)$ that receives together with the input x, a random string r of bits, that it uses to read random bits from. Let us redefine **RP**:

Definition 6.1.10. The class **RP** (for Randomized Polynomial time) consists of all languages L that have a deterministic algorithm $\mathbf{Alg}(x, r)$ with worst case polynomial running time such that for any input $x \in \Sigma^*$,

- $x \in L \implies \mathbf{Alg}(x, r) = 1$ for half the possible values of r.
- $x \notin L \implies \mathbf{Alg}(x, r) = 0$ for all values of r.

Let assume that we now want to minimize the number of random bits we use in the execution of the algorithm (Why?). If we run the algorithm t times, we have confidence 2^{-t} in our result, while using $t \log n$ random bits (assuming our random algorithm needs only $\log n$ bits in each execution). Similarly, let us choose two random numbers from $\mathbb{Z}_n$, and run $\mathbf{Alg}(x, a)$ and $\mathbf{Alg}(x, b)$, gaining us only confidence $1/4$ in the correctness of our results, while requiring $2 \log n$ bits.

Can we do better? Let us define $r_i = ai + b \mod n$, where a, b are random values as above (note, that we assume that n is prime), for $i = 1, \ldots, t$. Thus $Y = \sum_{i=1}^{t} \mathbf{Alg}(x, r_i)$ is a sum of random variables which are pairwise independent, as the r_i are pairwise independent. Assume, that $x \in L$, then we have $\mathbf{E}[Y] = t/2$, and $\sigma_Y^2 = \mathbf{V}[Y] = \sum_{i=1}^{t} \mathbf{V}[\mathbf{Alg}(x, r_i)] \leq t/4$, and $\sigma_Y \leq \sqrt{t}/2$. The probability that all those executions failed, corresponds to the event that $Y = 0$, and

$$\mathbf{Pr}\Big[Y = 0\Big] \leq \mathbf{Pr}\Big[\big|Y - \mathbf{E}[Y]\big| \geq \frac{t}{2}\Big] = \mathbf{Pr}\Big[\big|Y - \mathbf{E}[Y]\big| \geq \frac{\sqrt{t}}{2} \cdot \sqrt{t}\Big] \leq \frac{1}{t},$$

by the Chebyshev inequality. Thus we were able to "extract" from our random bits, much more than one would naturally suspect is possible. We thus get the following result.

Lemma 6.1.11. *Given an algorithm* **Alg** *in* **RP** *that uses* $\lg n$ *random bits, one can run it* t *times, such that the runs results in a new algorithm that fails with probability at most* $1/t$.

6.2. QuickSort is quick via direct argumentation

Consider a specific element α in the input array of n elements that is being sorted by **QuickSort**, and let X_i be the size of the recursive subproblem in the ith level of the recursion that contains x. If x thus not participate in such a subproblem in this level, that $X_i = 0$. It is easy to verify that

$$X_0 = n \quad \text{and} \quad \mathbf{E}\Big[X_i \,\Big|\, X_{i-1}\Big] \leq \frac{1}{2} \cdot \frac{3}{4} X_{i-1} + \frac{1}{2} X_{i-1} \leq \frac{7}{8} X_{i-1}.$$

As such, $\mathbf{E}[X_i] = \mathbf{E}\Big[\mathbf{E}[X_i]\Big] = (7/8)^i n$. In particular, we have by Markov's inequality that

$$\mathbf{Pr}\left[\begin{array}{c}\alpha \text{ participates in more than} \\ c \ln n \text{ levels of the recursion}\end{array}\right] = \mathbf{Pr}[X_{c \ln n} \geq 1] \leq \frac{\mathbf{E}[X_{c \ln n}]}{1} \leq (7/8)^{c \ln n} n \leq \frac{1}{n^{\beta+1}},$$

if $\Big(c \ln(8/7)\Big) \ln n \geq \beta \ln n \iff c \geq \beta / \ln(8/7)$. We conclude the following.

Theorem 6.2.1. *For any* $\beta \geq 1$*, we have that the running time of* **QuickSort** *sorting* n *elements is* $O(\beta n \log n)$*, with probability* $\geq 1 - 1/n^{\beta}$.

Proof: For $c = \beta / \ln(8/7)$, the probability that an element participates in at most $c \ln n$ levels of the recursion is at most $1/n^{\beta+1}$. Since there are n elements, by the union bound, this bounds the probability that any input number would participate in more than $c \ln n$ recursive calls. But that implies that the recursion depth of **QuickSort** is $\leq c \ln n$, which immediately implies the claim. ∎

What the above proof shows is that an element can not be too unlucky – if it participates in enough rounds, then, with high probability, the subproblem containing it would shrink significantly. This fairness of luck is one of the most important principles in randomized algorithms, and we next formalize it by proving a rather general theorem on the "concentration" of luck.

Chapter 7

Concentration of Random Variables – Chernoff's Inequality

By Sariel Har-Peled, April 14, 2016[①]

7.1. Concentration of mass and Chernoff's inequality

7.1.1. Example: Binomial distribution

Consider the binomial distribution $\mathrm{Bin}(n, 1/2)$ for various values of n as depicted in Figure 7.1 – here we think about the value of the variable as the number of heads in flipping a fair coin n times. Clearly, as the value of n increases the probability of getting a number of heads that is significantly smaller or larger than $n/2$ is tiny. Here we are interested in quantifying exactly how far can we divert from this expected value. Specifically, if $X \sim \mathrm{Bin}(n, 1/2)$, then we would be interested in bounding the probability $\mathbf{Pr}[X > n/2 + \Delta]$, where $\Delta = t\sigma_X = t\sqrt{n}/2$ (i.e., we are t standard deviations away from the expectation). For $t > 2$, this probability is roughly 2^{-t}, which is what we prove here.

More surprisingly, if you look only on the middle of the distribution, it looks the same after clipping away the uninteresting tails, see Figure 7.2; that is, it looks more and more like the normal distribution. This is a universal phenomena known the *central limit theorem* – every sum of nicely behaved random variables behaves like the normal distribution. We unfortunately need a more precise quantification of this behavior, thus the following.

7.1.2. A restricted case of Chernoff inequality via games

7.1.2.1. Chernoff games

7.1.2.1.1. The game. Consider the game where a player starts with $Y_0 = 1$ dollars. At every round, the player can bet a certain amount x (fractions are fine). With probability half she loses her bet, and with probability half she gains an amount equal to her bet. The player is not allowed to go all in – because if she looses then the game is over. So it is natural to ask what her optimal betting strategy is, such that in the end of the game she has as much money as possible.

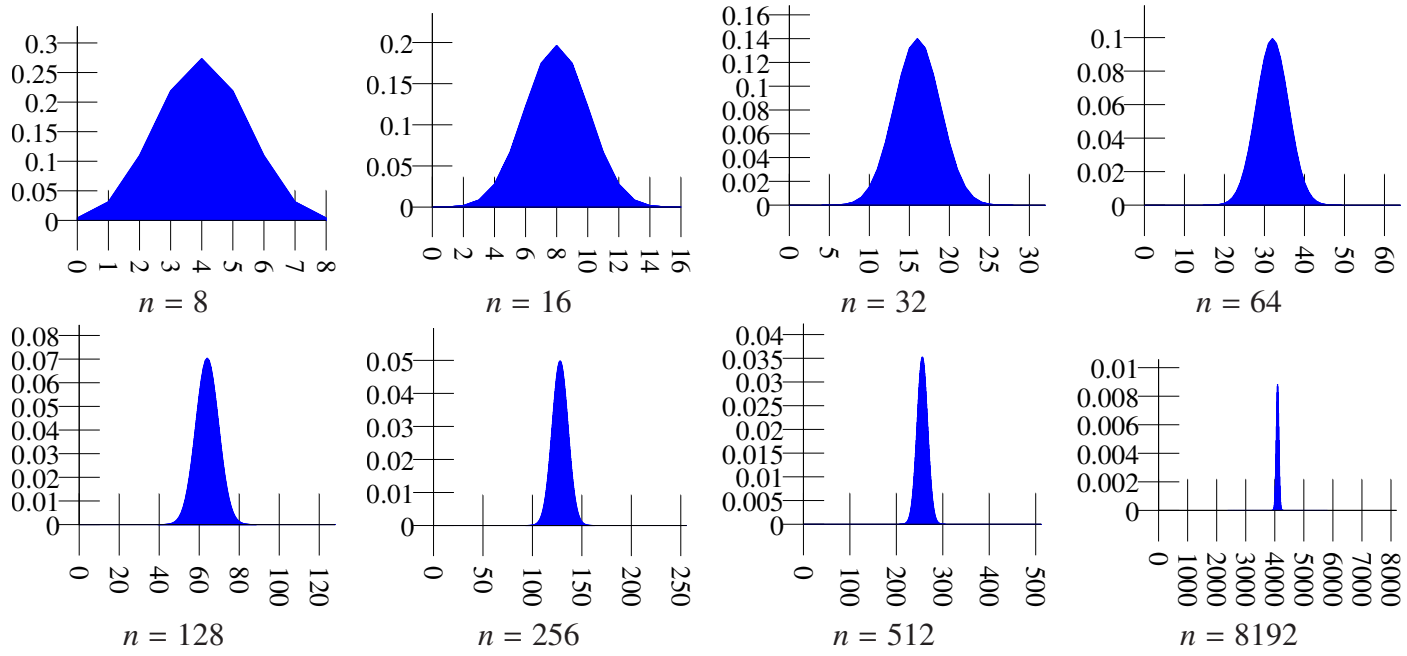


Figure 7.1: The binomial distribution for different values of n. It pretty quickly concentrates around its expectation.

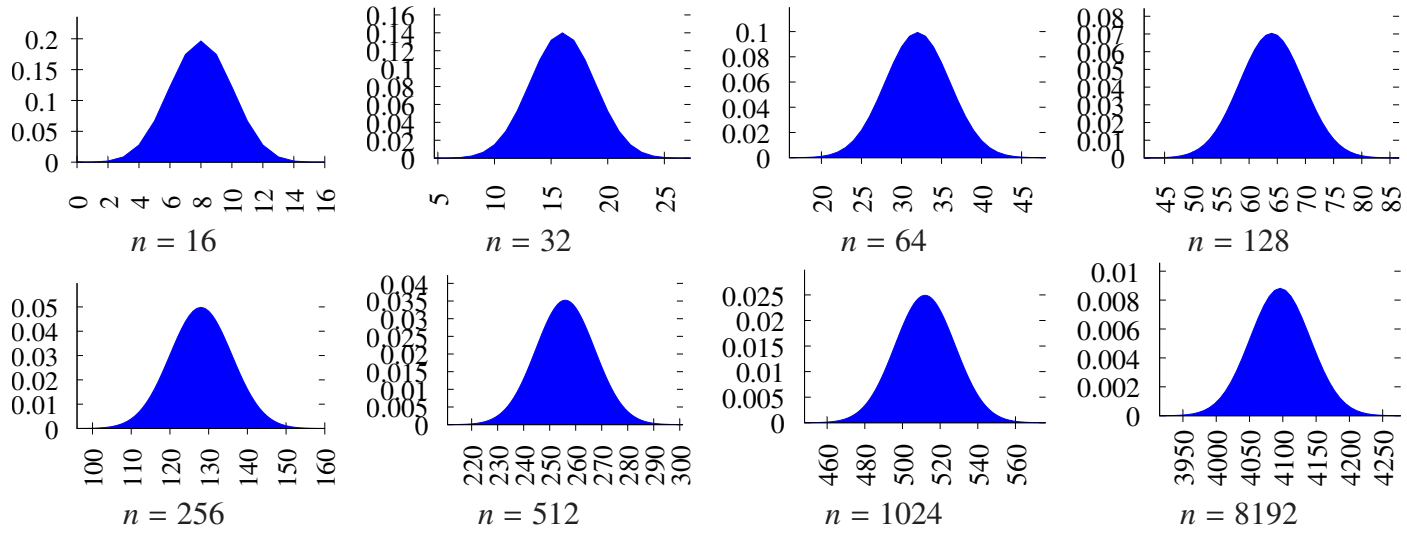


Figure 7.2: The "middle" of the binomial distribution for different values of n. It very quickly converges to the normal distribution (under appropriate rescaling and translation.

Values	Probabilities	Inequality	Ref
$-1, +1$	$\mathbf{Pr}[X_i = -1] =$ $\mathbf{Pr}[X_i = 1] = 1/2$	$\mathbf{Pr}\Big[Y \geq \Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big)$ $\mathbf{Pr}\Big[Y \leq -\Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big)$ $\mathbf{Pr}\Big[\lvert Y\rvert \geq \Delta\Big] \leq 2\exp\Big(-\Delta^2/2n\Big)$	Theorem 7.1.7$_{\text{p6}}$ Theorem 7.1.7$_{\text{p6}}$ Corollary 7.1.8$_{\text{p7}}$
$0, 1$	$\mathbf{Pr}[X_i = 0] =$ $\mathbf{Pr}[X_i = 1] = 1/2$	$\mathbf{Pr}\Big[\Big\lvert Y - \frac{n}{2}\Big\rvert \geq \Delta\Big] \leq 2\exp\Big(-2\Delta^2/n\Big)$	Corollary 7.1.9$_{\text{p7}}$
0,1	$\mathbf{Pr}[X_i = 0] = 1 - p_i$ $\mathbf{Pr}[X_i = 1] = p_i$	$\mathbf{Pr}\Big[Y > (1+\delta)\mu\Big] < \Big(\frac{e^\delta}{(1+\delta)^{1+\delta}}\Big)^\mu$	Theorem 7.3.2$_{\text{p12}}$
	For $\delta \leq 2e - 1$ $\delta \geq 2e - 1$ $\delta \geq e^2$	$\mathbf{Pr}[Y > (1+\delta)\mu] < \exp\Big(-\mu\delta^2/4\Big)$ $\mathbf{Pr}[Y > (1+\delta)\mu] < 2^{-\mu(1+\delta)}$ $\mathbf{Pr}[Y > (1+\delta)\mu] < \exp\Big(-(\mu\delta/2)\ln\delta\Big)$	Theorem 7.3.2$_{\text{p12}}$
	For $\delta \geq 0$	$\mathbf{Pr}[Y < (1-\delta)\mu] < \exp\Big(-\mu\delta^2/2\Big)$	Theorem 7.3.5$_{\text{p13}}$
$X_i \in [0, 1]$	X_is have arbitrary independent distributions.	$\mathbf{Pr}\Big[Y - \mu \geq \varepsilon\mu\Big] \leq \exp\Big(-\varepsilon^2\mu/4\Big)$ $\mathbf{Pr}\Big[Y - \mu \leq -\varepsilon\mu\Big] \leq \exp\Big(-\varepsilon^2\mu/2\Big).$	Theorem 7.4.5$_{\text{p15}}$
$X_i \in [a_i, b_i]$	X_is have arbitrary independent distributions.	$\mathbf{Pr}\Big[\Big\lvert Y - \mu\Big\rvert \geq \eta\Big] \leq 2\exp\Big(-\frac{2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$	Theorem 7.5.3$_{\text{p18}}$

Table 7.1: Summary of Chernoff type inequalities covered. Here we have n independent random variables $X_1, \ldots, X_n$, $Y = \sum_i X_i$ and $\mu = \mathbf{E}[Y]$.

7.1.2.1.2. Is the game pointless? So, let Y_{i-1} be the money the player has in the end of the $(i-1)$th round, and she bets an amount $\psi_i \le Y_{i-1}$ in the ith round. As such, in the end of the ith round, she has

$$Y_i = \begin{cases} Y_{i-1} - \psi_i & \text{LOSE: probability half} \\ Y_{i-1} + \psi_i & \text{WIN: probability half} \end{cases}$$

dollars. This game, in expectation, does not change the amount of money the player has. Indeed, we have

$$\mathbf{E}\Big[Y_i \;\Big|\; Y_{i-1}\Big] = \frac{1}{2}(Y_{i-1} - \psi_i) + \frac{1}{2}(Y_{i-1} + \psi_i) = Y_{i-1}.$$

And as such, we have that $\mathbf{E}[Y_i] = \mathbf{E}\Big[\mathbf{E}\Big[Y_i \Big| Y_{i-1}\Big]\Big] = \mathbf{E}[Y_{i-1}] = \cdots = \mathbf{E}[Y_0] = 1$. In particular, $\mathbf{E}[Y_n] = 1$ – namely, on average, independent of the player strategy she is not going to make any money in this game (and she is allowed to change her bets after every round). Unless, she is lucky[②]...

7.1.2.1.3. What about a lucky player? The player believes she will get lucky and wants to develop a strategy to take advantage of it. Formally, she believes that she can win, say, at least $(1+\delta)/2$ fraction of her bets (instead of the predicted $1/2$) – for example, if the bets are in the stock market, she can improve her chances by doing more research on the companies she is investing in[③]. Unfortunately, the player does not know which rounds she is going to be lucky in – so she still needs to be careful.

7.1.2.1.4. In a search of a good strategy. Of course, there are many safe strategies the player can use, from not playing at all, to risking only a tiny fraction of her money at each round. In other words, our quest here is to find the best strategy that extracts the maximum benefit for the player out of her inherent luck.

Here, we restrict ourselves to a simple strategy – at every round, the player would bet β fraction of her money, where β is a parameter to be determined. Specifically, in the end of the ith round, the player would have

$$Y_i = \begin{cases} (1-\beta)Y_{i-1} & \text{LOSE} \\ (1+\beta)Y_{i-1} & \text{WIN.} \end{cases}$$

By our assumption, the player is going to win in at least $M = (1+\delta)n/2$ rounds. Our purpose here is to figure out what the value of β should be so that player gets as rich as possible[④]. Now, if the player is successful in $\ge M$ rounds, out of the n rounds of the game, then the amount of money the player has, in the end of the game, is

$$\begin{aligned} Y_n &\ge (1-\beta)^{n-M}(1+\beta)^M = (1-\beta)^{n/2-(\delta/2)n}(1+\beta)^{n/2+(\delta/2)n} = \Big((1-\beta)(1+\beta)\Big)^{n/2-(\delta/2)n}(1+\beta)^{\delta n} \\ &= \Big(1-\beta^2\Big)^{n/2-(\delta/2)n}(1+\beta)^{\delta n} \ge \exp\Big(-2\beta^2\Big)^{n/2-(\delta/2)n}\exp(\beta/2)^{\delta n} = \exp\Big(\Big(-\beta^2+\beta^2\delta+\beta\delta/2\Big)n\Big). \end{aligned}$$

To maximize this quantity, we choose $\beta = \delta/4$ (there is a better choice, see Lemma 7.1.6, but we use this value for the simplicity of exposition). Thus, we have that $Y_n \ge \exp\left(\left(-\frac{\delta^2}{16} + \frac{\delta^3}{16} + \frac{\delta^2}{8}\right)n\right) \ge \exp\left(\frac{\delta^2}{16}n\right)$, proving the following.

Lemma 7.1.1. *Consider a Chernoff game with n rounds, starting with one dollar, where the player wins in $\ge (1+\delta)n/2$ of the rounds. If the player bets $\delta/4$ fraction of her current money, at all rounds, then in the end of the game the player would have at least $\exp\big(n\delta^2/16\big)$ dollars.*

[②] "I would rather have a general who was lucky than one who was good." – Napoleon Bonaparte.

[③] "I am a great believer in luck, and I find the harder I work, the more I have of it." – Thomas Jefferson.

[④] This optimal choice is known as Kelly criterion, see Remark 7.1.3.

Remark 7.1.2. Note, that Lemma 7.1.1 holds if the player wins any $\geq (1+\delta)n/2$ rounds. In particular, the statement does not require randomness by itself – for our application, however, it is more natural and interesting to think about the player wins as being randomly distributed.

Remark 7.1.3. Interestingly, the idea of choosing the best fraction to bet is an old and natural question arising in investments strategies, and the right fraction to use is known as *Kelly criterion*, going back to Kelly's work from 1956 [Kel56].

7.1.2.2. Chernoff's inequality

The above implies that if a player is lucky, then she is going to become filthy rich[5]. Intuitively, this should be a pretty rare event – because if the player is rich, then (on average) many other people have to be poor. We are thus ready for the kill.

Theorem 7.1.4 (Chernoff's inequality). *Let $X_1, \ldots, X_n$ be n independent random variables, where $X_i = 0$ or $X_i = 1$ with equal probability. Then, for any $\delta \in (0, 1/2)$, we have that*

$$\mathbf{Pr}\left[\sum_i X_i \geq (1+\delta)\frac{n}{2}\right] \leq \exp\left(-\frac{\delta^2}{16}n\right).$$

Proof: Imagine that we are playing the Chernoff game above, with $\beta = \delta/4$, starting with 1 dollar, and let Y_i be the amount of money in the end of the ith round. Here $X_i = 1$ indicates that the player won the ith round. We have, by Lemma 7.1.1 and Markov's inequality, that

$$\mathbf{Pr}\left[\sum_i X_i \geq (1+\delta)\frac{n}{2}\right] \leq \mathbf{Pr}\left[Y_n \geq \exp\left(\frac{n\delta^2}{16}\right)\right] \leq \frac{\mathbf{E}[Y_n]}{\exp(n\delta^2/16)} = \frac{1}{\exp(n\delta^2/16)} = \exp\left(-\frac{\delta^2}{16}n\right),$$

as claimed. ■

7.1.2.2.1. This is crazy – so intuition maybe? If the player is $(1+\delta)/2$-lucky then she can make a lot of money; specifically, at least $f(\delta) = \exp\left(n\delta^2/16\right)$ dollars by the end of the game. Namely, beating the odds has significant monetary value, and this value grows quickly with δ. Since we are in a "zero-sum" game settings, this event should be very rare indeed. Under this interpretation, of course, the player needs to know in advance the value of δ – so imagine that she guesses it somehow in advance, or she plays the game in parallel with all the possible values of δ, and she settles on the instance that maximizes her profit.

7.1.2.2.2. Can one do better? No, not really. Chernoff inequality is tight (this is a challenging homework exercise) up to the constant in the exponent. The best bound I know for this version of the inequality has $1/2$ instead of $1/16$ in the exponent. Note, however, that no real effort was taken to optimize the constants – this is not the purpose of this write-up.

7.1.2.3. Some low level boring calculations

Above, we used the following well known facts.

[5] Not that there is anything wrong with that – many of my friends are filthy,

Lemma 7.1.5. *(A) Markov's inequality. For any positive random variable X and $t > 0$, we have* $\mathbf{Pr}[X \geq t] \leq \mathbf{E}[X]/t$.

(B) For any two random variables X and Y, we have that $\mathbf{E}\Big[X\Big] = \mathbf{E}\Big[\mathbf{E}\Big[X\,\Big|Y\Big]\Big]$.

(C) For $x \in (0,1)$, $1 + x \geq e^{x/2}$.

(D) For $x \in (0, 1/2)$, $1 - x \geq e^{-2x}$.

Lemma 7.1.6. *The quantity* $\exp\Big(\Big(-\beta^2 + \beta^2\delta + \beta\delta/2\Big)n\Big)$ *is maximal for* $\beta = \frac{\delta}{4(1-\delta)}$.

Proof: We have to maximize $f(\beta) = -\beta^2 + \beta^2\delta + \beta\delta/2$ by choosing the correct value of β (as a function of δ, naturally). $f'(\beta) = -2\beta + 2\beta\delta + \delta/2 = 0 \quad\Longleftrightarrow\quad 2(\delta-1)\beta = -\delta/2 \quad\Longleftrightarrow\quad \beta = \frac{\delta}{4(1-\delta)}$. ■

7.1.3. Chernoff Inequality - A Special Case – the classical proof

Theorem 7.1.7. *Let $X_1, \ldots, X_n$ be n independent random variables, such that* $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = \frac{1}{2}$, *for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}\Big[Y \geq \Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big).$$

Proof: Clearly, for an arbitrary t, to specified shortly, we have

$$\mathbf{Pr}[Y \geq \Delta] = \mathbf{Pr}[\exp(tY) \geq \exp(t\Delta)] \leq \frac{\mathbf{E}[\exp(tY)]}{\exp(t\Delta)},$$

the first part follows by the fact that $\exp(\cdot)$ preserve ordering, and the second part follows by the Markov inequality.

Observe that

$$\begin{aligned}
\mathbf{E}[\exp(tX_i)] &= \frac{1}{2}e^{t} + \frac{1}{2}e^{-t} = \frac{e^t + e^{-t}}{2} \\
&= \frac{1}{2}\left(1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) \\
&+ \frac{1}{2}\left(1 - \frac{t}{1!} + \frac{t^2}{2!} - \frac{t^3}{3!} + \cdots\right) \\
&= \left(1 + \quad + \frac{t^2}{2!} + \quad + \cdots + \frac{t^{2k}}{(2k)!} + \cdots\right),
\end{aligned}$$

by the Taylor expansion of $\exp(\cdot)$. Note, that $(2k)! \geq (k!)2^k$, and thus

$$\mathbf{E}[\exp(tX_i)] = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{t^{2i}}{2^i(i!)} = \sum_{i=0}^{\infty} \frac{1}{i!}\left(\frac{t^2}{2}\right)^i = \exp\Big(t^2/2\Big),$$

again, by the Taylor expansion of $\exp(\cdot)$. Next, by the independence of the X_is, we have

$$\mathbf{E}[\exp(tY)] = \mathbf{E}\left[\exp\left(\sum_i tX_i\right)\right] = \mathbf{E}\left[\prod_i \exp(tX_i)\right] = \prod_{i=1}^{n} \mathbf{E}[\exp(tX_i)] \leq \prod_{i=1}^{n} e^{t^2/2} = e^{nt^2/2}.$$

We have $\mathbf{Pr}[Y \geq \Delta] \leq \dfrac{\exp\Big(nt^2/2\Big)}{\exp(t\Delta)} = \exp\Big(nt^2/2 - t\Delta\Big)$.

Next, by minimizing the above quantity for t, we set $t = \Delta/n$. We conclude,

$$\mathbf{Pr}[Y \geq \Delta] \leq \exp\left(\frac{n}{2}\left(\frac{\Delta}{n}\right)^2 - \frac{\Delta}{n}\Delta\right) = \exp\left(-\frac{\Delta^2}{2n}\right).$$ ■

By the symmetry of Y, we get the following:

Corollary 7.1.8. *Let $X_1, \ldots, X_n$ be n independent random variables, such that $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = \frac{1}{2}$, for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}[|Y| \geq \Delta] \leq 2e^{-\Delta^2/2n}.$$

Corollary 7.1.9. *Let $X_1, \ldots, X_n$ be n independent* coin flips*, such that $\mathbf{Pr}[X_i = 0] = \mathbf{Pr}[X_i = 1] = \frac{1}{2}$, for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}\left[\left|Y - \frac{n}{2}\right| \geq \Delta\right] \leq 2e^{-2\Delta^2/n}.$$

Remark 7.1.10. Before going any further, it is might be instrumental to understand what this inequalities imply. Consider then case where X_i is either zero or one with probability half. In this case $\mu = \mathbf{E}[Y] = n/2$. Set $\delta = t\sqrt{n}$ ($\sqrt{\mu}$ is approximately the standard deviation of X if $p_i = 1/2$). We have by

$$\mathbf{Pr}\left[\left|Y - \frac{n}{2}\right| \geq \Delta\right] \leq 2\exp\left(-2\Delta^2/n\right) = 2\exp\left(-2(t\sqrt{n})^2/n\right) = 2\exp\left(-2t^2\right).$$

Thus, Chernoff inequality implies exponential decay (i.e., $\leq 2^{-t}$) with t standard deviations, instead of just polynomial (i.e., $\leq 1/t^2$) by the Chebychev's inequality.

7.2. Applications of Chernoff's inequality

There is a zoo of Chernoff type inequalities, and prove some of them later on the chapter – while being very useful and technically interesting, they tend to numb to reader into boredom and submission. As such, we discuss applications of Chernoff's inequality here, and the interested reader can read the proofs of the more general forms only if they are interested in them.

7.2.1. QuickSort is Quick

We revisit **QuickSort**. We remind the reader that the running time of **QuickSort** is proportional to the number of comparisons performed by the algorithm. Next, consider an arbitrary element u being sorted. Consider the ith level recursive subproblem that contains u, and let S_i be the set of elements in this subproblems. We consider u to be *successful* in the ith level, if $|S_{i+1}| \leq |S_i|/2$. Namely, if u is successful, then the next level in the recursion involving u would include a considerably smaller subproblem. Let X_i be the indicator variable which is 1 if u is successful.

We first observe that if **QuickSort** is applied to an array with n elements, then u can be successful at most $T = \lceil \lg n \rceil$ times, before the subproblem it participates in is of size one, and the recursion stops. Thus, consider the indicator variable X_i which is 1 if u is successful in the ith level, and zero otherwise. Note that the X_is are independent, and $\mathbf{Pr}[X_i = 1] = 1/2$.

If u participates in v levels, then we have the random variables $X_1, X_2, \ldots, X_v$. To make things simpler, we will extend this series by adding independent random variables, such that $\mathbf{Pr}[`]X_i = 1 = 1/2$, for $i \geq v$. Thus, we have an infinite sequence of independent random variables, that are 0/1 and get 1 with probability $1/2$. The question is how many elements in the sequence we need to read, till we get T ones.

Lemma 7.2.1. *Let $X_1, X_2, \ldots$ be an infinite sequence of independent random 0/1 variables. Let M be an arbitrary parameter. Then the probability that we need to read more than $2M + 4t\sqrt{M}$ variables of this sequence till we collect M ones is at most $2\exp\left(-t^2\right)$, for $t \leq \sqrt{M}$. If $t \geq \sqrt{M}$ then this probability is at most $2\exp\left(-t\sqrt{M}\right)$.*

Proof: Consider the random variable $Y = \sum_{i=1}^{L} X_i$, where $L = 2M + 4t\sqrt{M}$. Its expectation is $L/2$, and using the Chernoff inequality, we get

$$\begin{aligned}\alpha = \mathbf{Pr}\left[Y \leq M\right] &\leq \mathbf{Pr}\left[\left|Y - \frac{L}{2}\right| \geq \frac{L}{2} - M\right] \leq 2\exp\left(-\frac{2}{L}\left(\frac{L}{2} - M\right)^2\right) \\ &\leq 2\exp\left(-\frac{2}{L}\left(M + 2t\sqrt{M} - M\right)^2\right) \leq 2\exp\left(-\frac{2}{L}\left(2t\sqrt{M}\right)^2\right) = 2\exp\left(-\frac{8t^2M}{L}\right),\end{aligned}$$

by Corollary 7.1.9. For $t \leq \sqrt{M}$ we have that $L = 2M + 4t\sqrt{M} \leq 8M$, as such in this case $\mathbf{Pr}[Y \leq M] \leq 2\exp\left(-t^2\right)$.

If $t \geq \sqrt{M}$, then $\alpha = 2\exp\left(-\dfrac{8t^2M}{2M + 4t\sqrt{M}}\right) \leq 2\exp\left(-\dfrac{8t^2M}{6t\sqrt{M}}\right) \leq 2\exp\left(-t\sqrt{M}\right)$. ■

Going back to the **QuickSort** problem, we have that if we sort n elements, the probability that u will participate in more than $L = (4 + c)\lceil \lg n \rceil = 2\lceil \lg n \rceil + 4c\sqrt{\lg n}\sqrt{\lg n}$, is smaller than $2\exp\left(-c\sqrt{\lg n}\sqrt{\lg n}\right) \leq 1/n^c$, by Lemma 7.2.1. There are n elements being sorted, and as such the probability that any element would participate in more than $(4 + c + 1)\lceil \lg n \rceil$ recursive calls is smaller than $1/n^c$.

Lemma 7.2.2. *For any $c > 0$, the probability that* **QuickSort** *performs more than $(6 + c)n \lg n$, is smaller than $1/n^c$.*

7.2.2. How many times can the minimum change?

Let $\Pi = \pi_1 \ldots \pi_n$ be a random permutation of $\{1, \ldots, n\}$. Let $\mathcal{E}_i$ be the event that π_i is the minimum number seen so far as we read Π; that is, $\mathcal{E}_i$ is the event that $\pi_i = \min_{k=1}^{i} \pi_k$. Let X_i be the indicator variable that is one if $\mathcal{E}_i$ happens. We already seen, and it is easy to verify, that $\mathbf{E}[X_i] = 1/i$. We are interested in how many times the minimum might change[6]; that is $Z = \sum_i X_i$, and how concentrated is the distribution of Z. The following is maybe surprising.

Lemma 7.2.3. *The events $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent (as such, variables $X_1, \ldots, X_n$ are independent).*

Proof: The trick is to think about the sampling process in a different way, and then the result readily follows. Indeed, we randomly pick a permutation of the given numbers, and set the first number to be π_n. We then, again, pick a random permutation of the remaining numbers and set the first number as the penultimate number (i.e., π_{n-1}) in the output permutation. We repeat this process till we generate the whole permutation.

Now, consider $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, and observe that $\mathbf{Pr}\left[\mathcal{E}_{i_k} \,\middle|\, \mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] = \mathbf{Pr}[\mathcal{E}_{i_k}]$, since by our thought experiment, $\mathcal{E}_{i_k}$ is determined before all the other variables $\mathcal{E}_{i_{k-1}}, \ldots, \mathcal{E}_{i_1}$, and these variables are inherently not effected by this event happening or not. As such, we have

$$\begin{aligned}\mathbf{Pr}\left[\mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \cap \ldots \cap \mathcal{E}_{i_k}\right] &= \mathbf{Pr}\left[\mathcal{E}_{i_k} \,\middle|\, \mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] \mathbf{Pr}[\mathcal{E}_{i_1} \cap \ldots \cap \mathcal{E}_{i_{k-1}}] \\ &= \mathbf{Pr}\left[\mathcal{E}_{i_k}\right] \mathbf{Pr}\left[\mathcal{E}_{i_1} \cap \mathcal{E}_{i_2} \cap \ldots \cap \mathcal{E}_{i_{k-1}}\right] = \prod_{j=1}^{k} \mathbf{Pr}\left[\mathcal{E}_{i_j}\right] = \prod_{j=1}^{k} \frac{1}{i_j},\end{aligned}$$

by induction. ■

[6] The answer, my friend, is blowing in the permutation.

Theorem 7.2.4. *Let $\Pi = \pi_1 \dots \pi_n$ be a random permutation of $1, \dots, n$, and let Z be the number of times, that π_i is the smallest number among $\pi_1, \dots, \pi_i$, for $i = 1, \dots, n$. Then, we have that for $t \geq 2e$ that $\mathbf{Pr}[Z > t \ln n] \leq 1/n^{t \ln 2}$, and for $t \in [1, 2e]$, we have that $\mathbf{Pr}[Z > t \ln n] \leq 1/n^{(t-1)^2/4}$.*

Proof: Follows readily from Chernoff's inequality, as $Z = \sum_i X_i$ is a sum of independent indicator variables, and, since by linearity of expectations, we have

$$\mu = \mathbf{E}\big[Z\big] = \sum_i \mathbf{E}\big[X_i\big] = \sum_{i=1}^{n} \frac{1}{i} \geq \int_{x=1}^{n+1} \frac{1}{x} \mathrm{d}x = \ln(n+1) \geq \ln n.$$

Next, we set $\delta = t - 1$, and use Theorem 7.3.2$_{\text{p12}}$. ■

7.2.3. Routing in a Parallel Computer

Let G be a graph of a network, where every node is a processor. The processor communicate by sending packets on the edges. Let $\big[0, \dots, N-1\big]$ denote be vertices (i.e., processors) of G, where $N = 2^n$, and G is the hypercube. As such, each processes is identified with a binary string $b_1 b_2 \dots b_n \in \{0,1\}^n$. Two nodes are connected if their binary string differs only in a single bit. Namely, G is the binary *hypercube* over n bits.

We want to investigate the best routing strategy for this topology of network. We assume that every processor need to send a message to a single other processor. This is represented by a permutation π, and we would like to figure out how to send the messages encoded by the permutation while create minimum delay/congestion.

Specifically, in our model, every edge has a FIFO queue[7] of the packets it has to transmit. At every clock tick, one message get sent. All the processors start sending the packets in their permutation in the same time.

A routing scheme is *oblivious* if every node that has to forward a packet, inspect the packet, and depending only on the content of the packet decides how to forward it. That is, such a routing scheme is local in nature, and does not take into account other considerations. Oblivious routing is of course a bad idea – it ignores congestion in the network, and might insist routing things through regions of the hypercube that are "gridlocked".

Theorem 7.2.5 ([KKT91]). *For any deterministic oblivious permutation routing algorithm on a network of N nodes each of out-degree n, there is a permutation for which the routing of the permutation takes $\Omega\Big(\sqrt{N/n}\Big)$ units of time (i.e., ticks).*

Proof: (Sketch.) The above is implied by a nice averaging argument – construct, for every possible destination, the routing tree of all packets to this specific node. Argue that there must be many edges in this tree that are highly congested in this tree (which is NOT the permutation routing we are looking for!). Now, by averaging, there must be a single edge that is congested in "many" of these trees. Pick a source-destination pair from each one of these trees that uses this edge, and complete it into a full permutation in the natural way. Clearly, the congestion of the resulting permutation is high. For the exact details see [KKT91]. ■

7.2.3.0.1. How do we send a packet? We use *bit fixing*. Namely, the packet from the i node, always go to the current adjacent node that have the first different bit as we scan the destination string $d(i)$. For example, packet from (0000) going to (1101), would pass through (1000), (1100), (1101).

7.2.3.0.2. The routing algorithm. We assume each edge have a FIFO queue. The routing algorithm is depicted in Figure 7.3.

[7] First in, first out queue. I sure hope you already knew that.

RandomRoute($v_0, \ldots, v_{N-1}$)

// v_i: Packet at node i to be routed to node $d(i)$.

(i) Pick a *random* intermediate destination $\sigma(i)$ from $[1, \ldots, N]$. Packet v_i travels to $\sigma(i)$.

// Here random sampling is done with ***replacement***.

// Several packets might travel to the same destination.

(ii) Wait till all the packets arrive to their intermediate destination.

(iii) Packet v_i travels from $\sigma(i)$ to its destination $d(i)$.

Figure 7.3: The routing algorithm

7.2.3.1. Analysis

We analyze only (i) as (iii) follows from the same analysis. In the following, let ρ_i denote the route taken by v_i in (i).

Exercise 7.2.6. Once a packet v_j that travel along a path ρ_j can not leave a path ρ_i, and then join it again later. Namely, $\rho_i \cap \rho_j$ is (maybe an empty) path.

Lemma 7.2.7. *Let the route of a message* **c** *follow the sequence of edges* $\pi = (e_1, e_2, \ldots, e_k)$. *Let* S *be the set of packets whose routes pass through at least one of* $(e_1, \ldots, e_k)$. *Then, the delay incurred by* **c** *is at most* $|S|$.

Proof: A packet in S is said to leave π at that time step at which it traverses an edge of π for the last time. If a packet is ready to follow edge e_j at time t, we define its *lag* at time t to be $t - j$. The lag of **c** is initially zero, and the delay incurred by **c** is its lag when it traverse e_k. We will show that each step at which the lag of **c** increases by one can be charged to a distinct member of S.

We argue that if the lag of **c** reaches $\ell + 1$, some packet in S leaves π with lag ℓ. When the lag of **c** increases from ℓ to $\ell + 1$, there must be at least one packet (from S) that wishes to traverse the same edge as **c** at that time step, since otherwise **c** would be permitted to traverse this edge and its lag would not increase. Thus, S contains at least one packet whose lag reach the value ℓ.

Let τ be the last time step at which any packet in S has lag ℓ. Thus there is a packet **d** ready to follow edge e_μ at τ, such that $\tau - \mu = \ell$. We argue that some packet of S leaves π at τ; this establishes the lemma since once a packet leaves π, it would never join it again and as such will never again delay **c**.

Since **d** is ready to follow e_μ at τ, some packet ω (which may be **d** itself) in S follows e_μ at time τ. Now ω leaves π at time τ; if not, some packet will follow $e_{\mu+1}$ at step $\mu + 1$ with lag still at ℓ, violating the maximality of τ. We charge to ω the increase in the lag of **c** from ℓ to $\ell + 1$; since ω leaves π, it will never be charged again. Thus, each member of S whose route intersects π is charge for at most one delay, establishing the lemma. ■

Let H_{ij} be an indicator variable that is 1 if ρ_i and ρ_j share an edge, and 0 otherwise. The total delay for v_i is at most $\leq \sum_j H_{ij}$.

Crucially, for a fixed i, the variables $H_{i1}, \ldots, H_{iN}$ are independent. Indeed, imagine first picking the destination of v_i, and let the associated path be ρ_i. Now, pick the destinations of all the other packets in the network. Since the sampling of destinations is done with replacements, whether or not, the path of v_j intersects ρ_i or not, is independent of whether v_k intersects ρ_i. Of course, the probabilities $\mathbf{Pr}\left[H_{ij} = 1\right]$ and $\mathbf{Pr}\left[H_{ik} = 1\right]$ are probably different. Confusingly, however, $H_{11}, \ldots, H_{NN}$ are not independent. Indeed, imagine k and j being close vertices on the hypercube. If $H_{ij} = 1$ then intuitively it means that ρ_i is traveling close to the vertex v_j, and as such there is a higher probability that $H_{ik} = 1$.

Let $\rho_i = (e_1, \ldots, e_k)$, and let $T(e)$ be the number of packets (i.e., paths) that pass through e. We have that

$$\sum_{j=1}^{N} H_{ij} \leq \sum_{j=1}^{k} T(e_j) \quad \text{and thus} \quad \mathbf{E}\left[\sum_{j=1}^{N} H_{ij}\right] \leq \mathbf{E}\left[\sum_{j=1}^{k} T(e_j)\right].$$

Because of symmetry, the variables $T(e)$ have the same distribution for all the edges of G. On the other hand, the expected length of a path is $n/2$, there are N packets, and there are $Nn/2$ edges. We conclude $E[T(e)] = 1$. Thus

$$\mu = \mathbf{E}\left[\sum_{j=1}^{N} H_{ij}\right] \leq \mathbf{E}\left[\sum_{j=1}^{k} T(e_j)\right] = \mathbf{E}\Big[|\rho_i|\Big] \leq \frac{n}{2}.$$

By the Chernoff inequality, we have

$$\mathbf{Pr}\left[\sum_{j} H_{ij} > 7n\right] \leq \mathbf{Pr}\left[\sum_{j} H_{ij} > (1+13)\mu\right] < 2^{-13\mu} \leq 2^{-6n}.$$

Since there are $N = 2^n$ packets, we know that with probability $\leq 2^{-5n}$ all packets arrive to their temporary destination in a delay of most $7n$.

Theorem 7.2.8. *Each packet arrives to its destination in $\leq 14n$ stages, in probability at least $1 - 1/N$ (note that this is very conservative).*

7.2.4. Faraway Strings

Consider the Hamming distance between binary strings. It is natural to ask how many strings of length n can one have, such that any pair of them, is of Hamming distance at least t from each other. Consider two random strings, generated by picking at each bit randomly and independently. Thus, $\mathbf{E}[d_H(x,y)] = n/2$, where $d_H(x,y)$ denote the hamming distance between x and y. In particular, using the Chernoff inequality, we have that

$$\mathbf{Pr}[d_H(x,y) \leq n/2 - \Delta] \leq \exp\Big(-2\Delta^2/n\Big).$$

Next, consider generating M such string, where the value of M would be determined shortly. Clearly, the probability that any pair of strings are at distance at most $n/2 - \Delta$, is

$$\alpha \leq \binom{M}{2} \exp\Big(-2\Delta^2/n\Big) < M^2 \exp\Big(-2\Delta^2/n\Big).$$

If this probability is smaller than one, then there is some probability that all the M strings are of distance at least $n/2 - \Delta$ from each other. Namely, there exists a set of M strings such that every pair of them is far. We used here the fact that if an event has probability larger than zero, then it exists. Thus, set $\Delta = n/4$, and observe that

$$\alpha < M^2 \exp\Big(-2n^2/16n\Big) = M^2 \exp(-n/8).$$

Thus, for $M = \exp(n/16)$, we have that $\alpha < 1$. We conclude:

Lemma 7.2.9. *There exists a set of $\exp(n/16)$ binary strings of length n, such that any pair of them is at Hamming distance at least $n/4$ from each other.*

This is our first introduction to the beautiful technique known as the probabilistic method — we will hear more about it later in the course.

This result has also interesting interpretation in the Euclidean setting. Indeed, consider the sphere $\mathbb{S}$ of radius $\sqrt{n}/2$ centered at $(1/2, 1/2, \ldots, 1/2) \in \mathbb{R}^n$. Clearly, all the vertices of the binary hypercube $\{0,1\}^n$ lie on this sphere. As such, let P be the set of points on $\mathbb{S}$ that exists according to Lemma 7.2.9. A pair p, q of points of P have *Euclidean* distance at least $\sqrt{d_H(p,q)} = \sqrt{n}4 = \sqrt{n}/2$ from each other. We conclude:

Lemma 7.2.10. *Consider the unit hypersphere $\mathbb{S}$ in $\mathbb{R}^n$. The sphere $\mathbb{S}$ contains a set Q of points, such that each pair of points is at (Euclidean) distance at least one from each other, and $|Q| \geq \exp(n/16)$.*

7.3. The Chernoff Bound — General Case

Here we present the Chernoff bound in a more general settings.

Question 7.3.1. *Let $X_1, \ldots, X_n$ be n independent Bernoulli trials, where*

$$\mathbf{Pr}\Big[X_i = 1\Big] = p_i, \quad \text{and} \quad \mathbf{Pr}\Big[X_i = 0\Big] = q_i = 1 - p_i.$$

(Each X_i is known as a Poisson trials.) And let $X = \sum_{i=1}^{b} X_i$. $\mu = \mathbf{E}[X] = \sum_i p_i$. We are interested in the question of what is the probability that $X > (1+\delta)\mu$?

Theorem 7.3.2. *For any $\delta > 0$, we have $\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \left(\dfrac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$.*

Or in a more simplified form, we have:

$$\delta \leq 2e - 1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\Big(-\mu\delta^2/4\Big), \tag{7.1}$$

$$\delta > 2e - 1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < 2^{-\mu(1+\delta)}, \tag{7.2}$$

and $$\delta \geq e^2 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\left(-\frac{\mu\delta\ln\delta}{2}\right). \tag{7.3}$$

Proof: We have $\mathbf{Pr}[X > (1+\delta)\mu] = \mathbf{Pr}\Big[e^{tX} > e^{t(1+\delta)\mu}\Big]$. By the Markov inequality, we have:

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \frac{\mathbf{E}\Big[e^{tX}\Big]}{e^{t(1+\delta)\mu}}$$

On the other hand,

$$\mathbf{E}\Big[e^{tX}\Big] = \mathbf{E}\Big[e^{t(X_1+X_2\ldots+X_n)}\Big] = \mathbf{E}\Big[e^{tX_1}\Big]\cdots\mathbf{E}\Big[e^{tX_n}\Big].$$

Namely,

$$\mathbf{Pr}[X > (1+\delta)\mu] < \frac{\prod_{i=1}^n \mathbf{E}\Big[e^{tX_i}\Big]}{e^{t(1+\delta)\mu}} = \frac{\prod_{i=1}^n \Big((1-p_i)e^0 + p_i e^t\Big)}{e^{t(1+\delta)\mu}} = \frac{\prod_{i=1}^n (1 + p_i(e^t-1))}{e^{t(1+\delta)\mu}}.$$

Let $y = p_i(e^t - 1)$. We know that $1 + y < e^y$ (since $y > 0$). Thus,

$$\begin{aligned}
\mathbf{Pr}[X > (1+\delta)\mu] &< \frac{\prod_{i=1}^n \exp(p_i(e^t-1))}{e^{t(1+\delta)\mu}} = \frac{\exp(\sum_{i=1}^n p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\
&= \frac{\exp((e^t-1)\sum_{i=1}^n p_i)}{e^{t(1+\delta)\mu}} = \frac{\exp((e^t-1)\mu)}{e^{t(1+\delta)\mu}} = \left(\frac{\exp(e^t-1)}{e^{t(1+\delta)}}\right)^\mu \\
&= \left(\frac{\exp(\delta)}{(1+\delta)^{(1+\delta)}}\right)^\mu,
\end{aligned}$$

if we set $t = \log(1+\delta)$.

For the proof of the simplified form, see Section 7.3.1. ■

Definition 7.3.3. $F^+(\mu,\delta) = \left[\dfrac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^\mu$.

Example 7.3.4. Arkansas Aardvarks win a game with probability $1/3$. What is their probability to have a winning season with n games. By Chernoff inequality, this probability is smaller than

$$F^+(n/3, 1/2) = \left[\frac{e^{1/2}}{1.5^{1.5}}\right]^{n/3} = (0.89745)^{n/3} = 0.964577^n.$$

For $n = 40$, this probability is smaller than 0.236307. For $n = 100$ this is less than 0.027145. For $n = 1000$, this is smaller than $2.17221 \cdot 10^{-16}$ (which is pretty slim and shady). Namely, as the number of experiments is increases, the distribution converges to its expectation, and this converge is exponential.

Theorem 7.3.5. *Under the same assumptions as Theorem 7.3.2, we have:* $\mathbf{Pr}[X < (1-\delta)\mu] < \exp\left(-\mu\delta^2/2\right)$.

Definition 7.3.6. Let $F^-(\mu,\delta) = e^{-\mu\delta^2/2}$, and let $\Delta^-(\mu,\varepsilon)$ denote the quantity, which is what should be the value of δ, so that the probability is smaller than ε. We have that

$$\Delta^-(\mu,\varepsilon) = \sqrt{\frac{2\log 1/\varepsilon}{\mu}}.$$

And for large δ we have $\Delta^+(\mu,\varepsilon) < \dfrac{\log_2(1/\varepsilon)}{\mu} - 1$.

7.3.1. A More Convenient Form

Proof: **(of simplified form of Theorem 7.3.2$_{\text{p12}}$)** Eq. (7.2) is easy. Indeed, we have

$$\left[\frac{e}{1+\delta}\right]^{(1+\delta)\mu} \le \left[\frac{e}{1+2e-1}\right]^{(1+\delta)\mu} \le 2^{-(1+\delta)\mu},$$

since $\delta > 2e-1$. For the stronger version, Eq. (7.3), observe that

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu = \exp\Big(\mu\delta - \mu(1+\delta)\ln(1+\delta)\Big). \tag{7.4}$$

As such, we have

$$\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\Big(-\mu(1+\delta)(\ln(1+\delta)-1)\Big) \le \exp\left(-\mu\delta\ln\frac{1+\delta}{e}\right) \le \exp\left(-\frac{\mu\delta\ln\delta}{2}\right),$$

since for $x \ge e^2$ we have that $\dfrac{1+x}{e} \ge \sqrt{x} \iff \ln\dfrac{1+x}{e} \ge \dfrac{\ln x}{2}$.

As for Eq. (7.1), we prove this only for $\delta \le 1/2$. For details about the case $1/2 \le \delta \le 2e-1$, see [MR95]. The Taylor expansion of $\ln(1+\delta)$ is

$$\delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} - \frac{\delta^4}{4} + \cdot \ge \delta - \frac{\delta^2}{2},$$

for $\delta \leq 1$. Thus, plugging into Eq. (7.4), we have

$$\begin{aligned}\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] &< \exp\Big(\mu\Big[\delta - (1+\delta)(\delta - \delta^2/2)\Big]\Big) = \exp\Big(\mu\Big(\delta - \delta + \delta^2/2 - \delta^2 + \delta^3/2\Big)\Big) \\ &\leq \exp\Big(\mu\Big(-\delta^2/2 + \delta^3/2\Big)\Big) \leq \exp\Big(-\mu\delta^2/4\Big),\end{aligned}$$

for $\delta \leq 1/2$. ■

7.4. A special case of Hoeffding's inequality

In this section, we prove yet another version of Chernoff inequality, where each variable is randomly picked according to its own distribution in the range $[0, 1]$. We prove a more general version of this inequality in Section 7.5, but the version presented here does not follow from this generalization.

Theorem 7.4.1. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $X = \sum_{i=1}^n X_i$, and let $\mu = \mathbf{E}[X]$. We have that $\mathbf{Pr}\Big[X - \mu \geq \eta\Big] \leq \left(\frac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^{n-\mu-\eta}$.*

Proof: Let $s \geq 1$ be some arbitrary parameter. By the standard arguments, we have

$$\gamma = \mathbf{Pr}\Big[X \geq \mu + \eta\Big] = \mathbf{Pr}\Big[s^X \geq s^{\mu+\eta}\Big] \leq \frac{\mathbf{E}\Big[s^X\Big]}{s^{\mu+\eta}} = s^{-\mu-\eta}\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big].$$

By calculations, see Lemma 7.4.6 below, one can show that $\mathbf{E}\Big[s^{X_1}\Big] \leq 1 + (s-1)\,\mathbf{E}[X_i]$. As such, by the AM-GM inequality[8], we have that

$$\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big] \leq \prod_{i=1}^n \Big(1 + (s-1)\,\mathbf{E}[X_i]\Big) \leq \left(\frac{1}{n}\sum_{i=1}^n \Big(1 + (s-1)\,\mathbf{E}[X_i]\Big)\right)^n = \left(1 + (s-1)\frac{\mu}{n}\right)^n.$$

Setting $s = \dfrac{(\mu+\eta)(n-\mu)}{\mu(n-\mu-\eta)} = \dfrac{\mu n - \mu^2 + \eta n - \eta\mu}{\mu n - \mu^2 - \eta\mu}$ we have that

$$1 + (s-1)\frac{\mu}{n} = 1 + \frac{\eta n}{\mu n - \mu^2 - \eta\mu}\cdot\frac{\mu}{n} = 1 + \frac{\eta}{n-\mu-\eta} = \frac{n-\mu}{n-\mu-\eta}.$$

As such, we have that

$$\gamma \leq s^{-\mu-\eta}\prod_{i=1}^n \mathbf{E}\Big[s^{X_i}\Big] = \left(\frac{\mu(n-\mu-\eta)}{(\mu+\eta)(n-\mu)}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^n = \left(\frac{\mu}{(\mu+\eta)}\right)^{\mu+\eta}\left(\frac{n-\mu}{n-\mu-\eta}\right)^{n-\mu-\eta}.$$ ■

Remark 7.4.2. Setting $s = (\mu+\eta)/\mu$ in the proof of Theorem 7.4.1, we have

$$\mathbf{Pr}\Big[X - \mu \geq \eta\Big] \leq \left(\tfrac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(1 + \left(\tfrac{\mu+\eta}{\mu} - 1\right)\tfrac{\mu}{n}\right)^n = \left(\tfrac{\mu}{\mu+\eta}\right)^{\mu+\eta}\left(1 + \tfrac{\eta}{n}\right)^n.$$

[8] The inequality between arithmetic and geometric means: $(\sum_{i=1}^n x_i)/n \geq \sqrt[n]{x_1 \cdots x_n}$.

Corollary 7.4.3. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $\overline{X} = \sum_{i=1}^n X_i/n$, $p = \mathbf{E}\left[\overline{X}\right] = \mu/n$ and $q = 1-p$. Then, we have that $\mathbf{Pr}\left[\overline{X} - p \geq t\right] \leq \exp(nf(t))$, for*

$$f(t) = (p+t)\ln\frac{p}{p+t} + (q-t)\ln\frac{q}{q-t}. \tag{7.5}$$

Theorem 7.4.4. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $\overline{X} = (\sum_{i=1}^n X_i)/n$, and let $p = \mathbf{E}[X]$. We have that $\mathbf{Pr}\left[\overline{X} - p \geq t\right] \leq \exp\left(-2nt^2\right)$ and $\mathbf{Pr}\left[\overline{X} - p \leq -t\right] \leq \exp\left(-2nt^2\right)$.*

Proof: Let $p = \mu/n$, $q = 1-p$, and let $f(t)$ be the function from Eq. (7.5), for $t \in (-p, q)$. Now, we have that

$$\begin{aligned} f'(t) &= \ln\frac{p}{p+t} + (p+t)\frac{p+t}{p}\left(-\frac{p}{(p+t)^2}\right) - \ln\frac{q}{q-t} - (q-t)\frac{q-t}{q}\frac{q}{(q-t)^2} = \ln\frac{p}{p+t} - \ln\frac{q}{q-t} \\ &= \ln\frac{p(q-t)}{q(p+t)}. \end{aligned}$$

As for the second derivative, we have

$$f''(t) = \frac{\cancel{q(p+t)}}{p(q-t)} \cdot \frac{p}{q} \cdot \frac{(p+t)(-1) - (q-t)}{\cancel{(p+t)^2}}. = \frac{-p-t-q+t}{(q-t)(p+t)} = -\frac{1}{(q-t)(p+t)} \leq -4.$$

Indeed, $t \in (-p, q)$ and the denominator is minimized for $t = (q-p)/2$, and as such $(q-t)(p+t) \leq \Big(2q - (q-p)\Big)\Big(2p + (q-p)\Big)/4 = (p+q)^2/4 = 1/4$.

Now, $f(0) = 0$ and $f'(0) = 0$, and by Taylor's expansion, we have that $f(t) = f(0) + f'(0)t + \dfrac{f''(x)}{2}t^2 \leq -2t^2$, where x is between 0 and t.

The first bound now readily follows from plugging this bound into Corollary 7.4.3. The second bound follows by considering the random variants $Y_i = 1 - X_i$, for all i, and plugging this into the first bound. Indeed, for $\overline{Y} = 1 - \overline{X}$, we have that $q = \mathbf{E}\left[\overline{Y}\right]$, and then $\overline{X} - p \leq -t \iff t \leq p - \overline{X} \iff t \leq 1 - q - (1 - \overline{Y}) = \overline{Y} - q$. Thus, $\mathbf{Pr}\left[\overline{X} - p \leq -t\right] = \mathbf{Pr}\left[\overline{Y} - q \geq t\right] \leq \exp\left(-2nt^2\right)$. ■

Theorem 7.4.5. *Let $X_1, \ldots, X_n \in [0,1]$ be n independent random variables, let $X = (\sum_{i=1}^n X_i)$, and let $\mu = \mathbf{E}[X]$. We have that $\mathbf{Pr}[X - \mu \geq \varepsilon\mu] \leq \exp\left(-\varepsilon^2\mu/4\right)$ and $\mathbf{Pr}[X - \mu \leq -\varepsilon\mu] \leq \exp\left(-\varepsilon^2\mu/2\right)$.*

Proof: Let $p = \mu/n$, and let $g(x) = f(px)$, for $x \in [0,1]$ and $xp < q$. As before, computing the derivative of g, we have

$$g'(x) = pf'(xp) = p\ln\frac{p(q-xp)}{q(p+xp)} = p\ln\frac{q-xp}{q(1+x)} \leq p\ln\frac{1}{1+x} \leq -\frac{px}{2},$$

since $(q-xp)/q$ is maximized for $x = 0$, and$\ln\frac{1}{1+x} \leq -x/2$, for $x \in [0,1]$, as can be easily verified[9]. Now, $g(0) = f(0) = 0$, and by integration, we have that $g(x) = \int_{y=0}^{x} g'(y)\mathrm{d}y \leq \int_{y=0}^{x}(-py/2)\mathrm{d}y = -px^2/4$. Now, plugging into Corollary 7.4.3, we get that the desired probability $\mathbf{Pr}[X - \mu \geq \varepsilon\mu]$ is

$$\mathbf{Pr}\left[\overline{X} - p \geq \varepsilon p\right] \leq \exp(nf(\varepsilon p)) = \exp\left(ng(\varepsilon)\right) \leq \exp\left(-pn\varepsilon^2/4\right) = \exp\left(-\mu\varepsilon^2/4\right).$$

[9] Indeed, this is equivalent to $\frac{1}{1+x} \leq e^{-x/2} \iff e^{x/2} \leq 1 + x$, which readily holds for $x \in [0,1]$.

As for the other inequality, set $h(x) = g(-x) = f(-xp)$. Then

$$h'(x) = -pf'(-xp) = -p \ln \frac{p(q+xp)}{q(p-xp)} = p \ln \frac{q(1-x)}{q+xp} = p \ln \frac{q-xq}{q+xp} = p \ln\left(1 - x\frac{p+q}{q+xp}\right)$$
$$= p \ln\left(1 - x\frac{1}{q+xp}\right) \leq p \ln(1-x) \leq -px,$$

since $1 - x \leq e^{-x}$. By integration, as before, we conclude that $h(x) \leq -px^2/2$. Now, plugging into Corollary 7.4.3, we get $\mathbf{Pr}[X - \mu \leq -\varepsilon\mu] = \mathbf{Pr}\left[\overline{X} - p \leq -\varepsilon p\right] \leq \exp(nf(-\varepsilon p)) \leq \exp(nh(\varepsilon)) \leq \exp\left(-np\varepsilon^2/2\right) \leq \exp\left(-\mu\varepsilon^2/2\right)$. ■

7.4.1. Some technical lemmas

Lemma 7.4.6. *Let $X \in [0,1]$ be a random variable, and let $s \geq 1$. Then $\mathbf{E}\left[s^X\right] \leq 1 + (s-1)\,\mathbf{E}[X]$.*

Proof: For the sake of simplicity of exposition, assume that X is a discrete random variable, and that there is a value $\alpha \in (0, 1/2)$, such that $\beta = \mathbf{Pr}[X = \alpha] > 0$. Consider the modified random variable X', such that $\mathbf{Pr}[X' = 0] = \mathbf{Pr}[X = 0] + \beta/2$, and $\mathbf{Pr}[X' = 2\alpha] = \mathbf{Pr}[X = \alpha] + \beta/2$. Clearly, $\mathbf{E}[X] = \mathbf{E}[X']$. Next, observe that $\mathbf{E}\left[s^{X'}\right] - \mathbf{E}\left[s^X\right] = (\beta/2)(s^{2\alpha} + s^0) - \beta s^\alpha \geq 0$, by the convexity of s^x. We conclude that $\mathbf{E}\left[s^X\right]$ achieves its maximum if takes only the values 0 and 1. But then, we have that $\mathbf{E}\left[s^X\right] = \mathbf{Pr}[X = 0]\, s^0 + \mathbf{Pr}[X = 1]\, s^1 = (1 - \mathbf{E}[X]) + \mathbf{E}[X]\, s = 1 + (s-1)\,\mathbf{E}[X]$, as claimed. ■

7.5. Hoeffding's inequality

In this section, we prove a generalization of Chernoff's inequality. The proof is considerably more tedious, and it is included here for the sake of completeness.

Lemma 7.5.1. *Let X be a random variable. If $\mathbf{E}[X] = 0$ and $a \leq X \leq b$, then for any $s > 0$, we have $\mathbf{E}\left[e^{sX}\right] \leq \exp\left(s^2(b-a)^2/8\right)$.*

Proof: Let $a \leq x \leq b$ and observe that x can be written as a convex combination of a and b. In particular, we have

$$x = \lambda a + (1-\lambda)b \quad \text{for} \quad \lambda = \frac{b-x}{b-a} \in [0,1].$$

Since $s > 0$, the function $\exp(sx)$ is convex, and as such

$$e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb},$$

since we have that $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ if $f(\cdot)$ is a convex function. Thus, for a random variable X, by linearity of expectation, we have

$$\mathbf{E}\left[e^{sX}\right] \leq \mathbf{E}\left[\frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}\right] = \frac{b-\mathbf{E}[X]}{b-a}e^{sa} + \frac{\mathbf{E}[X]-a}{b-a}e^{sb}$$
$$= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb},$$

since $\mathbf{E}[X] = 0$.

Next, set $p = -\dfrac{a}{b-a}$ and observe that $1 - p = 1 + \dfrac{a}{b-a} = \dfrac{b}{b-a}$ and

$$-ps(b-a) = -\left(-\frac{a}{b-a}\right)s(b-a) = sa.$$

As such, we have

$$\begin{aligned}\mathbf{E}\left[e^{sX}\right] &\leq (1-p)e^{sa} + pe^{sb} = (1 - p + pe^{s(b-a)})e^{sa} \\ &= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} \\ &= \exp\left(-ps(b-a) + \ln\left(1 - p + pe^{s(b-a)}\right)\right) = \exp(-pu + \ln(1 - p + pe^{u})),\end{aligned}$$

for $u = s(b-a)$. Setting

$$\phi(u) = -pu + \ln(1 - p + pe^{u}),$$

we thus have $\mathbf{E}\left[e^{sX}\right] \leq \exp(\phi(u))$. To prove the claim, we will show that $\phi(u) \leq u^2/8 = s^2(b-a)^2/8$.

To see that, expand $\phi(u)$ about zero using Taylor's expansion. We have

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(\theta) \tag{7.6}$$

where $\theta \in [0, u]$, and notice that $\phi(0) = 0$. Furthermore, we have

$$\phi'(u) = -p + \frac{pe^{u}}{1 - p + pe^{u}},$$

and as such $\phi'(0) = -p + \frac{p}{1-p+p} = 0$. Now,

$$\phi''(u) = \frac{(1 - p + pe^{u})pe^{u} - (pe^{u})^2}{(1 - p + pe^{u})^2} = \frac{(1-p)pe^{u}}{(1 - p + pe^{u})^2}.$$

For any $x, y \geq 0$, we have $(x+y)^2 \geq 4xy$ as this is equivalent to $(x-y)^2 \geq 0$. Setting $x = 1 - p$ and $y = pe^{u}$, we have that

$$\phi''(u) = \frac{(1-p)pe^{u}}{(1 - p + pe^{u})^2} \leq \frac{(1-p)pe^{u}}{4(1-p)pe^{u}} = \frac{1}{4}.$$

Plugging this into Eq. (7.6), we get that

$$\phi(u) \leq \frac{1}{8}u^2 = \frac{1}{8}(s(b-a))^2 \quad \text{and} \quad \mathbf{E}\left[e^{sX}\right] \leq \exp(\phi(u)) \leq \exp\left(\frac{1}{8}(s(b-a))^2\right),$$

as claimed. ∎

Lemma 7.5.2. *Let X be a random variable. If $\mathbf{E}[X] = 0$ and $a \leq X \leq b$, then for any $s > 0$, we have*

$$\mathbf{Pr}[X > t] \leq \frac{\exp\left(\frac{s^2(b-a)^2}{8}\right)}{e^{st}}.$$

Proof: Using the same technique we used in proving Chernoff's inequality, we have that

$$\mathbf{Pr}[X > t] = \mathbf{Pr}\left[e^{sX} > e^{st}\right] \leq \frac{\mathbf{E}\left[e^{sX}\right]}{e^{st}} \leq \frac{\exp\left(\frac{s^2(b-a)^2}{8}\right)}{e^{st}}.$$ ∎

Theorem 7.5.3 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \in [a_i, b_i]$, for $i = 1, \ldots, n$. Then, for the random variable $S = X_1 + \cdots + X_n$ and any $\eta > 0$, we have*

$$\mathbf{Pr}\left[\left|S - \mathbf{E}[S]\right| \geq \eta\right] \leq 2\exp\left(-\frac{2\eta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof: Let $Z_i = X_i - \mathbf{E}[X_i]$, for $i = 1, \ldots, n$. Set $Z = \sum_{i=1}^{n} Z_i$, and observe that

$$\mathbf{Pr}[Z \geq \eta] = \mathbf{Pr}\left[e^{sZ} \geq e^{s\eta}\right] \leq \frac{\mathbf{E}[\exp(sZ)]}{\exp(s\eta)},$$

by Markov's inequality. Arguing as in the proof of Chernoff's inequality, we have

$$\mathbf{E}[\exp(sZ)] = \mathbf{E}\left[\prod_{i=1}^{n}\exp(sZ_i)\right] = \prod_{i=1}^{n}\mathbf{E}[\exp(sZ_i)] \leq \prod_{i=1}^{n}\exp\left(\frac{s^2(b_i - a_i)^2}{8}\right),$$

since the Z_is are independent and by Lemma 7.5.1. This implies that

$$\mathbf{Pr}[Z \geq \eta] \leq \exp(-s\eta)\prod_{i=1}^{n} e^{s^2(b_i - a_i)^2/8} = \exp\left(\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - s\eta\right).$$

The upper bound is minimized for $s = 4\eta/\left(\sum_i (b_i - a_i)^2\right)$, implying

$$\mathbf{Pr}[Z \geq \eta] \leq \exp\left(-\frac{2\eta^2}{\sum(b_i - a_i)^2}\right).$$

The claim now follows by the symmetry of the upper bound (i.e., apply the same proof to $-Z$). ■

7.6. Bibliographical notes

Some of the exposition here follows more or less the exposition in [MR95]. Exercise 7.7.1 (without the hint) is from [Mat99]. McDiarmid [McD89] provides a survey of Chernoff type inequalities, and Theorem 7.4.5 and Section 7.4 is taken from there (our proof has somewhat weaker constants).

Section 7.2.3 is based on Section 4.2 in [MR95]. A similar result to Theorem 7.2.8 is known for the case of the wrapped butterfly topology (which is similar to the hypercube topology but every node has a constant degree, and there is no clear symmetry). The interested reader is referred to [MU05].

A more general treatment of such inequalities and tools is provided by Dubhashi and Panconesi [DP09].

7.7. Exercises

Exercise 7.7.1 (Chernoff inequality is tight.). Let $S = \sum_{i=1}^{n} S_i$ be a sum of n independent random variables each attaining values $+1$ and -1 with equal probability. Let $P(n, \Delta) = \mathbf{Pr}[S > \Delta]$. Prove that for $\Delta \leq n/C$,

$$P(n, \Delta) \geq \frac{1}{C}\exp\left(-\frac{\Delta^2}{Cn}\right),$$

where C is a suitable constant. That is, the well-known Chernoff bound $P(n, \Delta) \leq \exp(-\Delta^2/2n))$ is close to the truth.

Exercise 7.7.2 (Chernoff inequality is tight by direct calculations.). For this question use only basic argumentation – do not use Stirling's formula, Chernoff inequality or any similar "heavy" machinery.

(A) Prove that $\sum_{i=0}^{n-k} \binom{2n}{i} \leq \frac{n}{4k^2} 2^{2n}$.

Hint: Consider flipping a coin $2n$ times. Write down explicitly the probability of this coin to have at most $n-k$ heads, and use Chebyshev inequality.

(B) Using (A), prove that $\binom{2n}{n} \geq 2^{2n}/4\sqrt{n}$ (which is a pretty good estimate).

(C) Prove that $\binom{2n}{n+i+1} = \left(1 - \frac{2i+1}{n+i+1}\right)\binom{2n}{n+i}$.

(D) Prove that $\binom{2n}{n+i} \leq \exp\left(\frac{-i(i-1)}{2n}\right)\binom{2n}{n}$.

(E) Prove that $\binom{2n}{n+i} \geq \exp\left(-\frac{8i^2}{n}\right)\binom{2n}{n}$.

(F) Using the above, prove that $\binom{2n}{n} \leq c\frac{2^{2n}}{\sqrt{n}}$ for some constant c (I got $c = 0.824...$ but any reasonable constant will do).

(G) Using the above, prove that

$$\sum_{i=t\sqrt{n}+1}^{(t+1)\sqrt{n}} \binom{2n}{n-i} \leq c2^{2n} \exp\left(-t^2/2\right).$$

In particular, conclude that when flipping fair coin $2n$ times, the probability to get less than $n - t\sqrt{n}$ heads (for t an integer) is smaller than $c' \exp\left(-t^2/2\right)$, for some constant c'.

(H) Let X be the number of heads in $2n$ coin flips. Prove that for any integer $t > 0$ and any $\delta > 0$ sufficiently small, it holds that $\mathbf{Pr}[X < (1-\delta)n] \geq \exp\left(-c''\delta^2 n\right)$, where c'' is some constant. Namely, the Chernoff inequality is tight in the worst case.

Exercise 7.7.3 (More binary strings. More!). To some extent, Lemma 7.2.9 is somewhat silly, as one can prove a better bound by direct argumentation. Indeed, for a fixed binary string x of length n, show a bound on the number of strings in the Hamming ball around x of radius $n/4$ (i.e., binary strings of distance at most $n/4$ from x). (Hint: interpret the special case of the Chernoff inequality as an inequality over binomial coefficients.)

Next, argue that the greedy algorithm which repeatedly pick a string which is in distance $\geq n/4$ from all strings picked so far, stops after picking at least $\exp(n/8)$ strings.

Exercise 7.7.4 (Tail inequality for geometric variables). Let $X_1, \ldots, X_m$ be m independent random variables with geometric distribution with probability p (i.e., $\mathbf{Pr}[X_i = j] = (1-p)^{j-1}p$). Let $Y = \sum_i X_i$, and let $\mu = \mathbf{E}[Y] = m/p$. Prove that $\mathbf{Pr}\left[Y \geq (1+\delta)\mu\right] \leq \exp\left(-m\delta^2/8\right)$.

Bibliography

[DP09] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[Kel56] J. L. Kelly. A new interpretation of information rate. *Bell Sys. Tech. J.*, 35(4):917–926, jul 1956.

[KKT91] C. Kaklamanis, D. Krizanc, and T. Tsantilas. Tight bounds for oblivious routing in the hypercube. *Math. sys. theory*, 24(1):223–232, 1991.

[Mat99] J. Matoušek. *Geometric Discrepancy*. Springer, 1999.

[McD89] C. McDiarmid. *Surveys in Combinatorics*, chapter On the method of bounded differences. Cambridge University Press, 1989.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.

Chapter 8

Martingales

By Sariel Har-Peled, December 30, 2015[①]

> *'After that he always chose out a "dog command" and sent them ahead. It had the task of informing the inhabitants in the village where we were going to stay overnight that no dog must be allowed to bark in the night otherwise it would be liquidated. I was also on one of those commands and when we came to a village in the region of Milevsko I got mixed up and told the mayor that every dog-owner whose dog barked in the night would be liquidated for strategic reasons. The mayor got frightened, immediately harnessed his horses and rode to headquarters to beg mercy for the whole village. They didn't let him in, the sentries nearly shot him and so he returned home, but before we got to the village everybody on his advice had tied rags round the dogs muzzles with the result that three of them went mad.'*
>
> – The good soldier Svejk, Jaroslav Hasek

8.1. Martingales

8.1.1. Preliminaries

Let X and Y be two random variables. Let $\rho(x,y) = \mathbf{Pr}[(X = x) \cap (Y = y)]$. Then,

$$\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big] = \frac{\rho(x,y)}{\mathbf{Pr}[Y = y]} = \frac{\rho(x,y)}{\sum_z \rho(z,y)}$$

and $\mathbf{E}\Big[X \;\Big|\; Y = y\Big] = \sum_x x \,\mathbf{Pr}\Big[X = x \;\Big|\; Y = y\Big] = \dfrac{\sum_x x\rho(x,y)}{\sum_z \rho(z,y)} = \dfrac{\sum_x x\rho(x,y)}{\mathbf{Pr}[Y = y]}.$

Definition 8.1.1. The *conditional expectation* of X given Y, is the random variable $\mathbf{E}\Big[X \;\Big|\; Y\Big]$ is the random variable $f(y) = \mathbf{E}\Big[X \;\Big|\; Y = y\Big]$.

Lemma 8.1.2. *For any two random variables X and Y, we have* $\mathbf{E}\Big[\mathbf{E}\Big[X \;\Big|\; Y\Big]\Big] = \mathbf{E}\Big[X\Big]$.

Proof: $\mathbf{E}\Big[\mathbf{E}\Big[X \;\Big|\; Y\Big]\Big] = \mathbf{E}_Y\Big[\mathbf{E}\Big[X \;\Big|\; Y = y\Big]\Big] = \sum_y \mathbf{Pr}[Y = y]\,\mathbf{E}\Big[X \;\Big|\; Y = y\Big]$

$$\begin{aligned}
&= \sum_y \mathbf{Pr}\Big[Y = y\Big]\,\frac{\sum_x x\,\mathbf{Pr}[X = x \cap Y = y]}{\mathbf{Pr}[Y = y]} \\
&= \sum_y \sum_x x\,\mathbf{Pr}\Big[X = x \cap Y = y\Big] = \sum_x x \sum_y \mathbf{Pr}\Big[X = x \cap Y = y\Big] \\
&= \sum_x x\,\mathbf{Pr}\Big[X = x\Big] = \mathbf{E}\Big[X\Big].
\end{aligned}$$

■

Lemma 8.1.3. *For any two random variables X and Y, we have* $\mathbf{E}\Big[Y \cdot \mathbf{E}\Big[X\Big|Y\Big]\Big] = \mathbf{E}\Big[XY\Big]$.

Proof: We have that $\mathbf{E}\Big[Y \cdot \mathbf{E}\Big[X \;\Big|\; Y\Big]\Big] = \sum_y \mathbf{Pr}[Y = y] \cdot y \cdot \mathbf{E}\Big[X \;\Big|\; Y = y\Big]$

$$= \sum_y \mathbf{Pr}[Y = y] \cdot y \cdot \frac{\sum_x x\,\mathbf{Pr}[X = x \cap Y = y]}{\mathbf{Pr}[Y = y]} = \sum_x \sum_y xy \cdot \mathbf{Pr}\Big[X = x \cap Y = y\Big] = \mathbf{E}\Big[XY\Big].$$

■

8.1.2. Martingales

Intuitively, martingales are a sequence of random variables describing a process, where the only thing that matters at the beginning of the ith step is where the process was in the end of the $(i-1)$th step. That is, it does not matter how the process arrived to a certain state, only that it is currently at this state.

Definition 8.1.4. A sequence of random variables $X_0, X_1, \ldots,$ is said to be a *martingale sequence* if for all $i > 0$, we have $\mathbf{E}\Big[X_i \;\Big|\; X_0, \ldots, X_{i-1}\Big] = X_{i-1}$.

Lemma 8.1.5. *Let $X_0, X_1, \ldots,$ be a martingale sequence. Then, for all $i \geq 0$, we have* $\mathbf{E}[X_i] = \mathbf{E}[X_0]$.

8.1.2.1. Examples of martingales

Example 8.1.6. An example of martingales is the sum of money after participating in a sequence of fair bets. That is, let X_i be the amount of money a gambler has after playing i rounds. In each round it either gains one dollar, or loses one dollar. Clearly, we have $\mathbf{E}\Big[X_i \;\Big|\; X_0, \ldots, X_{i-1}\Big] = \mathbf{E}\Big[X_i \;\Big|\; X_{i-1}\Big] = X_i$.

Example 8.1.7. Let $Y_i = X_i^2 - i$, where X_i is as defined in the above example. We claim that $Y_0, Y_1, \ldots$ is a martingale. Let us verify that this is true. Given Y_{i-1}, we have $Y_{i-1} = X_{i-1}^2 - (i-1)$. We have that

$$\begin{aligned}
\mathbf{E}\Big[Y_i \;\Big|\; Y_{i-1}\Big] &= \mathbf{E}\Big[X_i^2 - i \;\Big|\; X_{i-1}^2 - (i-1)\Big] = \frac{1}{2}\Big((X_{i-1}+1)^2 - i)\Big) + \frac{1}{2}\Big((X_{i-1}-1)^2 - i\Big) \\
&= X_{i-1}^2 + 1 - i = X_{i-1}^2 - (i-1) = Y_{i-1},
\end{aligned}$$

which implies that indeed it is a martingale.

Example 8.1.8. Let U be a urn with b black balls, and w white balls. We repeatedly select a ball and replace it by c balls having the same color. Let X_i be the fraction of black balls after the first i trials. We claim that the sequence $X_0, X_1, \ldots$ is a martingale.

Indeed, let $n_i = b + w + i(c-1)$ be the number of balls in the urn after the ith trial. Clearly,

$$\begin{aligned}\mathbf{E}\Big[X_i \;\Big|\; X_{i-1}, \ldots, X_0\Big] &= X_{i-1} \cdot \frac{(c-1) + X_{i-1}n_{i-1}}{n_i} + (1 - X_{i-1}) \cdot \frac{X_{i-1}n_{i-1}}{n_i} \\ &= \frac{X_{i-1}(c-1) + X_{i-1}n_{i-1}}{n_i} = X_{i-1}\frac{c-1+n_{i-1}}{n_i} = X_{i-1}\frac{n_i}{n_i} = X_{i-1}.\end{aligned}$$

Example 8.1.9. Let G be a random graph on the vertex set $V = \{1, \ldots, n\}$ obtained by independently choosing to include each possible edge with probability p. The underlying probability space is called $\mathbf{G}_{n,p}$. Arbitrarily label the $m = n(n-1)/2$ possible edges with the sequence $1, \ldots, m$. For $1 \le j \le m$, define the indicator random variable I_j, which takes values 1 if the edge j is present in G, and has value 0 otherwise. These indicator variables are independent and each takes value 1 with probability p.

Consider any real valued function f defined over the space of all graphs, e.g., the clique number, which is defined as being the size of the largest complete subgraph. The ***edge exposure martingale*** is defined to be the sequence of random variables $X_0, \ldots, X_m$ such that

$$X_i = \mathbf{E}\Big[f(G) \;\Big|\; I_1, \ldots, I_i\Big],$$

while $X_0 = \mathbf{E}[f(G)]$ and $X_m = f(G)$. This sequence of random variable begin a martingale follows immediately from a theorem that would be described in the next lecture.

One can define similarly a ***vertex exposure martingale***, where the graph G_i is the graph induced on the first i vertices of the random graph G.

Example 8.1.10 (The sheep of Mabinogion). The following is taken from medieval Welsh manuscript based on Celtic mythology:

> "And he came towards a valley, through which ran a river; and the borders of the valley were wooded, and on each side of the river were level meadows. And on one side of the river he saw a flock of white sheep, and on the other a flock of black sheep. And whenever one of the white sheep bleated, one of the black sheep would cross over and become white; and when one of the black sheep bleated, one of the white sheep would cross over and become black." – *Peredur the son of Evrawk*, from the *Mabinogion*.

More concretely, we start at time 0 with w_0 white sheep, and b_0 black sheep. At every iteration, a random sheep is picked, it bleats, and a sheep of the other color turns to this color. the game stops as soon as all the sheep have the same color. No sheep dies or get born during the game. Let X_i be the expected number of black sheep in the end of the game, after the ith iteration. For reasons that we would see later on, this sequence is a martingale.

The original question is somewhat more interesting – if we are allowed to take a way sheep in the end of each iteration, what is the optimal strategy to maximize X_i?

8.1.2.2. Azuma's inequality

A sequence of random variables $X_0, X_1, \ldots$ has *bounded differences* if $|X_i - X_{i-1}| \le \Delta$, for some fixed Δ.

Theorem 8.1.11 (Azuma's Inequality.). *Let $X_0, \ldots, X_m$ be a martingale with $X_0 = 0$, and $|X_{i+1} - X_i| \le 1$ for all $0 \le i < m$. Let $\lambda > 0$ be arbitrary. Then* $\mathbf{Pr}\Big[X_m > \lambda\sqrt{m}\Big] < \exp\Big(-\lambda^2/2\Big)$.

Proof: Let $\alpha = \lambda/\sqrt{m}$. Let $Y_i = X_i - X_{i-1}$, so that $|Y_i| \le 1$ and $\mathbf{E}\Big[Y_i \;\Big|\; X_0, \ldots, X_{i-1}\Big] = 0$.

We are interested in bounding $\mathbf{E}\Big[e^{\alpha Y_i} \;\Big|\; X_0, \ldots, X_{i-1}\Big]$. Note that, for $-1 \le x \le 1$, we have

$$e^{\alpha x} \le h(x) = \frac{e^{\alpha} + e^{-\alpha}}{2} + \frac{e^{\alpha} - e^{-\alpha}}{2} x,$$

as $e^{\alpha x}$ is a convex function, $h(-1) = e^{-\alpha}$, $h(1) = e^{\alpha}$, and $h(x)$ is a linear function. Thus,

$$\begin{aligned}
\mathbf{E}\Big[e^{\alpha Y_i} \;\Big|\; X_0, \ldots, X_{i-1}\Big] &\le \mathbf{E}\Big[h(Y_i) \;\Big|\; X_0, \ldots, X_{i-1}\Big] = h\Big(\mathbf{E}\Big[Y_i \;\Big|\; X_0, \ldots, X_{i-1}\Big]\Big) \\
&= h\Big(0\Big) = \frac{e^{\alpha} + e^{-\alpha}}{2} \\
&= \frac{(1 + \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \cdots) + (1 - \alpha + \frac{\alpha^2}{2!} - \frac{\alpha^3}{3!} + \cdots)}{2} \\
&= 1 + \frac{\alpha^2}{2} + \frac{\alpha^4}{4!} + \frac{\alpha^6}{6!} + \cdots \\
&\le 1 + \frac{1}{1!}\left(\frac{\alpha^2}{2}\right) + \frac{1}{2!}\left(\frac{\alpha^2}{2}\right)^2 + \frac{1}{3!}\left(\frac{\alpha^2}{2}\right)^3 + \cdots = \exp\Big(\alpha^2/2\Big),
\end{aligned}$$

as $(2i)! \ge 2^i i!$.

Hence, by Lemma 8.1.3, we have that

$$\begin{aligned}
\mathbf{E}\Big[e^{\alpha X_m}\Big] &= \mathbf{E}\left[\prod_{i=1}^{m} e^{\alpha Y_i}\right] = \mathbf{E}\left[\left(\prod_{i=1}^{m-1} e^{\alpha Y_i}\right) e^{\alpha Y_m}\right] \\
&= \mathbf{E}\left[\left(\prod_{i=1}^{m-1} e^{\alpha Y_i}\right) \mathbf{E}\Big[e^{\alpha Y_m} \;\Big|\; X_0, \ldots, X_{m-1}\Big]\right] \le e^{\alpha^2/2}\, \mathbf{E}\left[\prod_{i=1}^{m-1} e^{\alpha Y_i}\right] \\
&\le \exp\Big(m\alpha^2/2\Big).
\end{aligned}$$

Therefore, by Markov's inequality, we have

$$\begin{aligned}
\mathbf{Pr}\Big[X_m > \lambda\sqrt{m}\Big] = \mathbf{Pr}\Big[e^{\alpha X_m} > e^{\alpha\lambda\sqrt{m}}\Big] &= \frac{\mathbf{E}\Big[e^{\alpha X_m}\Big]}{e^{\alpha\lambda\sqrt{m}}} = e^{m\alpha^2/2 - \alpha\lambda\sqrt{m}} \\
&= \exp\Big(m(\lambda/\sqrt{m})^2/2 - (\lambda/\sqrt{m})\lambda\sqrt{m}\Big) = e^{-\lambda^2/2},
\end{aligned}$$

implying the result. ∎

Here is an alternative form.

Theorem 8.1.12 (Azuma's Inequality). *Let $X_0, \ldots, X_m$ be a martingale sequence such that and $|X_{i+1} - X_i| \le 1$ for all $0 \le i < m$. Let $\lambda > 0$ be arbitrary. Then $\mathbf{Pr}\Big[|X_m - X_0| > \lambda\sqrt{m}\Big] < 2\exp\Big(-\lambda^2/2\Big)$.*

Example 8.1.13. Let $\chi(H)$ be the chromatic number of a graph H. What is chromatic number of a random graph? How does this random variable behaves?

Consider the vertex exposure martingale, and let $X_i = \mathbf{E}\Big[\chi(G) \;\Big|\; G_i\Big]$. Again, without proving it, we claim that $X_0, \ldots, X_n = X$ is a martingale, and as such, we have: $\mathbf{Pr}\Big[|X_n - X_0| > \lambda\sqrt{n}\Big] \le e^{-\lambda^2/2}$. However, $X_0 = \mathbf{E}[\chi(G)]$, and $X_n = \mathbf{E}\Big[\chi(G) \;\Big|\; G_n\Big] = \chi(G)$. Thus,

$$\mathbf{Pr}\Big[\Big|\chi(G) - \mathbf{E}[\chi(G)]\Big| > \lambda\sqrt{n}\Big] \le e^{-\lambda^2/2}.$$

Namely, the chromatic number of a random graph is highly concentrated! And we do not even know what is the expectation of this variable!

Chapter 9

Martingales II

By Sariel Har-Peled, December 30, 2015[①]

> "The Electric Monk was a labor-saving device, like a dishwasher or a video recorder. Dishwashers washed tedious dishes for you, thus saving you the bother of washing them yourself, video recorders watched tedious television for you, thus saving you the bother of looking at it yourself; Electric Monks believed things for you, thus saving you what was becoming an increasingly onerous task, that of believing all the things the world expected you to believe."
>
> – — Dirk Gently's Holistic Detective Agency, Douglas Adams..

9.1. Filters and Martingales

Definition 9.1.1. A *σ-field* $(\Omega, \mathcal{F})$ consists of a sample space Ω (i.e., the atomic events) and a collection of subsets $\mathcal{F}$ satisfying the following conditions:

(A) $\emptyset \in \mathcal{F}$.
(B) $C \in \mathcal{F} \Rightarrow \overline{C} \in \mathcal{F}$.
(C) $C_1, C_2, \ldots \in \mathcal{F} \Rightarrow C_1 \cup C_2 \ldots \in \mathcal{F}$.

Definition 9.1.2. Given a σ-field $(\Omega, \mathcal{F})$, a *probability measure* $\mathbf{Pr} : \mathcal{F} \to \mathbb{R}^+$ is a function that satisfies the following conditions.

(A) $\forall A \in \mathcal{F}$, $0 \leq \mathbf{Pr}[A] \leq 1$.
(B) $\mathbf{Pr}[\Omega] = 1$.
(C) For mutually disjoint events $C_1, C_2, \ldots$, we have $\mathbf{Pr}\left[\cup_i C_i\right] = \sum_i \mathbf{Pr}\left[C_i\right]$.

Definition 9.1.3. A *probability space* $(\Omega, \mathcal{F}, \mathbf{Pr})$ consists of a σ-field $(\Omega, \mathcal{F})$ with a probability measure $\mathbf{Pr}$ defined on it.

Definition 9.1.4. Given a σ-field $(\Omega, \mathcal{F})$ with $\mathcal{F} = 2^\Omega$, a *filter* (also *filtration*) is a nested sequence $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ of subsets of 2^Ω, such that:

(A) $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
(B) $\mathcal{F}_n = 2^\Omega$.
(C) For $0 \leq i \leq n$, $(\Omega, \mathcal{F}_i)$ is a σ-field.

Definition 9.1.5. An *elementary event* or *atomic event* is a subset of a sample space that contains only one element of Ω.

Intuitively, when we consider a probability space, we usually consider a random variable X. The value of X is a function of the elementary event that happens in the probability space. Formally, a random variable is a mapping $X : \Omega \to \mathbb{R}$. Thus, each $\mathcal{F}_i$ defines a partition of Ω into *atomic events*. This partition is getting more and more refined as we progress down the filter.

Example 9.1.6. Consider an algorithm **Alg** that uses n random bits. As such, the underlying sample space is $\Omega = \left\{ b_1 b_2 \ldots b_n \;\middle|\; b_1, \ldots, b_n \in \{0,1\} \right\}$; that is, the set of all binary strings of length n. Next, let $\mathcal{F}_i$ be the σ-field generated by the partition of Ω into the atomic events B_w, where $w \in \{0,1\}^i$; here w is the string encoding the first i random bits used by the algorithm. Specifically,

$$B_w = \left\{ wx \;\middle|\; x \in \{0,1\}^{n-i} \right\},$$

and the set of atomic events in $\mathcal{F}_i$ is $\left\{ B_w \;\middle|\; w \in \{0,1\}^i \right\}$. The set $\mathcal{F}_i$ is the closure of this set of atomic events under complement and union. In particular, we conclude that $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_n$ form a filter.

Definition 9.1.7. A random variable X is said to be *$\mathcal{F}_i$-measurable* if for each $x \in \mathbb{R}$, the event $X \le x$ is in $\mathcal{F}_i$; that is, the set $\left\{ \omega \in \Omega \;\middle|\; X(\omega) \le x \right\}$ is in $\mathcal{F}_i$.

Example 9.1.8. Let $\mathcal{F}_0, \ldots, \mathcal{F}_n$ be the filter defined in Example 9.1.6. Let X be the parity of the n bits. Clearly, $X = 1$ is a valid event only in $\mathcal{F}_n$ (why?). Namely, it is only measurable in $\mathcal{F}_n$, but not in $\mathcal{F}_i$, for $i < n$.

As such, a random variable X is $\mathcal{F}_i$-measurable, only if it is a constant on the elementary events of $\mathcal{F}_i$. This gives us a new interpretation of what a filter is – its a sequence of refinements of the underlying probability space, that is achieved by splitting the atomic events of $\mathcal{F}_i$ into smaller atomic events in $\mathcal{F}_{i+1}$. Putting it explicitly, an atomic event $\mathcal{E}$ of $\mathcal{F}_i$, is a subset of 2^Σ. As we move to $\mathcal{F}_{i+1}$ the event $\mathcal{E}$ might now be split into several atomic (and disjoint events) $\mathcal{E}_1, \ldots, \mathcal{E}_k$. Now, naturally, the atomic event that really happens is an atomic event of $\mathcal{F}_n$. As we progress down the filter, we "zoom" into this event.

Definition 9.1.9 (Conditional expectation in a filter). Let $(\Omega, \mathcal{F})$ be any σ-field, and Y any random variable that takes on distinct values on the elementary events in $\mathcal{F}$. Then $\mathbf{E}[X|\mathcal{F}] = \mathbf{E}[X \mid Y]$.

9.2. Martingales

Definition 9.2.1. A sequence of random variables $Y_1, Y_2, \ldots$, is said to be a *martingale difference* sequence if for all $i \ge 0$, we have $\mathbf{E}\left[Y_i \;\middle|\; Y_1, \ldots, Y_{i-1} \right] = 0$.

Clearly, $X_1, \ldots,$ is a martingale sequence if and only if $Y_1, Y_2, \ldots,$ is a martingale difference sequence where $Y_i = X_i - X_{i-1}$.

Definition 9.2.2. A sequence of random variables $Y_1, Y_2, \ldots,$ is

$$\text{a } \textit{super martingale} \text{ sequence if} \quad \forall i \quad \mathbf{E}\left[Y_i \;\middle|\; Y_1, \ldots, Y_{i-1} \right] \le Y_{i-1},$$
$$\text{and a } \textit{sub martingale} \text{ sequence if} \quad \forall i \quad \mathbf{E}\left[Y_i \;\middle|\; Y_1, \ldots, Y_{i-1} \right] \ge Y_{i-1}.$$

9.2.1. Martingales – an alternative definition

Definition 9.2.3. Let $(\Omega, \mathcal{F}, \mathbf{Pr})$ be a probability space with a filter $\mathcal{F}_0, \mathcal{F}_1, \ldots$. Suppose that $X_0, X_1, \ldots$, are random variables such that, for all $i \geq 0$, X_i is $\mathcal{F}_i$-measurable. The sequence $X_0, \ldots, X_n$ is a *martingale* provided that, for all $i \geq 0$, we have $\mathbf{E}[X_{i+1} \mid \mathcal{F}_i] = X_i$.

Lemma 9.2.4. *Let $(\Omega, \mathcal{F})$ and $(\Omega, \mathcal{G})$ be two σ-fields such that $\mathcal{F} \subseteq \mathcal{G}$. Then, for any random variable X, $\mathbf{E}\Big[\mathbf{E}\Big[X \,\Big|\, \mathcal{G}\Big] \,\Big|\, \mathcal{F}\Big] = \mathbf{E}\Big[X \,\Big|\, \mathcal{F}\Big]$.*

Proof: $\mathbf{E}\Big[\mathbf{E}\Big[X \,\Big|\, \mathcal{G}\Big] \,\Big|\, \mathcal{F}\Big] = \mathbf{E}\Big[\mathbf{E}\Big[X \,\Big|\, G = g\Big] \,\Big|\, F = f\Big]$

$$
\begin{aligned}
&= \mathbf{E}\left[\frac{\sum_x x \mathbf{Pr}[X = x \cap G = g]}{\mathbf{Pr}[G = g]} \,\Big|\, F = f\right] = \sum_{g \in G} \frac{\frac{\sum_x x \mathbf{Pr}[X = x \cap G = g]}{\mathbf{Pr}[G = g]} \cdot \mathbf{Pr}[G = g \cap F = f]}{\mathbf{Pr}[F = f]} \\
&= \sum_{g \in \mathcal{G}, g \subseteq f} \frac{\frac{\sum_x x \mathbf{Pr}[X = x \cap G = g]}{\mathbf{Pr}[G = g]} \cdot \mathbf{Pr}[G = g \cap F = f]}{\mathbf{Pr}[F = f]} = \sum_{g \in \mathcal{G}, g \subseteq f} \frac{\frac{\sum_x x \mathbf{Pr}[X = x \cap G = g]}{\mathbf{Pr}[G = g]} \cdot \mathbf{Pr}[G = g]}{\mathbf{Pr}[F = f]} \\
&= \sum_{g \in \mathcal{G}, g \subseteq f} \frac{\sum_x x \mathbf{Pr}[X = x \cap G = g]}{\mathbf{Pr}[F = f]} = \frac{\sum_x x \Big(\sum_{g \in \mathcal{G}, g \subseteq f} \mathbf{Pr}[X = x \cap G = g]\Big)}{\mathbf{Pr}[F = f]} \\
&= \frac{\sum_x x \mathbf{Pr}[X = x \cap F = f]}{\mathbf{Pr}[F = f]} = \mathbf{E}\Big[X \,\Big|\, \mathcal{F}\Big]. \qquad \blacksquare
\end{aligned}
$$

Theorem 9.2.5. *Let $(\Omega, \mathcal{F}, \mathbf{Pr})$ be a probability space, and let $\mathcal{F}_0, \ldots, \mathcal{F}_n$ be a filter with respect to it. Let X be any random variable over this probability space and define $X_i = \mathbf{E}\Big[X \,\Big|\, \mathcal{F}_i\Big]$ then, the sequence $X_0, \ldots, X_n$ is a martingale.*

Proof: We need to show that $\mathbf{E}\Big[X_{i+1} \,\Big|\, \mathcal{F}_i\Big] = X_i$. Namely,

$$\mathbf{E}\Big[X_{i+1} \,\Big|\, \mathcal{F}_i\Big] = \mathbf{E}\Big[\mathbf{E}\Big[X \,\Big|\, \mathcal{F}_{i+1}\Big] \,\Big|\, \mathcal{F}_i\Big] = \mathbf{E}\Big[X \,\Big|\, \mathcal{F}_i\Big] = X_i,$$

by Lemma 9.2.4 and by definition of X_i. $\blacksquare$

Definition 9.2.6. Let $f : \mathcal{D}_1 \times \cdots \times \mathcal{D}_n \to \mathbb{R}$ be a real-valued function with a arguments from possibly distinct domains. The function f is said to satisfy the *Lipschitz condition* if for any $x_1 \in \mathcal{D}_1, \ldots, x_n \in \mathcal{D}_n$, and $i \in \{1, \ldots, n\}$ and any $y_i \in \mathcal{D}_i$,

$$\Big| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n) \Big| \leq 1.$$

Specifically, a function is *c-Lipschitz*, if the inequality holds with a constant c (instead of 1).

Definition 9.2.7. Let $X_1, \ldots, X_n$ be a sequence of *independent* random variables, and a function $f(X_1, \ldots, X_n)$ defined over them that such that f satisfies the Lipschitz condition. The *Doob martingale* sequence $Y_0, \ldots, Y_m$ is defined by $Y_0 = \mathbf{E}\Big[f(X_1, \ldots, X_n)\Big]$ and $Y_i = \mathbf{E}\Big[f(X_1, \ldots, X_n) \,\Big|\, X_1, \ldots, X_i\Big]$, for $i = 1, \ldots, n$.

Clearly, a Doob martingale $Y_0, \ldots, Y_n$ is a martingale, by Theorem 9.2.5. Furthermore, if $|X_i - X_{i-1}| \leq 1$, for $i = 1, \ldots, n$, then $|X_i - X_{i-1}| \leq 1$. and we can use Azuma's inequality on such a sequence.

9.3. Occupancy Revisited

We have m balls thrown independently and uniformly into n bins. Let Z denote the number of bins that remains empty in the end of the process. Let X_i be the bin chosen in the ith trial, and let $Z = F(X_1, \ldots, X_m)$, where F returns the number of empty bins given that m balls had thrown into bins $X_1, \ldots, X_m$. Clearly, we have by Azuma's inequality that $\mathbf{Pr}\Big[\big|Z - \mathbf{E}[Z]\big| > \lambda\sqrt{m}\Big] \leq 2e^{-\lambda^2/2}$.

The following is an extension of Azuma's inequality shown in class. We do not provide a proof but it is similar to what we saw.

Theorem 9.3.1 (Azuma's Inequality - Stronger Form). *Let $X_0, X_1, \ldots,$ be a martingale sequence such that for each k, $|X_k - X_{k-1}| \leq c_k$, where c_k may depend on k. Then, for all $t \geq 0$, and any $\lambda > 0$,*

$$\mathbf{Pr}\Big[|X_t - X_0| \geq \lambda\Big] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_{k=1}^{t} c_k^2}\right).$$

Theorem 9.3.2. *Let $r = m/n$, and Z_{end} be the number of empty bins when m balls are thrown randomly into n bins. Then $\mu = \mathbf{E}\Big[Z_{end}\Big] = n\Big(1 - \frac{1}{n}\Big)^m \approx ne^{-r}$, and for any $\lambda > 0$, we have*

$$\mathbf{Pr}\Big[\big|Z_{end} - \mu\big| \geq \lambda\Big] \leq 2\exp\left(-\frac{\lambda^2(n - 1/2)}{n^2 - \mu^2}\right).$$

Proof: Let $z(Y, t)$ be the expected number of empty bins, if there are Y empty bins in time t. Clearly,

$$z(Y, t) = Y\left(1 - \frac{1}{n}\right)^{m-t}.$$

In particular, $\mu = z(n, 0) = n\Big(1 - \frac{1}{n}\Big)^m$.

Let $\mathcal{F}_t$ be the σ-field generated by the bins chosen in the first t steps. Let Z_{end} be the number of empty bins at time m, and let $Z_t = \mathbf{E}\Big[Z_{\text{end}} \,\Big|\, \mathcal{F}_t\Big]$. Namely, Z_t is the expected number of empty bins after we know where the first t balls had been placed. The random variables $Z_0, Z_1, \ldots, Z_m$ form a martingale. Let Y_t be the number of empty bins after t balls where thrown. We have $Z_{t-1} = z(Y_{t-1}, t-1)$. Consider the ball thrown in the t-step. Clearly:

(A) With probability $1 - Y_{t-1}/n$ the ball falls into a non-empty bin. Then $Y_t = Y_{t-1}$, and $Z_t = z(Y_{t-1}, t)$. Thus,

$$\Delta_t = Z_t - Z_{t-1} = z(Y_{t-1}, t) - z(Y_{t-1}, t-1) = Y_{t-1}\left(\left(1 - \frac{1}{n}\right)^{m-t} - \left(1 - \frac{1}{n}\right)^{m-t+1}\right) = \frac{Y_{t-1}}{n}\left(1 - \frac{1}{n}\right)^{m-t} \leq \left(1 - \frac{1}{n}\right)^{m-t}.$$

(B) Otherwise, with probability Y_{t-1}/n the ball falls into an empty bin, and $Y_t = Y_{t-1} - 1$. Namely, $Z_t = z(Y_t - 1, t)$. And we have that

$$\begin{aligned}
\Delta_t = Z_t - Z_{t-1} &= z(Y_{t-1} - 1, t) - z(Y_{t-1}, t-1) = (Y_{t-1} - 1)\left(1 - \frac{1}{n}\right)^{m-t} - Y_{t-1}\left(1 - \frac{1}{n}\right)^{m-t+1} \\
&= \left(1 - \frac{1}{n}\right)^{m-t}\left(Y_{t-1} - 1 - Y_{t-1}\left(1 - \frac{1}{n}\right)\right) = \left(1 - \frac{1}{n}\right)^{m-t}\left(-1 + \frac{Y_{t-1}}{n}\right) = -\left(1 - \frac{1}{n}\right)^{m-t}\left(1 - \frac{Y_{t-1}}{n}\right) \\
&\geq -\left(1 - \frac{1}{n}\right)^{m-t}.
\end{aligned}$$

Thus, $Z_0, \ldots, Z_m$ is a martingale sequence, where $|Z_t - Z_{t-1}| \le |\Delta_t| \le c_t$, where $c_t = \left(1 - \frac{1}{n}\right)^{m-t}$. We have

$$\sum_{t=1}^{n} c_t^2 = \frac{1-(1-1/n)^{2m}}{1-(1-1/n)^2} = \frac{n^2\left(1-(1-1/n)^{2m}\right)}{2n-1} = \frac{n^2-\mu^2}{2n-1}.$$

Now, deploying Azuma's inequality, yield the result. ■

9.3.1. Lets verify this is indeed an improvement

Consider the case where $m = n \ln n$. Then, $\mu = n\left(1 - \frac{1}{n}\right)^m \le 1$. And using the "weak" Azuma's inequality implies that

$$\mathbf{Pr}\left[\left|Z_{\text{end}} - \mu\right| \ge \lambda\sqrt{n}\right] = \mathbf{Pr}\left[\left|Z_{\text{end}} - \mu\right| \ge \lambda\sqrt{\frac{n}{m}}\sqrt{m}\right] \le 2\exp\left(-\frac{\lambda^2 n}{2m}\right) = 2\exp\left(-\frac{\lambda^2}{2\ln n}\right),$$

which is interesting only if $\lambda > \sqrt{2\ln n}$. On the other hand, Theorem 9.3.2 implies that

$$\mathbf{Pr}\left[\left|Z_{\text{end}} - \mu\right| \ge \lambda\sqrt{n}\right] \le 2\exp\left(-\frac{\lambda^2 n(n-1/2)}{n^2-\mu^2}\right) \le 2\exp\left(-\lambda^2\right),$$

which is interesting for any $\lambda \ge 1$ (say).

9.4. Some useful estimates

Lemma 9.4.1. *For any $n \ge 2$, and $m \ge 1$, we have that $(1-1/n)^m \ge 1 - m/n$.*

Proof: Follows by induction. Indeed, for $m = 1$ the claim is immediate. For $m \ge 2$, we have

$$\left(1-\frac{1}{n}\right)^m = \left(1-\frac{1}{n}\right)\left(1-\frac{1}{n}\right)^{m-1} \ge \left(1-\frac{1}{n}\right)\left(1-\frac{m-1}{n}\right) \ge 1 - \frac{m}{n}.$$ ■

This implies the following.

Lemma 9.4.2. *For any $m \le n$, we have that $1 - m/n \le (1-1/n)^m \le \exp(-m/n)$.*

Chapter 10

The Probabilistic Method

By Sariel Har-Peled, December 30, 2015[①]

> "Shortly after the celebration of the four thousandth anniversary of the opening of space, Angary J. Gustible discovered Gustible's planet. The discovery turned out to be a tragic mistake.
>
> Gustible's planet was inhabited by highly intelligent life forms. They had moderate telepathic powers. They immediately mind-read Angary J. Gustible's entire mind and life history, and embarrassed him very deeply by making up an opera concerning his recent divorce."
>
> – – From Gustible's Planet, Cordwainer Smith.

10.1. Introduction

The probabilistic method is a combinatorial technique to use probabilistic algorithms to create objects having desirable properties, and furthermore, prove that such objects exist. The basic technique is based on two basic observations:

1. If $\mathbf{E}[X] = \mu$, then there exists a value x of X, such that $x \geq \mathbf{E}[X]$.

2. If the probability of event $\mathcal{E}$ is larger than zero, then $\mathcal{E}$ exists and it is not empty.

The surprising thing is that despite the elementary nature of those two observations, they lead to a powerful technique that leads to numerous nice and strong results. Including some elementary proofs of theorems that previously had very complicated and involved proofs.

The main proponent of the probabilistic method, was Paul Erdős. An excellent text on the topic is the book by Noga Alon and Joel Spencer [AS00].

This topic is worthy of its own course. The interested student is refereed to the course "Math 475 — The Probabilistic Method".

10.1.1. Examples

Theorem 10.1.1. *For any undirected graph $G(V, E)$ with n vertices and m edges, there is a partition of the vertex set V into two sets A and B such that*

$$\left|\left\{uv \in E \;\middle|\; u \in A \text{ and } v \in B\right\}\right| \geq \frac{m}{2}.$$

Proof: Consider the following experiment: randomly assign each vertex to A or B, independently and equal probability.

For an edge $e = uv$, the probability that one endpoint is in A, and the other in B is $1/2$, and let X_e be the indicator variable with value 1 if this happens. Clearly,

$$\mathbf{E}\left[\left|\left\{uv \in E \;\middle|\; (u, v) \in (A \times B) \cup (B \times A)\right\}\right|\right] = \sum_{e \in E(G)} \mathbf{E}[X_e] = \sum_{e \in E(G)} \frac{1}{2} = \frac{m}{2}.$$

Thus, there must be a partition of V that satisfies the theorem. ■

Definition 10.1.2. For a vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$, $\|v\|_\infty = \max_i |v_i|$.

Theorem 10.1.3. *Let M be an $n \times n$ binary matrix (i.e., each entry is either 0 or 1), then there always exists a vector $b \in \{-1, +1\}^n$ such that $\|\mathsf{M}b\|_\infty \leq 4\sqrt{n \log n}$.*

Proof: Let $v = (v_1, \ldots, v_n)$ be a row of M. Chose a random $b = (b_1, \ldots, b_n) \in \{-1, +1\}^n$. Let $i_1, \ldots, i_m$ be the indices such that $v_{i_j} = 1$, and let

$$Y = \langle v, b \rangle = \sum_{i=1}^{n} v_i b_i = \sum_{j=1}^{m} v_{i_j} b_{i_j} = \sum_{j=1}^{m} b_{i_j}.$$

As such Y is the sum of m independent random variables that accept values in $\{-1, +1\}$. Clearly,

$$\mathbf{E}[Y] = \mathbf{E}[\langle v, b \rangle] = \mathbf{E}\left[\sum_i v_i b_i\right] = \sum_i \mathbf{E}[v_i b_i] = \sum_i v_i \,\mathbf{E}[b_i] = 0.$$

By Chernoff inequality (Theorem 10.3.1) and the symmetry of Y, we have that, for $\Delta = 4\sqrt{n \ln n}$, it holds

$$\mathbf{Pr}[|Y| \geq \Delta] = 2\,\mathbf{Pr}[v \cdot b \geq \Delta] = 2\,\mathbf{Pr}\left[\sum_{j=1}^{m} b_{i_j} \geq \Delta\right] \leq 2\exp\left(-\frac{\Delta^2}{2m}\right) = 2\exp\left(-8\frac{n \ln n}{m}\right) \leq \frac{2}{n^8}.$$

Thus, the probability that any entry in $\mathsf{M}b$ exceeds $4\sqrt{n \ln n}$ is smaller than $2/n^7$. Thus, with probability at least $1 - 2/n^7$, all the entries of $\mathsf{M}b$ have value smaller than $4\sqrt{n \ln n}$.

In particular, there exists a vector $b \in \{-1, +1\}^n$ such that $\|\,\mathsf{M}b\,\|_\infty \leq 4\sqrt{n \ln n}$. ■

10.2. Maximum Satisfiability

In the MAX-SAT problem, we are given a binary formula F in *[CNF]* (Conjunctive normal form), and we would like to find an assignment that satisfies as many clauses as possible of F, for example $F = (x \vee y) \wedge (\overline{x} \vee z)$. Of course, an assignment satisfying all the clauses of the formula, and thus F itself, would be even better – but this problem is of course NPC. As such, we are looking for how well can be we do when we relax the problem to maximizing the number of clauses to be satisfied..

Theorem 10.2.1. *For any set of m clauses, there is a truth assignment of variables that satisfies at least $m/2$ clauses.*

Proof: Assign every variable a random value. Clearly, a clause with k variables, has probability $1 - 2^{-k}$ to be satisfied. Using linearity of expectation, and the fact that every clause has at least one variable, it follows, that $\mathbf{E}[X] = m/2$, where X is the random variable counting the number of clauses being satisfied. In particular, there exists an assignment for which $X \geq m/2$. ■

For an instant I, let $m_{\text{opt}}(I)$, denote the maximum number of clauses that can be satisfied by the "best" assignment. For an algorithm **Alg**, let $m_{\textbf{Alg}}(I)$ denote the number of clauses satisfied computed by the algorithm **Alg**. The *approximation factor* of **Alg**, is $m_{\textbf{Alg}}(I)/m_{\text{opt}}(I)$. Clearly, the algorithm of Theorem 10.2.1 provides us with $1/2$-approximation algorithm.

For every clause, C_j in the given instance, let $z_j \in \{0, 1\}$ be a variable indicating whether C_j is satisfied or not. Similarly, let $x_i = 1$ if the ith variable is being assigned the value TRUE. Let C_j^+ be indices of the variables that appear in C_j in the positive, and C_j^- the indices of the variables that appear in the negative. Clearly, to solve MAX-SAT, we need to solve:

$$\begin{aligned} \textit{maximize} \quad & \sum_{j=1}^{m} z_j \\ \textit{subject to} \quad & x_i, z_j \in \{0,1\} \text{ for all } i, j \\ & \sum_{i \in C_j^+} x_i + \sum_{i \in C_j^-} (1 - x_i) \geq z_j \text{ for all } j. \end{aligned}$$

We relax this into the following linear program:

$$\begin{aligned} \textit{maximize} \quad & \sum_{j=1}^{m} z_j \\ \textit{subject to} \quad & 0 \leq y_i, z_j \leq 1 \text{ for all } i, j \\ & \sum_{i \in C_j^+} y_i + \sum_{i \in C_j^-} (1 - y_i) \geq z_j \text{ for all } j. \end{aligned}$$

Which can be solved in polynomial time. Let $\widehat{t}$ denote the values assigned to the variable t by the linear-programming solution. Clearly, $\sum_{j=1}^{m} \widehat{z_j}$ is an upper bound on the number of clauses of I that can be satisfied.

We set the variable y_i to 1 with probability $\widehat{y_i}$. This is *randomized rounding*.

Lemma 10.2.2. *Let C_j be a clause with k literals. The probability that it is satisfied by randomized rounding is at least $\beta_k \widehat{z_j} \geq (1 - 1/e)\widehat{z_j}$, where*

$$\beta_k = 1 - \left(1 - \frac{1}{k}\right)^k.$$

Proof: Assume $C_j = y_1 \vee v_2 \ldots \vee v_k$. By the LP, we have $\widehat{y_1} + \cdots + \widehat{y_k} \geq \widehat{z_j}$. Furthermore, the probability that C_j is not satisfied is $\prod_{i=1}^{k}(1-\widehat{y_i})$. Note that $1 - \prod_{i=1}^{k}(1-\widehat{y_i})$ is minimized when all the $\widehat{y_i}$'s are equal (by symmetry). Namely, when $\widehat{y_i} = \widehat{z_j}/k$. Consider the function $f(x) = 1-(1-x/k)^k$. This is a concave function, which is larger than $g(x) = \beta_k x$ for all $0 \leq x \leq 1$, as can be easily verified, by checking the inequality at $x = 0$ and $x = 1$.

Thus,

$$\mathbf{Pr}\Big[C_j \text{ is satisfied}\Big] = 1 - \prod_{i=1}^{k}(1-\widehat{y_i}) \geq f\left(\widehat{z_j}\right) \geq \beta_k \widehat{z_j}.$$

The second part of the inequality, follows from the fact that $\beta_k \geq 1 - 1/e$, for all $k \geq 0$. Indeed, for $k = 1, 2$ the claim trivially holds. Furthermore,

$$1 - \left(1 - \frac{1}{k}\right)^k \geq 1 - \frac{1}{e} \quad \Leftrightarrow \quad \left(1 - \frac{1}{k}\right)^k \leq \frac{1}{e},$$

but this holds since $1 - x \leq e^{-x}$ implies that $1 - \frac{1}{k} \leq e^{-1/k}$, and as such $\left(1 - \frac{1}{k}\right)^k \leq e^{-k/k} = 1/e$. ■

Theorem 10.2.3. *Given an instance I of MAX-SAT, the expected number of clauses satisfied by linear programming and randomized rounding is at least $(1-1/e) \approx 0.632 m_{\text{opt}}(I)$, where $m_{\text{opt}}(I)$ is the maximum number of clauses that can be satisfied on that instance.*

Theorem 10.2.4. *Given an instance I of MAX-SAT, let n_1 be the expected number of clauses satisfied by randomized assignment, and let n_2 be the expected number of clauses satisfied by linear programming followed by randomized rounding. Then,* $\max(n_1, n_2) \geq (3/4)\sum_j \widehat{z_j} \geq (3/4) m_{\text{opt}}(I)$.

Proof: It is enough to show that $(n_1 + n_2)/2 \geq \frac{3}{4}\sum_j \widehat{z_j}$. Let S_k denote the set of clauses that contain k literals. We know that

$$n_1 = \sum_k \sum_{C_j \in S_k} \left(1 - 2^{-k}\right) \geq \sum_k \sum_{C_j \in S_k} \left(1 - 2^{-k}\right)\widehat{z_j}.$$

By Lemma 10.2.2 we have $n_2 \geq \sum_k \sum_{C_j \in S_k} \beta_k \widehat{z_j}$. Thus,

$$\frac{n_1 + n_2}{2} \geq \sum_k \sum_{C_j \in S_k} \frac{1 - 2^{-k} + \beta_k}{2} \widehat{z_j}.$$

One can verify that $\left(1 - 2^{-k}\right) + \beta_k \geq 3/2$, for all k. [2] Thus, we have

$$\frac{n_1 + n_2}{2} \geq \frac{3}{4} \sum_k \sum_{C_j \in S_k} \widehat{z_j} = \frac{3}{4} \sum_j \widehat{z_j}.$$

■

[2] Indeed, by the proof of Lemma 10.2.2, we have that $\beta_k \geq 1 - 1/e$. Thus, $\left(1 - 2^{-k}\right) + \beta_k \geq 2 - 1/e - 2^{-k} \geq 3/2$ for $k \geq 3$. Thus, we only need to check the inequality for $k = 1$ and $k = 2$, which can be done directly.

10.3. From previous lectures

Theorem 10.3.1. *Let $X_1, \ldots, X_n$ be n independent random variables, such that $\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[X_i = -1] = \frac{1}{2}$, for $i = 1, \ldots, n$. Let $Y = \sum_{i=1}^{n} X_i$. Then, for any $\Delta > 0$, we have*

$$\mathbf{Pr}\Big[Y \geq \Delta\Big] \leq \exp\Big(-\Delta^2/2n\Big).$$

Bibliography

[AS00] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley InterScience, 2nd edition, 2000.

Chapter 11

The Probabilistic Method II

By Sariel Har-Peled, December 30, 2015[1]

> "Today I know that everything watches, that nothing goes unseen, and that even wallpaper has a better memory than ours. It isn't God in His heaven that sees all. A kitchen chair, a coat-hanger a half-filled ash tray, or the wood replica of a woman name Niobe, can perfectly well serve as an unforgetting witness to every one of our acts."
>
> – Gunter Grass, The tin drum.

11.1. Expanding Graphs

In this lecture, we are going to discuss *expanding graphs*.

Definition 11.1.1. An (n, d, α, c) *OR-concentrator* is a bipartite multigraph $\mathsf{G}(L, R, E)$, with the independent sets of vertices L and R each of cardinality n, such that

(i) Every vertex in L has degree at most d.

(ii) Any subset S of vertices of L, with $|S| \leq \alpha n$ has at least $c\,|S|$ neighbors in R.

A good (n, d, α, c) OR-concentrator should have d as small as possible[2], and c as large as possible.

Theorem 11.1.2. *There is an integer n_0, such that for all $n \geq n_0$, there is an $(n, 18, 1/3, 2)$ OR-concentrator.*

Proof: Let every vertex of L choose neighbors by sampling (with replacement) d vertices independently and uniformly from R. We discard multiple parallel edges in the resulting graph.

Let $\mathcal{E}_s$ be the event that a subset of s vertices of L has fewer than cs neighbors in R. Clearly,

$$\mathbf{Pr}\left[\mathcal{E}_s\right] \leq \binom{n}{s}\binom{n}{cs}\left(\frac{cs}{n}\right)^{ds} \leq \left(\frac{ne}{s}\right)^{s}\left(\frac{ne}{cs}\right)^{cs}\left(\frac{cs}{n}\right)^{ds} = \left(\left(\frac{s}{n}\right)^{d-c-1}\exp(1+c)c^{d-c}\right)^{s},$$

[2] Or smaller!

since $\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$. Setting $\alpha = 1/3$ using $s \leq \alpha n$, and $c = 2$, we have

$$\begin{aligned}\mathbf{Pr}\left[\mathcal{E}_s\right] &\leq \left(\left(\frac{1}{3}\right)^{d-c-1} e^{1+c} c^{d-c}\right)^s \leq \left(\left(\frac{1}{3}\right)^d 3^{1+c} e^{1+c} c^{d-c}\right)^s \leq \left(\left(\frac{1}{3}\right)^d 3^{1+c} e^{1+c} c^d\right)^s \\ &\leq \left(\left(\frac{c}{3}\right)^d (3e)^{1+c}\right)^s \leq \left(\left(\frac{2}{3}\right)^{18} (3e)^{1+2}\right)^s \leq \left(0.4\right)^s,\end{aligned}$$

as $c = 2$ and $d = 18$. Thus,

$$\sum_{s\geq 1} \mathbf{Pr}\left[\mathcal{E}_s\right] \leq \sum_{s\geq 1} (0.4)^s < 1.$$

It thus follows that the random graph we generated has the required properties with positive probability. ■

11.2. Probability Amplification

Let **Alg** be an algorithm in **RP**, such that given x, **Alg** picks a random number r from the range $\mathbb{Z}_n = \{0, \ldots, n-1\}$, for a suitable choice of a prime n, and computes a binary value **Alg**(x, r) with the following properties:

(A) If $x \in L$, then **Alg**$(x, r) = 1$ for at least half the possible values of r.

(B) If $x \notin L$, then **Alg**$(x, r) = 0$ for all possible choices of r.

Next, we show that using $\lg^2 n$ bits[③] one can achieve $1/n^{\lg n}$ confidence, compared with the naive $1/n$, and the $1/t$ confidence achieved by t (dependent) executions of the algorithm using two-point sampling.

Theorem 11.2.1. *For n large enough, there exists a bipartite graph $\mathsf{G}(V, R, E)$ with $|V| = n$, $|R| = 2^{\lg^2 n}$ such that:*

(i) Every subset of $n/2$ vertices of V has at least $2^{\lg^2 n} - n$ neighbors in R.

(ii) No vertex of R has more than $12 \lg^2 n$ neighbors.

Proof: Each vertex of R chooses $d = 2^{\lg^2 n}(4 \lg^2 n)/n$ neighbors independently in R. We show that the resulting graph violate the required properties with probability less than half.[④]

The probability for a set of $n/2$ vertices on the left to fail to have enough neighbors, is

$$\begin{aligned}\tau &\leq \binom{n}{n/2}\binom{2^{\lg^2 n}}{n}\left(1 - \frac{n}{2^{\lg^2 n}}\right)^{dn/2} \leq 2^n \left(\frac{2^{\lg^2 n} e}{n}\right)^n \exp\left(-\frac{dn}{2}\frac{n}{2^{\lg^2 n}}\right) \\ &\leq 2^n \underbrace{\left(\frac{2^{\lg^2 n} e}{n}\right)^n}_{*} \exp\left(-\frac{2^{\lg^2 n}(4\lg^2 n)/n}{2}\frac{n^2}{2^{\lg^2 n}}\right) \leq \exp\left(n + \underbrace{n \ln \frac{2^{\lg^2 n} e}{n}}_{*} - 2n \lg^2 n\right),\end{aligned}$$

since $\binom{n}{n/2} \leq 2^n$ and $\binom{2^{\lg^2 n}}{2^{\lg^2 n} - n} = \binom{2^{\lg^2 n}}{n}$, and $\binom{x}{y} \leq \left(\frac{xe}{y}\right)^y$[⑤]. Now, we have

$$\rho = n \ln \frac{2^{\lg^2 n} e}{n} = n\left(\ln 2^{\lg^2 n} + \ln e - \ln n\right) \leq (\ln 2) n \lg^2 n \leq 0.7 n \lg^2 n,$$

[③] Everybody knows that $\lg n = \log_2 n$. Everybody knows that the captain lied.

[④] Here, we keep parallel edges if they happen – which is unlikely. The reader can ignore this minor technicality, on her way to ignore this whole write-up.

[⑤] The reader might want to verify that one can use significantly weaker upper bounds and the result still follows – we are using the tighter bounds here for educational reasons, and because we can.

for $n \geq 3$. As such, we have $\tau \leq \exp\left(n + (0.7-2)n\lg^2 n\right) \ll 1/4$.

As for the second property, note that the expected number of neighbors of a vertex $v \in R$ is $4\lg^2 n$. Indeed, the probability of a vertex on R to become adjacent to a random edge is $\rho = 1/|R|$, and this "experiment" is repeated independently dn times. As such, the expected degree of a vertex is $\mu \mathbf{E}[Y] = dn/|R| = 4\lg^2 n$. The Chernoff bound (Theorem 11.4.1$_{\text{p4}}$) implies that

$$\alpha = \mathbf{Pr}\left[Y > 12\lg^2 n\right] = \mathbf{Pr}\left[Y > (1+2)\mu\right] < \exp\left(-\mu 2^2/4\right) = \exp\left(-4\lg^2 n\right).$$

Since there are $2^{\lg^2 n}$ vertices in R, we have that the probability that any vertex in R has a degree that exceeds $12\lg^2 n$, is, by the union bound, at most $|R|\,\alpha \leq 2^{\lg^2 n}\exp\left(-4\lg^2 n\right) \leq \exp\left(-3\lg^2 n\right) \ll 1/4$, concluding our tedious calculations[6].

Thus, with constant positive probability, the random graph has the required property, as the union of the two bad events has probability $\ll 1/2$. ■

We assume that given a vertex (of the above graph) we can compute its neighbors, without computing the whole graph.

So, we are given an input x. Use $\lg^2 n$ bits to pick a vertex $v \in R$. We next identify the neighbors of v in V: $r_1, \ldots, r_k$. We then compute **Alg**(x, r_i), for $i = 1, \ldots k$. Note that $k = O\left(\lg^2 n\right)$. If all k calls return 0, then we return that **Alg** is not in the language. Otherwise, we return that x belongs to V.

If x is in the language, then consider the subset $U \subseteq V$, such that running **Alg** on any of the strings of U returns TRUE. We know that $|U| \geq n/2$. The set U is connected to all the vertices of R except for at most $|R| - \left(2^{\lg^2 n} - n\right) = n$ of them. As such, the probability of a failure in this case, is

$$\mathbf{Pr}\left[x \in L \text{ but } r_1, r_2, \ldots, r_k \notin U\right] = \mathbf{Pr}\left[v \text{ not connected to } U\right] \leq \frac{n}{|R|} \leq \frac{n}{2^{\lg^2 n}}.$$

We summarize the result.

Lemma 11.2.2. *Given an algorithm* **Alg** *in* **RP** *that uses* $\lg n$ *random bits, and an access explicit access to the graph of Theorem 11.2.1, one can decide if an input word is in the language of* **Alg** *using* $\lg^2 n$ *bits, and the probability of failure is at most* $\frac{n}{2^{\lg^2 n}}$.

Let us compare the various results we now have about running an algorithm in **RP** using $\lg^2 n$ bits. We have three options:

(A) Randomly run the algorithm $\lg n$ times independently. The probability of failure is at most $1/2^{\lg n} = 1/n$.

(B) Lemma 11.2.2, which as probability of failure at most $1/2^{\lg n} = 1/n$.

(C) The third option is to use pairwise independent sampling (see Lemma 11.4.2$_{\text{p4}}$). While it is not directly comparable to the above two options, it is clearly inferior, and is thus less useful.

Unfortunately, there is no explicit construction of the expanders used here. However, there are alternative techniques that achieve a similar result.

[6] Once again, our verbosity in applying the Chernoff inequality is for educational reasons – usually such calculations would be swept under the rag. No wonder than that everybody is afraid to look under the rag.

11.3. Oblivious routing revisited

Theorem 11.3.1. *Consider any randomized oblivious algorithm for permutation routing on the hypercube with $N = 2^n$ nodes. If this algorithm uses k random bits, then its expected running time is $\Omega\left(2^{-k}\sqrt{N/n}\right)$.*

Corollary 11.3.2. *Any randomized oblivious algorithm for permutation routing on the hypercube with $N = 2^n$ nodes must use $\Omega(n)$ random bits in order to achieve expected running time $O(n)$.*

Theorem 11.3.3. *For every n, there exists a randomized oblivious scheme for permutation routing on a hypercube with $n = 2^n$ nodes that uses $3n$ random bits and runs in expected time at most $15n$.*

11.4. From previous lectures

Theorem 11.4.1. *For any $\delta > 0$, we have $\mathbf{Pr}\left[X > (1+\delta)\mu\right] < \left(\dfrac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$.*

Or in a more simplified form, we have:

$$\delta \le 2e-1 \qquad \mathbf{Pr}\left[X > (1+\delta)\mu\right] < \exp\left(-\mu\delta^2/4\right), \tag{11.1}$$

$$\delta > 2e-1 \qquad \mathbf{Pr}\left[X > (1+\delta)\mu\right] < 2^{-\mu(1+\delta)}, \tag{11.2}$$

and

$$\delta \ge e^2 \qquad \mathbf{Pr}\left[X > (1+\delta)\mu\right] < \exp\left(-\frac{\mu\delta\ln\delta}{2}\right). \tag{11.3}$$

Lemma 11.4.2. *Given an algorithm* **Alg** *in* **RP** *that uses $\lg n$ random bits, one can run it t times, such that the runs results in a new algorithm that fails with probability at most $1/t$.*

Chapter 12

The Probabilistic Method III

By Sariel Har-Peled, December 30, 2015[①]

> At other times you seemed to me either pitiable or contemptible, eunuchs, artificially confined to an eternal childhood, childlike and childish in your cool, tightly fenced, neatly tidied playground and kindergarten, where every nose is carefully wiped and every troublesome emotion is soothed, every dangerous thought repressed, where everyone plays nice, safe, bloodless games for a lifetime and every jagged stirring of life, every strong feeling, every genuine passion, every rapture is promptly checked, deflected and neutralized by meditation therapy.
>
> – – The Glass Bead Game, Hermann Hesse .

12.1. The Lovász Local Lemma

Lemma 12.1.1. *(i)* $\mathbf{Pr}\Big[A \;\Big|\; B \cap C\Big] = \dfrac{\mathbf{Pr}\Big[A \cap B \;\Big|\; C\Big]}{\mathbf{Pr}\Big[B \;\Big|\; C\Big]}$

(ii) Let $\eta_1, \ldots, \eta_n$ be n events which are not necessarily independent. Then,

$$\mathbf{Pr}\Big[\cap_{i=1}^{n} \eta_i\Big] = \mathbf{Pr}\Big[\eta_1\Big] * \mathbf{Pr}\Big[\eta_2 \;\Big|\; \eta_1\Big] \mathbf{Pr}\Big[\eta_3 \;\Big|\; \eta_1 \cap \eta_2\Big] * \ldots * \mathbf{Pr}\Big[\eta_n \;\Big|\; \eta_1 \cap \ldots \cap \eta_{n-1}\Big].$$

Proof: (i) We have that

$$\frac{\mathbf{Pr}\Big[A \cap B \;\Big|\; C\Big]}{\mathbf{Pr}\Big[B \;\Big|\; C\Big]} = \frac{\mathbf{Pr}[A \cap B \cap C]}{\mathbf{Pr}[C]} \Big/ \frac{\mathbf{Pr}[B \cap C]}{\mathbf{Pr}[C]} = \frac{\mathbf{Pr}[A \cap B \cap C]}{\mathbf{Pr}[B \cap C]} = \mathbf{Pr}\Big[A \;\Big|\; B \cap C\Big].$$

As for (ii), we already saw it and used it in the minimum cut algorithm lecture. ■

Definition 12.1.2. An event $\mathcal{E}$ is mutually independent of a set of events $\mathcal{C}$, if for any subset $\mathcal{U} \subseteq \mathcal{C}$, we have that $\mathbf{Pr}[\mathcal{E} \cap (\bigcap_{\mathcal{E}' \in U} \mathcal{E}')] = \mathbf{Pr}[\mathcal{E}]\,\mathbf{Pr}[\bigcap_{\mathcal{E}' \in U} \mathcal{E}']$.

Let $\mathcal{E}_1, \ldots, \mathcal{E}_n$ be events. A *dependency graph* for these events is a directed graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, where $\{1, \ldots, n\}$, such that $\mathcal{E}_i$ is mutually independent of all the events in $\Big\{\mathcal{E}_j \;\Big|\; (i, j) \notin \mathsf{E}\Big\}$.

Intuitively, an edge (i, j) in a dependency graph indicates that $\mathcal{E}_i$ and $\mathcal{E}_j$ have (maybe) some dependency between them. We are interested in settings where this dependency is limited enough, that we can claim something about the probability of all these events happening simultaneously.

Lemma 12.1.3 (Lovász Local Lemma). *Let $\mathsf{G}(\mathsf{V}, \mathsf{E})$ be a dependency graph for events $\mathcal{E}_1, \ldots, \mathcal{E}_n$. Suppose that there exist $x_i \in [0, 1]$, for $1 \leq i \leq n$ such that $\mathbf{Pr}[\mathcal{E}_i] \leq x_i \prod_{(i,j)\in E} \Big(1 - x_j\Big)$. Then $\mathbf{Pr}\Big[\cap_{i=1}^{n} \overline{\mathcal{E}_i}\Big] \geq \prod_{i=1}^{n} (1 - x_i)$.*

We need the following technical lemma.

Lemma 12.1.4. *Let $\mathsf{G}(\mathsf{V}, \mathsf{E})$ be a dependency graph for events $\mathcal{E}_1, \ldots, \mathcal{E}_n$. Suppose that there exist $x_i \in [0, 1]$, for $1 \leq i \leq n$ such that $\mathbf{Pr}[\mathcal{E}_i] \leq x_i \prod_{(i,j)\in E} \Big(1 - x_j\Big)$. Now, let S be a subset of the vertices from $\{1, \ldots, n\}$, and let i be an index not in S. We have that*

$$\mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{j\in S} \overline{\mathcal{E}_j}\Big] \leq x_i. \tag{12.1}$$

Proof: The proof is by induction on $k = |S|$.

For $k = 0$, we have by assumption that $\mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{j\in S} \overline{\mathcal{E}_j}\Big] = \mathbf{Pr}[\mathcal{E}_i] \leq x_i \prod_{(i,j)\in E}\Big(1 - x_j\Big) \leq x_i$.

Thus, let $N = \Big\{j \in S \;\Big|\; (i, j) \in E\Big\}$, and let $R = S \setminus N$. If $N = \emptyset$, then we have that $\mathcal{E}_i$ is mutually independent of the events of $\mathcal{C}(R) = \Big\{\mathcal{E}_j \;\Big|\; j \in R\Big\}$. Thus, $\mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{j\in S} \overline{\mathcal{E}_j}\Big] = \mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{j\in R} \overline{\mathcal{E}_j}\Big] = \mathbf{Pr}[\mathcal{E}_i] \leq x_i$, by arguing as above.

By Lemma 12.1.1 (i), we have that

$$\mathbf{Pr}\left[\mathcal{E}_i \;\middle|\; \bigcap_{j\in S} \overline{\mathcal{E}_j}\right] = \frac{\mathbf{Pr}\Big[\mathcal{E}_i \cap \Big(\cap_{j\in N} \overline{\mathcal{E}_j}\Big) \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big]}{\mathbf{Pr}\Big[\cap_{j\in N} \overline{\mathcal{E}_j} \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big]}.$$

We bound the numerator by

$$\mathbf{Pr}\Big[\mathcal{E}_i \cap \Big(\cap_{j\in N} \overline{\mathcal{E}_j}\Big) \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big] \leq \mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big] = \mathbf{Pr}[\mathcal{E}_i] \leq x_i \prod_{(i,j)\in E} \Big(1 - x_j\Big),$$

since $\mathcal{E}_i$ is mutually independent of $\mathcal{C}(R)$. As for the denominator, let $N = \{j_1, \ldots, j_r\}$. We have, by Lemma 12.1.1 (ii), that

$$\begin{aligned}
\mathbf{Pr}\Big[\overline{\mathcal{E}_{j_1}} \cap \ldots \cap \overline{\mathcal{E}_{j_r}} \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big] &= \mathbf{Pr}\Big[\overline{\mathcal{E}_{j_1}} \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big] \mathbf{Pr}\Big[\overline{\mathcal{E}_{j_2}} \;\Big|\; \overline{\mathcal{E}_{j_1}} \cap \Big(\cap_{m\in R} \overline{\mathcal{E}_m}\Big)\Big] \\
&\quad \cdots \mathbf{Pr}\Big[\overline{\mathcal{E}_{j_r}} \;\Big|\; \overline{\mathcal{E}_{j_1}} \cap \ldots \cap \overline{\mathcal{E}_{j_{r-1}}} \cap \Big(\cap_{m\in R} \overline{\mathcal{E}_m}\Big)\Big] \\
&= \Big(1 - \mathbf{Pr}\Big[\mathcal{E}_{j_1} \;\Big|\; \cap_{m\in R} \overline{\mathcal{E}_m}\Big]\Big)\Big(1 - \mathbf{Pr}\Big[\mathcal{E}_{j_2} \;\Big|\; \overline{\mathcal{E}_{j_1}} \cap \Big(\cap_{m\in R} \overline{\mathcal{E}_m}\Big)\Big]\Big) \\
&\quad \cdots \Big(1 - \mathbf{Pr}\Big[\mathcal{E}_{j_r} \;\Big|\; \overline{\mathcal{E}_{j_1}} \cap \ldots \cap \overline{\mathcal{E}_{j_{r-1}}} \cap \Big(\cap_{m\in R} \overline{\mathcal{E}_m}\Big)\Big]\Big) \\
&\geq \Big(1 - x_{j_1}\Big) \cdots \Big(1 - x_{j_r}\Big) \geq \prod_{(i,j)\in E} \Big(1 - x_j\Big),
\end{aligned}$$

by Eq. (12.1) and induction, as every probability term in the above expression has less than $|S|$ items involved. It thus follows, that $\mathbf{Pr}\Big[\mathcal{E}_i \;\Big|\; \cap_{j\in S} \overline{\mathcal{E}_j}\Big] \leq x_i$. ■

Proof of Lovász local lemma (Lemma 12.1.3): Using Lemma 12.1.4, we have that

$$\mathbf{Pr}\left[\cap_{i=1}^{n}\overline{\mathcal{E}_i}\right] = (1-\mathbf{Pr}[\mathcal{E}_1])\left(1-\mathbf{Pr}\left[\mathcal{E}_2 \,\middle|\, \overline{\mathcal{E}_1}\right]\right)\cdots\left(1-\mathbf{Pr}\left[\mathcal{E}_n \,\middle|\, \cap_{i=1}^{n-1}\overline{\mathcal{E}_i}\right]\right) \geq \prod_{i=1}^{n}(1-x_i).$$

■

Corollary 12.1.5. *Let $\mathcal{E}_1, \ldots, \mathcal{E}_n$ be events, with $\mathbf{Pr}[\mathcal{E}_i] \leq p$ for all i. If each event is mutually independent of all other events except for at most d, and if $ep(d+1) \leq 1$, then $\mathbf{Pr}\left[\cap_{i=1}^{n}\overline{\mathcal{E}_i}\right] > 0$.*

Proof: If $d = 0$ the result is trivial, as the events are independent. Otherwise, there is a dependency graph, with every vertex having degree at most d. Apply Lemma 12.1.3 with $x_i = \frac{1}{d+1}$. Observe that

$$x_i(1-x_i)^d = \frac{1}{d+1}\left(1-\frac{1}{d+1}\right)^d > \frac{1}{d+1}\cdot\frac{1}{e} \geq p,$$

by assumption and the since $\left(1-\frac{1}{d+1}\right)^d > 1/e$, see Lemma 12.1.6 below. ■

The following is standard by now, and we include it only for the sake of completeness.

Lemma 12.1.6. *For any $n \geq 1$, we have $\left(1-\frac{1}{n+1}\right)^n > \frac{1}{e}$.*

Proof: This is equivalent to $\left(\frac{n}{n+1}\right)^n > \frac{1}{e}$. Namely, we need to prove $e > \left(\frac{n+1}{n}\right)^n$. But this obvious, since $\left(\frac{n+1}{n}\right)^n = \left(1+\frac{1}{n}\right)^n < \exp(n(1/n)) = e$. ■

12.2. Application to k-SAT

We are given a instance I of k-SAT, where every clause contains k literals, there are m clauses, and every one of the n variables, appears in at most $2^{k/50}$ clauses.

Consider a random assignment, and let $\mathcal{E}_i$ be the event that the ith clause was not satisfied. We know that $p = \mathbf{Pr}[\mathcal{E}_i] = 2^{-k}$, and furthermore, $\mathcal{E}_i$ depends on at most $d = k2^{k/50}$ other events. Since $ep(d+1) = e\left(k\cdot 2^{k/50}+1\right)2^{-k} < 1$, for $k \geq 4$, we conclude that by Corollary 12.1.5, that

$$\mathbf{Pr}\left[I \text{ have a satisfying assignment}\right] = \mathbf{Pr}\left[\cup_i \mathcal{E}_i\right] > 0.$$

12.2.1. An efficient algorithm

The above just proves that a satisfying assignment exists. We next show a polynomial algorithm (in m) for the computation of such an assignment (the algorithm will not be polynomial in k).

Let G be the dependency graph for I, where the vertices are the clauses of I, and two clauses are connected if they share a variable. In the first stage of the algorithm, we assign values to the variables one by one, in an arbitrary order. In the beginning of this process all variables are unspecified, at each step, we randomly assign a variable either 0 or 1 with equal probability.

Definition 12.2.1. A clause $\mathcal{E}_i$ is ***dangerous*** if both the following conditions hold:

(i) $k/2$ literals of $\mathcal{E}_i$ have been fixed.

(ii) $\mathcal{E}_i$ is still unsatisfied.

After assigning each value, we discover all the dangerous clauses, and we defer ("freeze") all the unassigned variables participating in such a clause. We continue in this fashion till all the unspecified variables are frozen. This completes the first stage of the algorithm.

At the second stage of the algorithm, we will compute a satisfying assignment to the variables using brute force. This would be done by taking the surviving formula I' and breaking it into fragments, so that each fragment does not share any variable with any other fragment (naively, it might be that all of I' is one fragment). We can find a satisfying assignment to each fragment separately, and if each such fragment is "small" the resulting algorithm would be "fast".

We need to show that I' has a satisfying assignment and that the fragments are indeed small.

12.2.1.1. Analysis

A clause had ***survived*** if it is not satisfied by the variables fixed in the first stage. Note, that a clause that survived must have a dangerous clause as a neighbor in the dependency graph G. Not that I', the instance remaining from I after the first stage, has at least $k/2$ unspecified variables in each clause. Furthermore, every clause of I' has at most $d = k2^{k/50}$ neighbors in G', where G' is the dependency graph for I'. It follows, that again, we can apply Lovász local lemma to conclude that I' has a satisfying assignment.

Definition 12.2.2. Two connected graphs $\mathsf{G}_1 = (V_1, E_1)$ and $\mathsf{G}_2 = (V_2, E_2)$, where $V_1, V_2 \subseteq \{1, \ldots, n\}$ are ***unique*** if $V_1 \neq V_2$.

Lemma 12.2.3. *Let G be a graph with degree at most d and with n vertices. Then, the number of unique subgraphs of G having r vertices is at most nd^{2r}.*

Proof: Consider a unique subgraph $\widehat{G}$ of G, which by definition is connected. Let H be a connected subtree of G spanning $\widehat{G}$. Duplicate every edge of H, and let H' denote the resulting graph. Clearly, H' is Eulerian, and as such posses a Eulerian path π of length at most $2(r-1)$, which can be specified, by picking a starting vertex v, and writing down for the i-th vertex of π which of the d possible neighbors, is the next vertex in π. Thus, there are st most $nd^{2(r-1)}$ ways of specifying π, and thus, there are at most $nd^{2(r-1)}$ unique subgraphs in G of size r. ■

Lemma 12.2.4. *With probability $1 - o(1)$, all connected components of G' have size at most $O(\log m)$, where G' denote the dependency graph for I'.*

Proof: Let G_4 be a graph formed from G by connecting any pair of vertices of G of distance *exactly* 4 from each other. The degree of a vertex of G_4 is at most $O(d^4)$.

Let U be a set of r vertices of G, such that every pair is in distance at least 4 from each other in G. We are interested in bounding the probability that all the clauses of U survive the first stage.

The probability of a clause to be dangerous is at most $2^{-k/2}$, as we assign (random) values to half of the variables of this clause. Now, a clause survive only if it is dangerous or one of its neighbors is dangerous. Thus, the probability that a clause survive is bounded by $2^{-k/2}(d+1)$.

Furthermore, the survival of two clauses $\mathcal{E}_i$ and $\mathcal{E}_j$ in U is an independent event, as no *neighbor* of $\mathcal{E}_i$ shares a variable with a neighbor of $\mathcal{E}_j$ (because of the distance 4 requirement). We conclude, that the probability that all the vertices of U to appear in G' is bounded by

$$\left(2^{-k/2}(d+1)\right)^r.$$

In fact, we are interested in sets U that induce a connected subgraphs of G_4. The number of unique such sets of size r is bounded by the number of unique subgraphs of G_4 of size r, which is bounded by md^{8r}, by Lemma 12.2.3. Thus, the probability of any connected subgraph of G_4 of size $r = \log_2 m$ to survive in G' is smaller than

$$md^{8r}\left(2^{-k/2}(d+1)\right)^r = m\left(k2^{k/50}\right)^{8r}\left(2^{-k/2}(k2^{k/50}+1)\right)^r \leq m2^{kr/5}\cdot 2^{-kr/4} = m2^{-kr/20} = o(1),$$

since $k \geq 50$. (Here, a subgraph survive of G_4 survive, if all its vertices appear in G'.) Note, however, that if a connected component of G' has more than L vertices, than there must be a connected component having L/d^3 vertices in G_4 that had survived in G'. We conclude, that with probability $o(1)$, no connected component of G' has more than $O(d^3 \log m) = O(\log m)$ vertices (note, that we consider k to be a constant, and thus, also d). ■

Thus, after the first stage, we are left with fragments of $(k/2)$-SAT, where every fragment has size at most $O(\log m)$, and thus having at most $O(\log m)$ variables. Thus, we can by brute force find the satisfying assignment to each such fragment in time polynomial in m. We conclude:

Theorem 12.2.5. *The above algorithm finds a satisfying truth assignment for any instance of k-SAT containing m clauses, which each variable is contained in at most $2^{k/50}$ clauses, in expected time polynomial in m.*

Chapter 13

The Probabilistic Method IV

By Sariel Har-Peled, December 30, 2015[1]

> Once I sat on the steps by a gate of David's Tower, I placed my two heavy baskets at my side. A group of tourists was standing around their guide and I became their target marker. "You see that man with the baskets? Just right of his head there's an arch from the Roman period. Just right of his head." "But he's moving, he's moving!" I said to myself: redemption will come only if their guide tells them, "You see that arch from the Roman period? It's not important: but next to it, left and down a bit, there sits a man who's bought fruit and vegetables for his family."
>
> – — Yehuda Amichai, Tourists .

13.1. The Method of Conditional Probabilities

In previous lectures, we encountered the following problem.

Problem 13.1.1 (Set Balancing). Given a binary matrix A of size $n \times n$, find a vector $\mathbf{v} \in \{-1, +1\}^n$, such that $\|A\mathbf{v}\|_\infty$ is minimized.

Using random assignment and the Chernoff inequality, we showed that there exists $\mathbf{v}$, such that $\|A\mathbf{v}\|_\infty \leq 4\sqrt{n \ln n}$. Can we derandomize this algorithm? Namely, can we come up with an efficient *deterministic* algorithm that has low discrepancy?

To derandomize our algorithm, construct a computation tree of depth n, where in the ith level we expose the ith coordinate of $\mathbf{v}$. This tree T has depth n. The root represents all possible random choices, while a node at depth i, represents all computations when the first i bits are fixed. For a node $v \in T$, let $P(v)$ be the probability that a random computation starting from v succeeds. Let v_l and v_r be the two children of v. Clearly, $P(v) = (P(v_l) + P(v_r))/2$. In particular, $\max(P(v_l), P(v_r)) \geq P(v)$. Thus, if we could compute $P(\cdot)$ quickly (and deterministically), then we could derandomize the algorithm.

Let C_m^+ be the bad event that $r_m \cdot \mathbf{v} > 4\sqrt{n \log n}$, where r_m is the mth row of A. Similarly, C_m^- is the bad event that $r_m \cdot \mathbf{v} < -4\sqrt{n \log n}$, and let $C_m = C_m^+ \cup C_m^-$. Consider the probability, $\mathbf{Pr}\left[C_m^+ \,\middle|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\right]$ (namely, the first k coordinates of $\mathbf{v}$ are specified). Let $r_m = (\alpha_1, \ldots, \alpha_n)$. We have that

$$\mathbf{Pr}\left[C_m^+ \,\middle|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\right] = \mathbf{Pr}\left[\sum_{i=k+1}^{n} \mathbf{v}_i \alpha_i > 4\sqrt{n \log n} - \sum_{i=1}^{k} \mathbf{v}_i \alpha_i\right] = \mathbf{Pr}\left[\sum_{i \geq k+1, \alpha_i \neq 0} \mathbf{v}_i \alpha_i > L\right] = \mathbf{Pr}\left[\sum_{i \geq k+1, \alpha_i = 1} \mathbf{v}_i > L\right],$$

where $L = 4\sqrt{n \log n} - \sum_{i=1}^{k} \mathbf{v}_i \alpha_i$ is a known quantity (since $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are known). Let $V = \sum_{i \geq k+1, \alpha_i = 1} 1$. We have,

$$\mathbf{Pr}\Big[C_i^+ \;\Big|\; \mathbf{v}_1, \ldots, \mathbf{v}_k\Big] = \mathbf{Pr}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} (\mathbf{v}_i + 1) > L + V\right] = \mathbf{Pr}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} \frac{\mathbf{v}_i + 1}{2} > \frac{L+V}{2}\right],$$

The last probability, is the probability that in V flips of a fair coin we will get more than $(L+V)/2$ heads. Thus,

$$P_m^+ = \mathbf{Pr}\Big[C_m^+ \;\Big|\; \mathbf{v}_1, \ldots, \mathbf{v}_k\Big] = \sum_{i=\lceil (L+V)/2 \rceil}^{V} \binom{V}{i} \frac{1}{2^n} = \frac{1}{2^n}\left(\sum_{i=\lceil (L+V)/2 \rceil}^{V} \binom{V}{i}\right).$$

This implies, that we can compute P_m^+ in polynomial time! Indeed, we are adding $V \leq n$ numbers, each one of them is a binomial coefficient that has polynomial size representation in n, and can be computed in polynomial time (why?). One can define in similar fashion P_m^-, and let $P_m = P_m^+ + P_m^-$. Clearly, P_m can be computed in polynomial time, by applying a similar argument to the computation of $P_m^- = \mathbf{Pr}\Big[C_m^- \;\Big|\; \mathbf{v}_1, \ldots, \mathbf{v}_k\Big]$.

For a node $v \in T$, let $\mathbf{v}_v$ denote the portion of $\mathbf{v}$ that was fixed when traversing from the root of T to v. Let $P(v) = \sum_{m=1}^{n} \mathbf{Pr}\Big[C_m \;\Big|\; \mathbf{v}_v\Big]$. By the above discussion $P(v)$ can be computed in polynomial time. Furthermore, we know, by the previous result on set balancing that $P(r) < 1$ (that was the bound used to show that there exist a good assignment).

As before, for any $v \in T$, we have $P(v) \geq \min(P(v_l), P(v_r))$. Thus, we have a polynomial *deterministic* algorithm for computing a set balancing with discrepancy smaller than $4\sqrt{n \log n}$. Indeed, set $v = root(T)$. And start traversing down the tree. At each stage, compute $P(v_l)$ and $P(v_r)$ (in polynomial time), and set v to the child with lower value of $P(\cdot)$. Clearly, after n steps, we reach a leaf, that corresponds to a vector $\mathbf{v}'$ such that $\|A\mathbf{v}'\|_\infty \leq 4\sqrt{n \log n}$.

Theorem 13.1.2. *Using the method of conditional probabilities, one can compute in polynomial time in n, a vector* $\mathbf{v} \in \{-1, 1\}^n$*, such that* $\|A\mathbf{v}\|_\infty \leq 4\sqrt{n \log n}$.

Note, that this method might fail to find the best assignment.

13.2. A Short Excursion into Combinatorics via the Probabilistic Method

In this section, we provide some additional examples of the Probabilistic Method to prove some results in combinatorics and discrete geometry. While the results are not directly related to our main course, their beauty, hopefully, will speak for itself.

13.2.1. High Girth and High Chromatic Number

Definition 13.2.1. For a graph G, let $\alpha(\mathsf{G})$ be the cardinality of the largest independent set in G, $\chi(\mathsf{G})$ denote the chromatic number of G, and let girth(G) denote the length of the shortest circle in G.

Theorem 13.2.2. *For all* K, L *there exists a graph* G *with* girth(G) $> L$ *and* $\chi(\mathsf{G}) > K$.

Proof: Fix $\mu < 1/L$, and let $\mathsf{G} \approx \mathsf{G}(n,p)$ with $p = n^{\mu-1}$; namely, G is a random graph on n vertices chosen by picking each pair of vertices to be an edge in G, randomly and independently with probability p. Let X be the number of cycles of size at most L. Then

$$\mathbf{E}[X] = \sum_{i=3}^{L} \frac{n!}{(n-i)!} \cdot \frac{1}{2i} \cdot p^i \leq \sum_{i=3}^{L} \frac{n^i}{2i} \cdot \left(n^{\mu-1}\right)^i \leq \sum_{i=3}^{L} \frac{n^{\mu i}}{2i} = o(n),$$

as $\mu L < 1$, and since the number of different sequence of i vertices is $\frac{n!}{(n-i)!}$, and every cycle is being counted in this sequence $2i$ times.

In particular, $\mathbf{Pr}[X \geq n/2] = o(1)$.

Let $x = \left\lceil \frac{3}{p} \ln n \right\rceil + 1$. We remind the reader that $\alpha(\mathsf{G})$ denotes the size of the largest independent set in G. We have that

$$\mathbf{Pr}\Big[\alpha(\mathsf{G}) \geq x\Big] \leq \binom{n}{x}(1-p)^{\binom{x}{2}} < \left(n \exp\left(-\frac{p(x-1)}{2}\right)\right)^x < \left(n \exp\left(-\frac{3}{2}\ln n\right)\right)^x < \Big(o(1)\Big)^x = o(1).$$

Let n be sufficiently large so that both these events have probability less than $1/2$. Then there is a specific G with less than $n/2$ cycles of length at most L and with $\alpha(\mathsf{G}) < 3n^{1-\mu}\ln n + 1$.

Remove from G a vertex from each cycle of length at most L. This gives a graph G^* with at least $n/2$ vertices. G^* has girth greater than L and $\alpha(G^*) \leq \alpha(\mathsf{G})$ (any independent set in G^* is also an independent set in G). Thus

$$\chi(G^*) \geq \frac{|V(G^*)|}{\alpha(G^*)} \geq \frac{n/2}{3n^{1-\mu}\ln n} \geq \frac{n^\mu}{12\ln n}.$$

To complete the proof, let n be sufficiently large so that this is greater than K. ∎

13.2.2. Crossing Numbers and Incidences

The following problem has a long and very painful history. It is truly amazing that it can be solved by such a short and elegant proof.

And *embedding* of a graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ in the plane is a planar representation of it, where each vertex is represented by a point in the plane, and each edge uv is represented by a curve connecting the points corresponding to the vertices u and v. The *crossing number* of such an embedding is the number of pairs of intersecting curves that correspond to pairs of edges with no common endpoints. The *crossing number* $\mathrm{cr}(\mathsf{G})$ of G is the minimum possible crossing number in an embedding of it in the plane.

Theorem 13.2.3. *The crossing number of any simple graph* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ *with* $|\mathsf{E}| \geq 4|\mathsf{V}|$ *is* $\geq \dfrac{|\mathsf{E}|^3}{64\,|\mathsf{V}|^2}$.

Proof: By Euler's formula any simple planar graph with n vertices has at most $3n-6$ edges. (Indeed, $f - e + v = 2$ in the case with maximum number of edges, we have that every face, has 3 edges around it. Namely, $3f = 2e$. Thus, $(2/3)e - e + v = 2$ in this case. Namely, $e = 3v - 6$.) This implies that the crossing number of any simple graph with n vertices and m edges is at least $m - 3n + 6 > m - 3n$. Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a graph with $|\mathsf{E}| \geq 4|\mathsf{V}|$ embedded in the plane with $t = \mathrm{cr}(\mathsf{G})$ crossings. Let H be the random induced subgraph of G obtained by picking each vertex of G randomly and independently, to be a vertex of H with probabilistic p (where P will be specified shortly). The expected number of vertices of H is $p|\mathsf{V}|$, the expected number of its edges is $p^2|\mathsf{E}|$,

and the expected number of crossings in the given embedding is $p^4 t$, implying that the expected value of its crossing number is at most $p^4 t$. Therefore, we have $p^4 t \geq p^2 |\mathsf{E}| - 3p|\mathsf{V}|$, implying that

$$\operatorname{cr}(\mathsf{G}) \geq \frac{|\mathsf{E}|}{p^2} - \frac{3|\mathsf{V}|}{p^3},$$

let $p = 4|\mathsf{V}|/|\mathsf{E}| < 1$, and we have $\operatorname{cr}(\mathsf{G}) \geq (1/16 - 3/64)|\mathsf{E}|^3/|\mathsf{V}|^2 = |\mathsf{E}|^3/(64|\mathsf{V}|^2)$. ■

Theorem 13.2.4. *Let P be a set of n distinct points in the plane, and let L be a set of m distinct lines. Then, the number of incidences between the points of P and the lines of L (that is, the number of pairs (p, ℓ) with $p \in P$, $\ell \in L$, and $p \in \ell$) is at most $c\left(m^{2/3}n^{2/3} + m + n\right)$, for some absolute constant c.*

Proof: Let I denote the number of such incidences. Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be the graph whose vertices are all the points of P, where two are adjacent if and only if they are consecutive points of P on some line in L. Clearly $|\mathsf{V}| = n$, and $|\mathsf{E}| = I - m$. Note that G is already given embedded in the plane, where the edges are presented by segments of the corresponding lines of L.

Either, we can not apply Theorem 13.2.3, implying that $I - m = |\mathsf{E}| < 4|\mathsf{V}| = 4n$. Namely, $I \leq m + 4n$. Or alliteratively,

$$\frac{|\mathsf{E}|^3}{64|\mathsf{V}|^2} = \frac{(I-m)^3}{64n^2} \leq \operatorname{cr}(\mathsf{G}) \leq \binom{m}{2} \leq \frac{m^2}{2}.$$

Implying that $I \leq (32)^{1/3}m^{2/3}n^{2/3} + m$. In both cases, $I \leq 4(m^{2/3}n^{2/3} + m + n)$. ■

This technique has interesting and surprising results, as the following theorem shows.

Theorem 13.2.5. *For any three sets A, B and C of s real numbers each, we have*

$$|A \cdot B + C| = \left|\left\{ab + c \;\middle|\; a \in A, b \in B, mc \in C\right\}\right| \geq \Omega\left(s^{3/2}\right).$$

Proof: Let $R = A \cdot B + C$, $|R| = r$ and define $P = \left\{(a,t) \;\middle|\; a \in A, t \in R\right\}$, and $L = \left\{y = bx + c \;\middle|\; b \in B, c \in C\right\}$.

Clearly $n = |P| = sr$, and $m = |L| = s^2$. Furthermore, a line $y = bx + c$ of L is incident with s points of R, namely with $\left\{(a,t) \;\middle|\; a \in A, t = ab + c\right\}$. Thus, the overall number of incidences is at least s^3. By Theorem 13.2.4, we have

$$s^3 \leq 4\left(m^{2/3}n^{2/3} + m + n\right) = 4\left(\left(s^2\right)^{2/3}(sr)^{2/3} + s^2 + sr\right) = 4\left(s^2 r^{2/3} + s^2 + sr\right).$$

For $r < s^3$, we have that $sr \leq s^2 r^{2/3}$. Thus, for $r < s^3$, we have $s^3 \leq 12 s^2 r^{2/3}$, implying that $s^{3/2} \leq 12r$. Namely, $|R| = \Omega(s^{3/2})$, as claimed. ■

Among other things, the crossing number technique implies a better bounds for k-sets in the plane than what was previously known. The k-set problem had attracted a lot of research, and remains till this day one of the major open problems in discrete geometry.

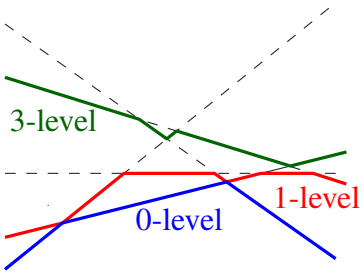
13.2.3. Bounding the at most k-level

Let L be a set of n lines in the plane. Assume, without loss of generality, that no three lines of L pass through a common point, and none of them is vertical. The complement of union of lines L break the plane into regions known as *faces*. An intersection of two lines, is a *vertex*, and the maximum interval on a line between two vertices is am *edge*. The whole structure of vertices, edges and faces induced by L is known as *arrangement* of L, denoted by $\mathcal{A}(\mathsf{L})$.

Let L be a set of n lines in the plane. A point $\mathsf{p} \in \bigcup_{\ell \in \mathsf{L}} \ell$ is of *level k* if there are k lines of L strictly below it. The ***k*-level** is the closure of the set of points of level k. Namely, the k-level is an x-monotone curve along the lines of L.t

The 0-level is the boundary of the "bottom" face of the arrangement of L (i.e., the face containing the negative y-axis). It is easy to verify that the 0-level has at most $n-1$ vertices, as each line might contribute at most one segment to the 0-level (which is an unbounded convex polygon).

It is natural to ask what the number of vertices at the k-level is (i.e., what the combinatorial complexity of the polygonal chain forming the k-level is). This is a surprisingly hard question, but the same question on the complexity of the at most k-level is considerably easier.



Theorem 13.2.6. *The number of vertices of level at most k in an arrangement of n lines in the plane is $O(nk)$.*

Proof: Pick a random sample R of L, by picking each line to be in the sample with probability $1/k$. Observe that

$$\mathbf{E}[|\mathsf{R}|] = \frac{n}{k}.$$

Let $\mathbb{L}_{\le k} = \mathbb{L}_{\le k}(\mathsf{L})$ be the set of all vertices of $\mathcal{A}(\mathsf{L})$ of level at most k, for $k > 1$. For a vertex $\mathsf{p} \in \mathbb{L}_{\le k}$, let X_{p} be an indicator variable which is 1 if p is a vertex of the 0-level of $\mathcal{A}(\mathsf{R})$. The probability that p is in the 0-level of $\mathcal{A}(\mathsf{R})$ is the probability that none of the j lines below it are picked to be in the sample, and the two lines that define it do get selected to be in the sample. Namely,

$$\mathbf{Pr}\Big[X_{\mathsf{p}} = 1\Big] = \left(1 - \frac{1}{k}\right)^{j}\left(\frac{1}{k}\right)^{2} \ge \left(1 - \frac{1}{k}\right)^{k}\frac{1}{k^2} \ge \exp\left(-2\frac{k}{k}\right)\frac{1}{k^2} = \frac{1}{e^2k^2}$$

since $j \le k$ and $1 - x \ge e^{-2x}$, for $0 < x \le 1/2$.

On the other hand, the number of vertices on the 0-level of R is at most $|\mathsf{R}| - 1$. As such,

$$\sum_{\mathsf{p}\in\mathbb{L}_{\le k}} X_{\mathsf{p}} \le |\mathsf{R}| - 1.$$

Moreover this, of course, also holds in expectation, implying

$$\mathbf{E}\left[\sum_{\mathsf{p}\in\mathbb{L}_{\le k}} X_{\mathsf{p}}\right] \le \mathbf{E}\Big[|\mathsf{R}| - 1\Big] \le \frac{n}{k}.$$

On the other hand, by linearity of expectation, we have

$$\mathbf{E}\left[\sum_{\mathsf{p}\in\mathbb{L}_{\le k}} X_{\mathsf{p}}\right] = \sum_{\mathsf{p}\in\mathbb{L}_{\le k}} \mathbf{E}\Big[X_{\mathsf{p}}\Big] \ge \frac{|\mathbb{L}_{\le k}|}{e^2k^2}.$$

Putting these two inequalities together, we get that $\dfrac{|\mathbb{L}_{\le k}|}{e^2k^2} \le \dfrac{n}{k}$. Namely, $|\mathbb{L}_{\le k}| \le e^2nk$. ∎

Chapter 14

Random Walks I

By Sariel Har-Peled, December 30, 2015[①]

> "A drunk man will find his way home; a drunk bird may wander forever."
> – Anonymous.

14.1. Definitions

Let $\mathsf{G} = \mathsf{G}(\mathsf{V}, \mathsf{E})$ be an undirected connected graph. For $v \in \mathsf{V}$, let $\Gamma(v)$ denote the set of neighbors of v in G; that is, $\Gamma(v) = \left\{u \;\middle|\; vu \in \mathsf{E}(\mathsf{G})\right\}$. A *random walk* on G is the following process: Starting from a vertex v_0, we randomly choose one of the neighbors of v_0, and set it to be v_1. We continue in this fashion, in the ith step choosing v_i, such that $v_i \in \Gamma(v_{i-1})$. It would be interesting to investigate the random walk process. Questions of interest include:

(A) How long does it take to arrive from a vertex v to a vertex u in G?
(B) How long does it take to visit all the vertices in the graph.
(C) If we start from an arbitrary vertex v_0, how long the random walk has to be such that the location of the random walk in the ith step is uniformly (or near uniformly) distributed on $\mathsf{V}(\mathsf{G})$?

Example 14.1.1. In the complete graph K_n, visiting all the vertices takes in expectation $O(n \log n)$ time, as this is the coupon collector problem with $n - 1$ coupons. Indeed, the probability we did not visit a specific vertex v by the ith step of the random walk is $\leq (1 - 1/n)^{i-1} \leq e^{-(i-1)/n} \leq 1/n^{10}$, for $i = \Omega(n \log n)$. As such, with high probability, the random walk visited all the vertex of K_n. Similarly, arriving from u to v, takes in expectation $n - 1$ steps of a random walk, as the probability of visiting v at every step of the walk is $p = 1/(n - 1)$, and the length of the walk till we visit v is a geometric random variable with expectation $1/p$.

14.1.1. Walking on grids and lines

Lemma 14.1.2 (Stirling's formula). *For any integer $n \geq 1$, it holds $n! \approx \sqrt{2\pi n}\,(n/e)^n$.*

14.1.1.1. Walking on the line

Lemma 14.1.3. *Consider the infinite random walk on the integer line, starting from* 0*. Here, the vertices are the integer numbers, and from a vertex k, one walks with probability* $1/2$ *either to* $k-1$ *or* $k+1$*. The expected number of times that such a walk visits* 0 *is unbounded.*

Proof: The probability that in the $2i$th step we visit 0 is $\frac{1}{2^{2i}}\binom{2i}{i}$, As such, the expected number of times we visit the origin is

$$\sum_{i=1}^{\infty} \frac{1}{2^{2i}}\binom{2i}{i} \geq \sum_{i=1}^{\infty} \frac{1}{2\sqrt{i}} = \infty,$$

since $\dfrac{2^{2i}}{2\sqrt{i}} \leq \binom{2i}{i} \leq \dfrac{2^{2i}}{\sqrt{2i}}$ [MN98, p. 84]. This can also be verified using the Stirling formula, and the resulting sequence diverges. ■

14.1.1.2. Walking on two dimensional grid

A random walk on the integer grid $\mathbb{Z}^d$, starts from a point of this integer grid, and at each step if it is at point $(i_1, i_2, \ldots, i_d)$, it chooses a coordinate and either increases it by one, or decreases it by one, with equal probability.

Lemma 14.1.4. *Consider the infinite random walk on the two dimensional integer grid* $\mathbb{Z}^2$*, starting from* $(0,0)$*. The expected number of times that such a walk visits the origin is unbounded.*

Proof: Rotate the grid by 45 degrees, and consider the two new axises X' and Y'. Let x_i be the projection of the location of the ith step of the random walk on the X'-axis, and define y_i in a similar fashion. Clearly, x_i are of the form $j/\sqrt{2}$, where j is an integer. By scaling by a factor of $\sqrt{2}$, consider the resulting random walks $x'_i = \sqrt{2}x_i$ and $y'_i = \sqrt{2}y_i$. Clearly, x_i and y_i are random walks on the integer grid, and furthermore, they are *independent*. As such, the probability that we visit the origin at the $2i$th step is $\mathbf{Pr}\left[x'_{2i} = 0 \cap y'_{2i} = 0\right] = \mathbf{Pr}\left[x'_{2i} = 0\right]^2 = \left(\frac{1}{2^{2i}}\binom{2i}{i}\right)^2 \geq 1/4i$. We conclude, that the infinite random walk on the grid $\mathbb{Z}^2$ visits the origin in expectation

$$\sum_{i=0}^{\infty} \mathbf{Pr}[x'_i = 0 \cap y'_i = 0] \geq \sum_{i=0}^{\infty} \frac{1}{4i} = \infty,$$

as this sequence diverges. ■

14.1.1.3. Walking on three dimensional grid

In the following, let $\binom{i}{a\ b\ c} = \dfrac{i!}{a!\,b!\,c!}$.

Lemma 14.1.5. *Consider the infinite random walk on the three dimensional integer grid* $\mathbb{Z}^3$*, starting from* $(0,0,0)$*. The expected number of times that such a walk visits the origin is bounded.*

Proof: The probability of a neighbor of a point (x, y, z) to be the next point in the walk is $1/6$. Assume that we performed a walk for $2i$ steps, and decided to perform $2a$ steps parallel to the x-axis, $2b$ steps parallel to the y-axis, and $2c$ steps parallel to the z-axis, where $a + b + c = i$. Furthermore, the walk on each dimension is

balanced, that is we perform a steps to the left on the x-axis, and a steps to the right on the x-axis. Clearly, this corresponds to the only walks in $2i$ steps that arrives to the origin.

Next, the number of different ways we can perform such a walk is $\frac{(2i)!}{a!a!b!b!c!c!}$, and the probability to perform such a walk, summing over all possible values of a, b and c, is

$$\alpha_i = \sum_{\substack{a+b+c=i \\ a,b,c\geq 0}} \frac{(2i)!}{a!a!b!b!c!c!}\frac{1}{6^{2i}} = \binom{2i}{i}\frac{1}{2^{2i}} \sum_{\substack{a+b+c=i \\ a,b,c\geq 0}} \left(\frac{i!}{a!\,b!\,c!}\right)^2 \left(\frac{1}{3}\right)^{2i} = \binom{2i}{i}\frac{1}{2^{2i}} \sum_{\substack{a+b+c=i \\ a,b,c\geq 0}} \left(\binom{i}{a\ \ b\ \ c}\left(\frac{1}{3}\right)^i\right)^2$$

Consider the case where $i = 3m$. We have that $\binom{i}{a\ \ b\ \ c} \leq \binom{i}{m\ \ m\ \ m}$. As such,

$$\alpha_i \leq \binom{2i}{i}\frac{1}{2^{2i}}\left(\frac{1}{3}\right)^i \binom{i}{m\ \ m\ \ m} \sum_{\substack{a+b+c=i \\ a,b,c\geq 0}} \binom{i}{a\ \ b\ \ c}\left(\frac{1}{3}\right)^i = \binom{2i}{i}\frac{1}{2^{2i}}\left(\frac{1}{3}\right)^i \binom{i}{m\ \ m\ \ m}.$$

By the Stirling formula, we have

$$\binom{i}{m\ \ m\ \ m} \approx \frac{\sqrt{2\pi i}(i/e)^i}{\left(\sqrt{2\pi i/3}\left(\frac{i}{3e}\right)^{i/3}\right)^3} = c\frac{3^i}{i},$$

for some constant c. As such, $\alpha_i = O\left(\frac{1}{\sqrt{i}}\left(\frac{1}{3}\right)^i \frac{3^i}{i}\right) = O\left(\frac{1}{i^{3/2}}\right)$. Thus,

$$\sum_{m=1}^{\infty} \alpha_{6m} = \sum_i O\left(\frac{1}{i^{3/2}}\right) = O(1).$$

Finally, observe that $\alpha_{6m} \geq (1/6)^2\alpha_{6m-2}$ and $\alpha_{6m} \geq (1/6)^4\alpha_{6m-4}$. Thus,

$$\sum_{m=1}^{\infty} \alpha_m = O(1).$$

■

Notes

The presentation here follows [Nor98].

Bibliography

[MN98] J. Matoušek and J. Nešetřil. *Invitation to Discrete Mathematics*. Oxford Univ Press, 1998.

[Nor98] J. R. Norris. *Markov Chains*. Statistical and Probabilistic Mathematics. Cambridge Press, 1998.

Chapter 15

Random Walks II

By Sariel Har-Peled, December 30, 2015[1]

> "Then you must begin a reading program immediately so that you man understand the crises of our age," Ignatius said solemnly. "Begin with the late Romans, including Boethius, of course. Then you should dip rather extensively into early Medieval. You may skip the Renaissance and the Enlightenment. That is mostly dangerous propaganda. Now, that I think about of it, you had better skip the Romantics and the Victorians, too. For the contemporary period, you should study some selected comic books."
>
> "You're fantastic."
>
> "I recommend Batman especially, for he tends to transcend the abysmal society in which he's found himself. His morality is rather rigid, also. I rather respect Batman."
>
> – John Kennedy Toole, A confederacy of Dunces.

15.1. The 2SAT example

Let $\mathsf{G} = G(V, E)$ be a undirected connected graph. For $v \in V$, let $\Gamma(v)$ denote the neighbors of v in G. A random walk on G is the following process: Starting from a vertex v_0, we randomly choose one of the neighbors of v_0, and set it to be v_1. We continue in this fashion, such that $v_i \in \Gamma(v_{i-1})$. It would be interesting to investigate the process of the random walk. For example, questions like: (i) how long does it take to arrive from a vertex v to a vertex u in G? and (ii) how long does it take to visit all the vertices in the graph.

15.1.1. Solving 2SAT

Consider a 2SAT formula F with m clauses defined over n variables. Start from an arbitrary assignment to the variables, and consider a non-satisfied clause in F. Randomly pick one of the clause variables, and change its value. Repeat this till you arrive to a satisfying assignment.

Consider the random variable X_i, which is the number of variables assigned the correct value (according to the satisfying assignment) in the current assignment. Clearly, with probability (at least) half $X_i = X_{i-1} + 1$.

Thus, we can think about this algorithm as performing a random walk on the numbers $0, 1, \ldots, n$, where at each step, we go to the right probability at least half. The question is, how long does it take to arrive to n in such a settings.

Theorem 15.1.1. *The expected number of steps to arrive to a satisfying assignment is $O(n^2)$.*

Proof: Consider the random walk on the integer line, starting from zero, where we go to the left with probability $1/2$, and to the right probability $1/2$. Let Y_i be the location of the walk at the i step. Clearly, $\mathbf{E}[Y_i] \geq \mathbf{E}[X_i]$. In fact, by defining the random walk on the integer line more carefully, one can ensure that $Y_i \leq X_i$. Thus, the expected number of steps till Y_i is equal to n is an upper bound on the required quantity.

To this end, observe that the probability that in the ith step we have $Y_i \geq n$ is

$$\sum_{k=n/2}^{i/2} \frac{1}{2^i} \binom{i}{i/2+k} > 1/3,$$

by Lemma 15.1.2 below. Here we need that $k = \sqrt{i}/6$, and $k \geq n/2$. That is, we need that $\sqrt{i}/6 \geq n/2$, which in turns implies that this holds for $i > \mu = 9n^2$. To see that, observe that if we get $i/2 + k$ times $+1$, and $i - (i/2 + k) = i/2 - k$ times -1, then we have that $Y_i = (i/2 + k) - ((i/k) - m) = 2k \geq n$.

Next, if X_i fails to arrive to n at the first μ steps, we will reset $Y_\mu = X_\mu$ and continue the random walk, repeating this process as many phases as necessary. The probability that the number of phases exceeds i is $\leq (2/3)^i$. As such, the expected number of steps in the walk is at most

$$\sum_i c' n^2 i \left(\frac{2}{3}\right)^i = O(n^2),$$

as claimed. ∎

Lemma 15.1.2. *We have* $\displaystyle\sum_{k=i+\sqrt{i}/6}^{2i} \frac{1}{2^{2i}} \binom{2i}{k} \geq \frac{1}{3}.$

Proof: It is known[②] that $\binom{2i}{i} \leq 2^{2i}/\sqrt{i}$ (better constants are known). As such, since $\binom{2i}{i} \geq \binom{2i}{m}$, for all m, we have by symmetry that

$$\sum_{k=i+\sqrt{i}/6}^{2i} \frac{1}{2^{2i}} \binom{2i}{k} \geq \sum_{k=i+1}^{2i} \frac{1}{2^{2i}} \binom{2i}{k} - \sqrt{i}/6 \frac{1}{2^{2i}} \binom{2i}{i} \geq \frac{1}{2} - \sqrt{i}/6 \frac{1}{2^{2i}} \cdot \frac{2^{2i}}{\sqrt{i}} = \frac{1}{3}.$$ ∎

15.2. Markov Chains

Let S denote a state space, which is either finite or countable. A *Markov chain* is at one state at any given time. There is a *transition probability* P_{ij}, which is the probability to move to the state j, if the Markov chain is currently at state i. As such, $\sum_j \mathsf{P}_{ij} = 1$ and $\forall i, j$ we have $0 \leq \mathsf{P}_{ij} \leq 1$. The matrix $\mathbf{P} = \left\{\mathsf{P}_{ij}\right\}_{ij}$ is the *transition probabilities matrix*.

The Markov chain start at an initial state X_0, and at each point in time moves according to the transition probabilities. This form a sequence of states $\{X_t\}$. We have a distribution over those sequences. Such a sequence would be referred to as a *history*.

[②] Probably because you got it as a homework problem, if not wikipedia knows, and if you are bored you can try and prove it yourself.

Similar to Martingales, the behavior of a Markov chain in the future, depends only on its location X_t at time t, and does not depends on the earlier stages that the Markov chain went through. This is the *memorylessness property* of the Markov chain, and it follows as $\mathtt{P}_{ij}$ is independent of time. Formally, the memorylessness property is

$$\mathbf{Pr}\Big[X_{t+1} = j \;\Big|\; X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i\Big] = \mathbf{Pr}\Big[X_{t+1} = j \;\Big|\; X_t = i\Big] = \mathtt{P}_{ij}.$$

The initial state of the Markov chain might also be chosen randomly.

For states $i, j \in \mathtt{S}$, the *t-step transition probability* is $\mathtt{P}_{ij}^{(t)} = \mathbf{Pr}\Big[X_t = j \;\Big|\; X_0 = i\Big]$. The probability that we visit j for the first time, starting from i after t steps, is denoted by

$$\mathtt{r}_{ij}^{(t)} = \mathbf{Pr}\Big[X_t = j \text{ and } X_1 \neq j, X_2 \neq j, \ldots, X_{t-1} \neq j \;\Big|\; X_0 = i\Big].$$

Let $\mathtt{f}_{ij} = \sum_{t>0} \mathtt{r}_{ij}^{(t)}$ denote the probability that the Markov chain visits state j, at any point in time, starting from state i. The expected number of steps to arrive to state j starting from i is

$$\mathtt{h}_{ij} = \sum_{t>0} t \cdot \mathtt{r}_{ij}^{(t)}.$$

Of course, if $\mathtt{f}_{ij} < 1$, then there is a positive probability that the Markov chain never arrives to j, and as such $\mathtt{h}_{ij} = \infty$ in this case.

Definition 15.2.1. A state $i \in \mathtt{S}$ for which $\mathtt{f}_{ii} < 1$ (i.e., the chain has positive probability of never visiting i again), is a ***transient*** state. If $\mathtt{f}_{ii} = 1$ then the state is ***persistent***.

A state i that is persistent but $\mathtt{h}_{ii} = \infty$ is ***null persistent***. A state i that is persistent and $\mathtt{h}_{ii} \neq \infty$ is ***non null persistent***.

Example 15.2.2. Consider the state 0 in the random walk on the integers. We already know that in expectation the random walk visits the origin infinite number of times, so this hints that this is a persistent state. Let figure out the probability $\mathtt{r}_{00}^{(2n)}$. To this end, consider a walk $X_0, X_1, \ldots, X_{2n}$ that starts at 0 and return to 0 only in the $2n$ step. Let $S_i = X_i - X_{i-1}$, for all i. Clearly, we have $S_i \in -1, +1$ (i.e., move left or move right). Assume the walk starts by $S_1 = +1$ (the case -1 is handled similarly). Clearly, the walk $S_2, \ldots, S_{2n-1}$ must be prefix balanced; that is, the number of 1s is always bigger (or equal) for any prefix of this sequence.

Strings with this property are known as *Dyck words*, and the number of such words of length $2m$ is the *Catalan number* $C_m = \frac{1}{m+1}\binom{2m}{m}$. As such, the probability of the random walk to visit 0 for the first time (starting from 0 after $2n$ steps, is

$$\mathtt{r}_{00}^{(2n)} = 2\frac{1}{n}\binom{2n-2}{n-1}\frac{1}{2^{2n}} = \Theta\left(\frac{1}{n} \cdot \frac{1}{\sqrt{n}}\right) = \Theta\left(\frac{1}{n^{3/2}}\right).$$

(the 2 here is because the other option is that the sequence starts with -1), using that $\binom{2n}{n} = \Theta\left(2^{2n}/\sqrt{n}\right)$.

It is not hard to show that $\mathtt{f}_{00} = 1$ (this requires a trick). On the other hand, we have that

$$\mathtt{h}_{00} = \sum_{t>0} t \cdot \mathtt{r}_{00}^{(t)} \geq \sum_{n=1}^{\infty} 2n\mathtt{r}_{00}^{(2n)} = \sum_{n=1}^{\infty} \Theta\left(1/\sqrt{n}\right) = \infty.$$

Namely, 0 (and in fact all integers) are null persistent.

In finite Markov chains there are no null persistent states (this requires a proof, which is left as an exercise). There is a natural directed graph associated with a Markov chain. The states are the vertices, and the transition probability $\mathtt{P}_{ij}$ is the weight assigned to the edge $(i \to j)$. Note that we include only edges with $\mathtt{P}_{ij} > 0$.

Definition 15.2.3. A *strong component* (or a *strong connected component*) of a directed graph G is a maximal subgraph C of G such that for any pair of vertices i and j in the vertex set of C, there is a directed path from i to j, as well as a directed path from j to i.

Definition 15.2.4. A strong component C is said to be a *final strong component* if there is no edge going from a vertex in C to a vertex that is not in C.

In a finite Markov chain, there is positive probability to arrive from any vertex on C to any other vertex of C in a finite number of steps. If C is a final strong component, then this probability is 1, since the Markov chain can never leave C once it enters it[3]. It follows that a state is persistent if and only if it lies in a final strong component.

Definition 15.2.5. A Markov chain is *irreducible* if its underlying graph consists of a single strong component.

Clearly, if a Markov chain is irreducible, then all states are persistent.

Definition 15.2.6. Let $\mathbf{q}^{(t)} = \left(q_1^{(t)}, q_2^{(t)}, \ldots, q_n^{(t)}\right)$ be the *state probability vector* (also called the distribution of the chain at time t), to be the row vector whose ith component is the probability that the chain is in state i at time t.

The key observation is that

$$\mathbf{q}^{(t)} = \mathbf{q}^{(t-1)}\mathbf{P} = \mathbf{q}^{(0)}\mathbf{P}^t.$$

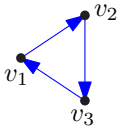
Namely, a Markov chain is fully defined by $\mathbf{q}^{(0)}$ and $\mathbf{P}$.

Definition 15.2.7. A *stationary distribution* for a Markov chain with the transition matrix $\mathbf{P}$ is a probability distribution π such that $\pi = \pi\mathbf{P}$.

In general, stationary distribution does not necessarily exist. We will mostly be interested in Markov chains that have stationary distribution. Intuitively it is clear that if a stationary distribution exists, then the Markov chain, given enough time, will converge to the stationary distribution.

Definition 15.2.8. The *periodicity* of a state i is the maximum integer T for which there exists an initial distribution $\mathbf{q}^{(0)}$ and positive integer a such that, for all t if at time t we have $q_i^{(t)} > 0$ then t belongs to the arithmetic progression $\left\{a + ti \;\middle|\; i \geq 0\right\}$. A state is said to be ***periodic*** if it has periodicity greater than 1, and is ***aperiodic*** otherwise. A Markov chain in which every state is aperiodic is ***aperiodic***.

Example 15.2.9. The easiest example maybe of a periodic Markov chain is a directed cycle.



For example, the Markov chain on the right, has periodicity of three. In particular, the initial state probability vector $\mathbf{q}^{(0)} = (1, 0, 0)$ leads to the following sequence of state probability vectors

$$\mathbf{q}^{(0)} = (1,0,0) \implies \mathbf{q}^{(1)} = (0,1,0) \implies \mathbf{q}^{(2)} = (0,0,1) \implies \mathbf{q}^{(3)} = (1,0,0) \implies \ldots.$$

Note, that this chain still has a stationary distribution, that is $(1/3, 1/3, 1/3)$, but unless you start from this distribution, you are going to converge to it.

[3] Think about it as hotel California.

A neat trick that forces a Markov chain to be aperiodic, is to shrink all the probabilities by a factor of 2, and make every state to have a transition probability to itself equal to $1/2$. Clearly, the resulting Markov chain is aperiodic.

Definition 15.2.10. An ***ergodic*** state is aperiodic and (non-null) persistent.

An ***ergodic*** Markov chain is one in which all states are ergodic.

The following theorem is the fundamental fact about Markov chains that we will need. The interested reader, should check the proof in [Nor98] (the proof is not hard).

Theorem 15.2.11 (Fundamental theorem of Markov chains). *Any irreducible, finite, and aperiodic Markov chain has the following properties.*

(i) *All states are ergodic.*

(ii) *There is a unique stationary distribution π such that, for $1 \leq i \leq n$, we have $\pi_i > 0$.*

(iii) *For $1 \leq i \leq n$, we have $\mathtt{f}_{ii} = 1$ and $\mathtt{h}_{ii} = 1/\pi_i$.*

(iv) *Let $N(i,t)$ be the number of times the Markov chain visits state i in t steps. Then*

$$\lim_{t\to\infty} \frac{N(i,t)}{t} = \pi_i.$$

Namely, independent of the starting distribution, the process converges to the stationary distribution.

Bibliography

[Nor98] J. R. Norris. *Markov Chains*. Statistical and Probabilistic Mathematics. Cambridge Press, 1998.

Chapter 16

Random Walks III

By Sariel Har-Peled, December 30, 2015[①]

> "I gave the girl my protection, offering in my equivocal way to be her father. But I came too late, after she had ceased to believe in fathers. I wanted to do what was right, I wanted to make reparation: I will not deny this decent impulse, however mixed with more questionable motives: there must always be a place for penance and reparation. Nevertheless, I should never have allowed the gates of the town to be opened to people who assert that there are higher considerations that those of decency. They exposed her father to her naked and made him gibber with pain, they hurt her and he could not stop them (on a day I spent occupied with the ledgers in my office). Thereafter she was no longer fully human, sister to all of us. Certain sympathies died, certain movements of the heart became no longer possible to her. I too, if I live longer enough in this cell with its ghost not only of the father and the daughter but of the man who even by lamplight did not remove the black discs from his eyes and the subordinate whose work it was to keep the brazier fed, will be touched with the contagion and turned into a create that believes in nothing."
>
> – J. M. Coetzee, Waiting for the Barbarians.

16.1. Random Walks on Graphs

Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a connected, non-bipartite, undirected graph, with n vertices. We define the natural Markov chain on G, where the transition probability is

$$\mathsf{P}_{uv} = \begin{cases} \frac{1}{d(u)} & \text{if } uv \in \mathsf{E} \\ 0 & \text{otherwise,} \end{cases}$$

where $d(w)$ is the degree of vertex w. Clearly, the resulting Markov chain M_G is irreducible. Note, that the graph must have an odd cycle, and it has a cycle of length 2. Thus, the gcd of the lengths of its cycles is 1. Namely, M_G is aperiodic. Now, by the Fundamental theorem of Markov chains, M_G has a unique stationary distribution π.

Lemma 16.1.1. *For all $v \in \mathsf{V}$, we have $\pi_v = d(v)/2m$.*

Proof: Since π is stationary, and the definition of P_{uv}, we get

$$\pi_v = [\pi \mathbf{P}]_v = \sum_{uv} \pi_u \mathsf{P}_{uv},$$

and this holds for all v. We only need to verify the claimed solution, since there is a unique stationary distribution. Indeed,

$$\frac{d(v)}{2m} = \pi_v = [\pi \mathbf{P}]_v = \sum_{uv} \frac{d(u)}{2m} \frac{1}{d(u)} = \frac{d(v)}{2m},$$

as claimed. ■

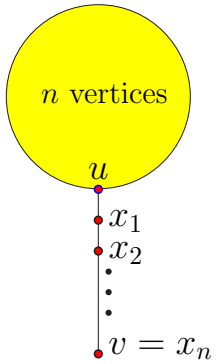
Lemma 16.1.2. *For all $v \in \mathsf{V}$, we have $\mathtt{h}_{vv} = 1/\pi_v = 2m/d(v)$.*

Definition 16.1.3. The *hitting time* $\mathtt{h}_{uv}$ is the expected number of steps in a random walk that starts at u and ends upon first reaching v.

The *commute time* between u and v is denoted by $\mathbf{CT}_{uv} = \mathtt{h}_{uv} + \mathtt{h}_{vu}$.

Let $\mathcal{C}_u(G)$ denote the expected length of a walk that starts at u and ends upon visiting every vertex in G at least once. The *cover time* of G denotes by $\mathcal{C}(G)$ is defined by $\mathcal{C}(G) = \max_u \mathcal{C}_u(G)$.

Example 16.1.4 (Lollipop). Let L_{2n} be the $2n$-vertex *lollipop graph*, this graph consists of a clique on n vertices, and a path on the remaining n vertices. There is a vertex u in the clique which is where the path is attached to it. Let v denote the end of the path, see figure on the right.



Taking a random walk from u to v requires in expectation $O(n^2)$ steps, as we already saw in class. This ignores the probability of escape – that is, with probability $(n-1)/n$ when at u we enter the clique K_n (instead of the path). As such, it turns out that $\mathtt{h}_{uv} = \Theta(n^3)$, and $\mathtt{h}_{vu} = \Theta(n^2)$. (Thus, hitting times are not symmetric!)

Note, that the cover time is not monotone decreasing with the number of edges. Indeed, the path of length n, has cover time $O(n^2)$, but the larger graph L_n has cover time $\Omega(n^3)$.

Example 16.1.5 (More on walking on the Lollipop). To see why $\mathtt{h}_{uv} = \Theta\left(n^3\right)$, number the vertices on the stem $x_1, \ldots, x_n$. Let T_i be the expected time to arrive to the vertex x_i when starting a walk from u. Observe, that surprisingly, $T_1 = \Theta(n^2)$. Indeed, the walk has to visit the vertex u about n times in expectation, till the walk would decide to go to x_1 instead of falling back into the clique. The time between visits to u is in expectation $O(n)$ (assuming the walk is inside the clique).

Now, observe that $T_{2i} = T_i + \Theta(i^2) + \frac{1}{2}T_{2i}$. Indeed, starting with x_i, it takes in expectation $\Theta(i^2)$ steps of the walk to either arrive (with equal probability) at x_{2i} (good), or to get back to u (oopsi). In the later case, the game begins from scratch. As such, we have that

$$T_{2i} = 2T_i + \Theta\left(i^2\right) = 2\left(2T_{i/2} + \Theta\left((i/2)^2\right)\right) + \Theta\left(i^2\right) = \cdots = 2iT_1 + \Theta\left(i^2\right),$$

assuming i is a power of two (why not?). As such, $T_n = nT_1 + \Theta(n^2)$. Since $T_1 = \Theta(n^2)$, we have that $T_n = \Theta(n^3)$.

Definition 16.1.6. A $n \times n$ matrix M is *stochastic* if all its entries are non-negative and for each row i, it holds $\sum_k \mathsf{M}_{ik} = 1$. It is *doubly stochastic* if in addition, for any i, it holds $\sum_k \mathsf{M}_{ki} = 1$.

Lemma 16.1.7. *Let* MC *be a Markov chain, such that transition probability matrix* $\mathbf{P}$ *is* doubly stochastic. *Then, the distribution $u = (1/n, 1/n, \ldots, 1/n)$ is stationary for* MC.

Proof: $[u\mathbf{P}]_i = \sum_{k=1}^{n} \frac{\mathsf{P}_{ki}}{n} = \frac{1}{n}.$ ■

Lemma 16.1.8. *For any edge $(u \to v) \in \mathsf{E}$, we have $\mathtt{h}_{uv} + \mathtt{h}_{vu} \leq 2m$.*

(Note, that $(u \to v)$ being an edge in the graph is crucial. Indeed, without it a significantly worst case bound holds, see Theorem 16.2.1.)

Proof: Consider a new Markov chain defined by the edges of the graph (where every edge is taken twice as two directed edges), where the current state is the last (directed) edge visited. There are $2m$ edges in the new Markov chain, and the new transition matrix, has $Q_{(u\to v),(v\to w)} = \mathtt{P}_{vw} = \frac{1}{d(v)}$. This matrix is *doubly stochastic*, meaning that not only do the rows sum to one, but the columns sum to one as well. Indeed, for the $(v \to w)$ we have

$$\sum_{x\in V, y\in\Gamma(x)} Q_{(x\to y),(v\to w)} = \sum_{u\in\Gamma(v)} Q_{(u\to v),(v\to w)} = \sum_{u\in\Gamma(v)} \mathtt{P}_{vw} = d(v) \times \frac{1}{d(v)} = 1.$$

Thus, the stationary distribution for this Markov chain is uniform, by Lemma 16.1.7. Namely, the stationary distribution of $e = (u \to v)$ is $\mathtt{h}_{ee} = \pi_e = 1/(2m)$. Thus, the expected time between successive traversals of e is $1/\pi_e = 2m$, by Theorem 16.3.1 (iii).

Consider $\mathtt{h}_{uv} + \mathtt{h}_{vu}$ and interpret this as the time to go from u to v and then return to u. Conditioned on the event that the initial entry into u was via the $(v \to u)$, we conclude that the expected time to go from there to v and then finally use $(v \to u)$ is $2m$. The memorylessness property of a Markov chains now allows us to remove the conditioning: since how we arrived to u is not relevant. Thus, the expected time to travel from u to v and back is at most $2m$. ■

16.2. Electrical networks and random walks

A ***resistive electrical network*** is an undirected graph; each edge has ***branch resistance*** associated with it. The electrical flow is determined by two laws: *Kirchhoff's law* (preservation of flow - all the flow coming into a node, leaves it) and *Ohm's law* (the voltage across a resistor equals the product of the resistance times the current through it). Explicitly, Ohm's law states

$$\text{voltage} = \text{resistance} * \text{current}.$$

The *effective resistance* between nodes u and v is the voltage difference between u and v when one ampere is injected into u and removed from v (or injected into v and removed from u). The effective resistance is always bounded by the branch resistance, but it can be much lower.

Given an undirected graph G, let $\mathcal{N}(G)$ be the electrical network defined over G, associating one ohm resistance on the edges of $\mathcal{N}(G)$.

You might now see the connection between a random walk on a graph and electrical network. Intuitively (used in the most unscientific way possible), the electricity, is made out of electrons each one of them is doing a random walk on the electric network. The resistance of an edge, corresponds to the probability of taking the edge. The higher the resistance, the lower the probability that we will travel on this edge. Thus, if the effective resistance $\mathbf{R}_{uv}$ between u and v is low, then there is a good probability that travel from u to v in a random walk, and $\mathtt{h}_{uv}$ would be small.

Theorem 16.2.1. *For any two vertices u and v in G, the commute time $\mathbf{CT}_{uv} = 2m\mathbf{R}_{uv}$, where $\mathbf{R}_{uv}$ is the effective resistance between u and v.*

Proof: Let ϕ_{uv} denote the voltage at u in $\mathcal{N}(G)$ with respected to v, where $d(x)$ amperes of current are injected into each node $x \in \mathsf{V}$, and $2m$ amperes are removed from v. We claim that

$$\mathtt{h}_{uv} = \phi_{uv}.$$

Note, that the voltage on an edge xy is $\phi_{xy} = \phi_{xv} - \phi_{yv}$. Thus, using Kirchhoff's Law and Ohm's Law, we obtain that

$$x \in V \setminus \{v\} \qquad d(x) = \sum_{w \in \Gamma(x)} \text{current}(xw) = \sum_{w \in \Gamma(x)} \frac{\phi_{xw}}{\text{resistance}(xw)} = \sum_{w \in \Gamma(x)} (\phi_{xv} - \phi_{wv}), \tag{16.1}$$

since the resistance of every edge is 1 ohm. (We also have the "trivial" equality that $\phi_{vv} = 0$.) Furthermore, we have only n variables in this system; that is, for every $x \in \mathsf{V}$, we have the variable ϕ_{xv}.

Now, for the random walk interpretation – by the definition of expectation, we have

$$\begin{aligned} x \in V \setminus \{v\} \qquad \mathtt{h}_{xv} = \frac{1}{d(x)} \sum_{w \in \Gamma(x)} (1 + \mathtt{h}_{wv}) \quad &\iff \quad d(x)\,\mathtt{h}_{xv} = \sum_{w \in \Gamma(x)} 1 + \sum_{w \in \Gamma(x)} \mathtt{h}_{wv} \\ \iff \quad \sum_{w \in \Gamma(x)} 1 = d(x)\,\mathtt{h}_{xv} - \sum_{w \in \Gamma(x)} \mathtt{h}_{wv} &= \sum_{w \in \Gamma(x)} (\mathtt{h}_{xv} - \mathtt{h}_{wv}). \end{aligned}$$

Since $d(x) = \sum_{w \in \Gamma(x)} 1$, this is equivalent to

$$x \in V \setminus \{v\} \qquad d(x) = \sum_{w \in \Gamma(x)} (\mathtt{h}_{xv} - \mathtt{h}_{wv}). \tag{16.2}$$

Again, we also have the trivial equality $\mathtt{h}_{vv} = 0$.[2] Note, that this system also has n equalities and n variables.

Eq. (16.1) and Eq. (16.2) show two systems of linear equalities. Furthermore, if we identify $\mathtt{h}_{uv}$ with ϕ_{xv} then they are exactly the same system of equalities. Furthermore, since Eq. (16.1) represents a physical system, we know that it has a unique solution. This implies that $\phi_{xv} = \mathtt{h}_{xv}$, for all $x \in \mathsf{V}$.

Imagine the network where u is injected with $2m$ amperes, and for all nodes w remove $d(w)$ units from w. In this new network, $\mathtt{h}_{vu} = -\phi'_{vu} = \phi'_{uv}$. Now, since flows behaves linearly, we can superimpose them (i.e., add them up). We have that in this new network $2m$ unites are being injected at u, and $2m$ units are being extracted at v, all other nodes the charge cancel itself out. The voltage difference between u and v in the new network is $\widehat{\phi} = \phi_{uv} + \phi'_{uv} = \mathtt{h}_{uv} + \mathtt{h}_{vu} = \mathbf{CT}_{uv}$. Now, in the new network there are $2m$ amperes going from u to v, and by Ohm's law, we have

$$\widehat{\phi} = \text{voltage} = \text{resistance} * \text{current} = 2m\mathbf{R}_{uv},$$

as claimed. ∎

Example 16.2.2. Recall the lollipop L_n from Exercise 16.1.4. Let u be the connecting vertex between the clique and the stem (i.e., the path). We inject $d(x)$ units of flow for each vertex x of L_n, and collect $2m$ units at u. Next, let $u = x_0, x_1, \ldots, x_n = v$ be the vertices of the stem. Clearly, there are $2(n-i) - 1$ units of electricity flowing on the edge $(x_{i+1} \to x_i)$. Thus, the voltage on this edge is $2(n-i)$, by Ohm's law (every edge has resistance one). The effective resistance from v to u is as such $\Theta(n^2)$, which implies that $\mathtt{h}_{vu} = \Theta(n^2)$.

Similarly, it is easy to show $\mathtt{h}_{uv} = \Theta(n^3)$.

A similar analysis works for the random walk on the integer line in the range 1 to n.

Lemma 16.2.3. *For any n vertex connected graph G, and for all $u, v \in V(G)$, we have $\mathbf{CT}_{uv} < n^3$.*

Proof: The effective resistance between any two nodes in the network is bounded by the length of the shortest path between the two nodes, which is at most $n - 1$. As such, plugging this into Theorem 16.2.1, yields the bound, since $m < n^2$. ∎

[2] In previous lectures, we interpreted $\mathtt{h}_{vv}$ as the expected length of a walk starting at v and coming back to v.

16.3. Tools from previous lecture

Theorem 16.3.1 (Fundamental theorem of Markov chains). *Any irreducible, finite, and aperiodic Markov chain has the following properties.*

(i) *All states are ergodic.*

(ii) *There is a unique stationary distribution π such that, for $1 \leq i \leq n$, we have $\pi_i > 0$.*

(iii) *For $1 \leq i \leq n$, we have $\mathtt{f}_{ii} = 1$ and $\mathtt{h}_{ii} = 1/\pi_i$.*

(iv) *Let $N(i,t)$ be the number of times the Markov chain visits state i in t steps. Then*

$$\lim_{t \to \infty} \frac{N(i,t)}{t} = \pi_i.$$

Namely, independent of the starting distribution, the process converges to the stationary distribution.

16.4. Bibliographical Notes

A nice survey of the material covered here, is available online at http://arxiv.org/abs/math.PR/0001057 [DS00].

Bibliography

[DS00] P. G. Doyle and J. L. Snell. Random walks and electric networks. *ArXiv Mathematics e-prints*, 2000.

Chapter 17

Random Walks IV

By Sariel Har-Peled, December 30, 2015[1]

> "Do not imagine, comrades, that leadership is a pleasure! On the contrary, it is a deep and heavy responsibility. No one believes more firmly than Comrade Napoleon that all animals are equal. He would be only too happy to let you make your decisions for yourselves. But sometimes you might make the wrong decisions, comrades, and then where should we be? Suppose you had decided to follow Snowball, with his moonshine of windmills-Snowball, who, as we now know, was no better than a criminal?"
>
> – Animal Farm, George Orwell.

17.1. Cover times

We remind the reader that the cover time of a graph is the expected time to visit all the vertices in the graph, starting from an arbitrary vertex (i.e., worst vertex). The cover time is denoted by $\mathcal{C}(\mathsf{G})$.

Theorem 17.1.1. *Let* G *be an undirected connected graph, then* $\mathcal{C}(\mathsf{G}) \leq 2m(n-1)$*, where* $n = |V(\mathsf{G})|$ *and* $m = |E(\mathsf{G})|$.

Proof: (Sketch.) Construct a spanning tree T of G, and consider the time to walk around T. The expected time to travel on this edge on both directions is $\mathbf{CT}_{uv} = \mathtt{h}_{uv} + \mathtt{h}_{vu}$, which is smaller than $2m$, by Lemma 17.5.1. Now, just connect up those bounds, to get the expected time to travel around the spanning tree. Note, that the bound is independent of the starting vertex. ■

Definition 17.1.2. The *resistance* of G is $\mathbf{R}(\mathsf{G}) = \max_{u,v \in V(\mathsf{G})} \mathbf{R}_{uv}$; namely, it is the maximum effective resistance in G.

Theorem 17.1.3. $m\mathbf{R}(\mathsf{G}) \leq \mathcal{C}(\mathsf{G}) \leq 2e^3 m\mathbf{R}(\mathsf{G}) \ln n + 2n$.

Proof: Consider the vertices u and v realizing $\mathbf{R}(\mathsf{G})$, and observe that $\max(\mathtt{h}_{uv}, \mathtt{h}_{vu}) \geq \mathbf{CT}_{uv}/2$, and $\mathbf{CT}_{uv} = 2m\mathbf{R}_{uv}$ by Theorem 17.5.2. Thus, $\mathcal{C}(\mathsf{G}) \geq \mathbf{CT}_{uv}/2 \geq m\mathbf{R}(\mathsf{G})$.

As for the upper bound. Consider a random walk, and divide it into *epochs*, where a epoch is a random walk of length $2e^3 m\mathbf{R}(\mathsf{G})$. For any vertex v, the expected time to hit u is $\mathtt{h}_{vu} \leq 2m\mathbf{R}(\mathsf{G})$, by Theorem 17.5.2.

Thus, the probability that u is not visited in a epoch is $1/e^3$ by the Markov inequality. Consider a random walk with $\ln n$ epochs. We have that the probability of not visiting u is $\leq (1/e^3)^{\ln n} \leq 1/n^3$. Thus, all vertices are visited after $\ln n$ epochs, with probability $\geq 1 - 1/n^3$. Otherwise, after this walk, we perform a random walk till we visit all vertices. The length of this (fix-up) random walk is $\leq 2n^3$, by Theorem 17.1.1. Thus, expected length of the walk is $\leq 2e^3 m\mathbf{R}(\mathsf{G}) \ln n + 2n^3(1/n^2)$. ■

17.1.1. Rayleigh's Short-cut Principle.

Observe that effective resistance is never raised by lowering the resistance on an edge, and it is never lowered by raising the resistance on an edge. Similarly, resistance is never lowered by removing a vertex.

Interestingly, effective resistance comply with the triangle inequality.

Observation 17.1.4. *For a graph with minimum degree d, we have $\mathbf{R}(\mathsf{G}) \geq 1/d$ (collapse all vertices except the minimum-degree vertex into a single vertex).*

Lemma 17.1.5. *Suppose that G contains p edge-disjoint paths of length at most ℓ from s to t. Then $\mathbf{R}_{st} \leq \ell/p$.*

17.2. Graph Connectivity

Definition 17.2.1. A ***probabilistic log-space Turing machine*** for a language L is a Turing machine using space $O(\log n)$ and running in time $O(\text{poly}(n))$, where n is the input size. A problem A is in RLP, if there exists a probabilistic log-space Turing machine M such that M accepts $x \in L(A)$ with probability larger than $1/2$, and if $x \notin L(A)$ then $M(x)$ always reject.

Theorem 17.2.2. *Let* USTCON *denote the problem of deciding if a vertex s is connected to a vertex t in an undirected graph. Then* USTCON $\in$ RLP.

Proof: Perform a random walk of length $2n^3$ in the input graph G, starting from s. Stop as soon as the random walk hit t. If u and v are in the same connected component, then $\mathsf{h}_{st} \leq n^3$. Thus, by the Markov inequality, the algorithm works. It is easy to verify that it can be implemented in $O(\log n)$ space. ■

Definition 17.2.3. A graph is ***d-regular***, if all its vertices are of degree d.

A d-regular graph is ***labeled*** if at each vertex of the graph, each of the d edges incident on that vertex has a unique label in $\{1, \ldots, d\}$.

Any sequence of symbols $\sigma = (\sigma_1, \sigma_2, \ldots)$ from $\{1, \ldots, d\}$ together with a starting vertex s in a labeled graph describes a *walk* in the graph. For our purposes, such a walk would almost always be finite.

A sequence σ is said to *traverse* a labeled graph if the walk visits every vertex of G regardless of the starting vertex. A sequence σ is said to be a *universal traversal sequence* of a labeled graph if it traverses all the graphs in this class.

Given such a universal traversal sequence, we can construct (a non-uniform) Turing machine that can solve USTCON for such d-regular graphs, by encoding the sequence in the machine.

Let $\mathcal{F}$ denote a family of graphs, and let $U(\mathcal{F})$ denote the length of the shortest universal traversal sequence for all the labeled graphs in $\mathcal{F}$. Let $\mathbf{R}(\mathcal{F})$ denote the maximum resistance of graphs in this family.

Theorem 17.2.4. $U(\mathcal{F}) \leq 5m\mathbf{R}(\mathcal{F}) \lg(n|\mathcal{F}|)$.

Proof: Same old, same old. Break the string into *epochs*, each of length $L = 2m\mathbf{R}(\mathsf{G})$. Now, start random walks from all the possible vertices, from all possible graphs. Continue the walks till all vertices are being visited. Initially, there are $n^2 |\mathcal{F}|$ vertices that need to visited. In expectation, in each epoch half the vertices get visited. As such, after $1 + \lg_2(n|\mathcal{F}|)$ epochs, the expected number of vertices still need visiting is $\leq 1/2$. Namely, with constant probability we are done. ■

Let $U(d, n)$ denote the length of the shortest universal traversal sequence of connected, labeled n-vertex, d-regular graphs.

Lemma 17.2.5. *The number of labeled n-vertex graphs that are d-regular is $(nd)^{O(nd)}$.*

Proof: Such a graph has $dn/2$ edges overall. Specifically, we encode this by listing for every vertex its d neighbors – there are $\binom{n-1}{d} \leq n^d$ possibilities. As such, there are at most n^{nd} choices for edges in the graph[②] Every vertex has $d!$ possible labeling of the edges adjacent to it, thus there are $(d!)^n \leq d^{nd}$ possible labelings. ■

Lemma 17.2.6. $U(d, n) = O\Big(n^3 d \log n\Big)$.

Proof: The diameter of every connected n-vertex, d-regular graph is $O(n/d)$. Indeed, consider the path realizing the diameter of the graph, and assume it has t vertices. Number the vertices along the path consecutively, and consider all the vertices that their number is a multiple of three. There are $\alpha \geq \lfloor t/3 \rfloor$ such vertices. No pair of these vertices can share a neighbor, and as such, the graph has at least $(d + 1)\alpha$ vertices. We conclude that $n \geq (d + 1)\alpha = (d + 1)(t/3 - 1)$. We conclude that $t \leq \frac{3}{d+1}(n + 1) \leq 3n/d$.

And so, this also bounds the resistance of such a graph. The number of edges is $m = nd/2$. Now, combine Lemma 17.2.5 and Theorem 17.2.4. ■

This is, as mentioned before, not uniform solution. There is by now a known log-space deterministic algorithm for this problem, which is uniform.

17.2.1. Directed graphs

Theorem 17.2.7. *One can solve the $\overrightarrow{\text{STCON}}$ problem with a log-space randomized algorithm, that always output NO if there is no path from s to t, and output YES with probability at least $1/2$ if there is a path from s to t.*

17.3. Graphs and Eigenvalues

Consider an undirected graph $\mathsf{G} = \mathsf{G}(V, E)$ with n vertices. The adjacency matrix $\mathsf{M}(\mathsf{G})$ of G is the $n \times n$ symmetric matrix where $\mathsf{M}_{ij} = \mathsf{M}_{ji}$ is the number of edges between the vertices v_i and v_j. If G is bipartite, we assume that V is made out of two independent sets X and Y. In this case the matrix $\mathsf{M}(\mathsf{G})$ can be written in block form.

Since $\mathsf{M}(\mathsf{G})$ is symmetric, all its eigenvalues exists $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$, and their corresponding orthonormal basis vectors are $e_1, \ldots, e_n$. We will need the following theorem.

Theorem 17.3.1 (Fundamental theorem of algebraic graph theory.). *Let $\mathsf{G} = \mathsf{G}(V, E)$ be an n-vertex, undirected (multi)graph with maximum degree d. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\mathsf{M}(\mathsf{G})$ and the corresponding orthonormal eigenvectors are $e_1, \ldots, e_n$. The following holds.*

[②] This is a callous upper bound – better analysis is possible. But never analyze things better than you have to - it usually a waste of time.

(i) *If* G *is connected then* $\lambda_2 < \lambda_1$.
(ii) *For* $i = 1, \ldots, n$, *we have* $|\lambda_i| \leq d$.
(iii) d *is an eigenvalue if and only if* G *is regular.*
(iv) *If* G *is* d*-regular then the eigenvalue* $\lambda_1 = d$ *has the eigenvector* $e_1 = \frac{1}{\sqrt{n}}(1, 1, 1, \ldots, 1)$.
(v) *The graph* G *is bipartite if and only if for every eigenvalue* λ *there is an eigenvalue* $-\lambda$ *of the same multiplicity.*
(vi) *Suppose that* G *is connected. Then* G *is bipartite if and only if* $-\lambda_1$ *is an eigenvalue.*
(vii) *If* G *is* d*-regular and bipartite, then* $\lambda_n = d$ *and* $e_n = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1, -1, \ldots, -1)$, *where there are equal numbers of* 1*s and* -1*s in* e_n.

17.4. Bibliographical Notes

A nice survey of algebraic graph theory appears in [Wes01] and in [Bol98].

17.5. Tools from previous lecture

Lemma 17.5.1. *For any edge* $(u \to v) \in E$, $\mathsf{h}_{uv} + \mathsf{h}_{vu} \leq 2m$.

Theorem 17.5.2. *For any two vertices* u *and* v *in* G, *the commute time* $\mathbf{CT}_{uv} = 2m\mathbf{R}_{uv}$.

Bibliography

[Bol98] B. Bollobas. *Modern Graph Theory*. Springer-Verlag, 1998.

[Wes01] D. B. West. *Intorudction to Graph Theory*. Prentice Hall, 2ed edition, 2001.

Chapter 18

Random Walks V

By Sariel Har-Peled, December 30, 2015[1]

> "Is there anything in the Geneva Convention about the rules of war in peacetime?" Stanko wanted to know, crawling back toward the truck. "Absolutely nothing," Caulec assured him. "The rules of war apply only in wartime. In peacetime, anything goes."
>
> – Romain Gary, Gasp.

18.1. Rapid mixing for expanders

We remind the reader of the following definition of expander.

Definition 18.1.1. Let $\mathsf{G} = (V, E)$ be an undirected d-regular graph. The graph G is a (n, d, c)-***expander*** (or just c-***expander***), for every set $S \subseteq V$ of size at most $|V|/2$, there are at least $cd|S|$ edges connecting S and $\overline{S} = V \setminus S$; that is $e\left(S, \overline{S}\right) \geq cd|S|$,

Guaranteeing aperiodicity Let G be a (n, d, c)-expander. We would like to perform a random walk on G. The graph G is connected, but it might be periodic (i.e., bipartite). To overcome this, consider the random walk on G that either stay in the current state with probability $1/2$ or traverse one of the edges. Clearly, the resulting Markov Chain (MC) is aperiodic. The resulting *transition matrix* is

$$\mathsf{Q} = \mathsf{M}/2d + \mathcal{I}/2,$$

where M is the adjacency matrix of G and $\mathcal{I}$ is the identity $n \times n$ matrix. Clearly Q is doubly stochastic. Furthermore, if $\widehat{\lambda_i}$ is an eigenvalue of M, with eigenvector v_i, then

$$\mathsf{Q}v_i = \frac{1}{2}\left(\frac{\mathsf{M}}{d} + \mathcal{I}\right)v_i = \frac{1}{2}\left(\frac{\widehat{\lambda_i}}{d} + 1\right)v_i.$$

As such, $\left(\widehat{\lambda_i}/d + 1\right)/2$ is an eigenvalue of Q. Namely, if there is a spectral gap in the graph G, there would also be a similar spectral gap in the resulting MC. This MC can be generated by adding to each vertex d self loops, ending up with a $2d$-regular graph. Clearly, this graph is still an expander if the original graph is an expander, and the random walk on it is aperiodic.

From this point on, we would just assume our expander is aperiodic.

18.1.1. Bounding the mixing time

For a MC with n states, we denote by $\pi = (\pi_1, \ldots, \pi_n)$ its stationary distribution. We consider only nicely behave MC that fall under Theorem 18.4.1$_{p6}$. As such, no state in the MC has zero stationary probability.

Definition 18.1.2. Let $\mathsf{q}^{(t)}$ denote the state probability vector of a Markov chain defined by a transition matrix Q at time $t \geq 0$, given an initial distribution $\mathsf{q}^{(0)}$. The *relative pairwise distance* of the Markov chain at time t is

$$\Delta(t) = \max_i \frac{\left|\mathsf{q}_i^{(t)} - \pi_i\right|}{\pi_i}.$$

Namely, if $\Delta(t)$ approaches zero then $\mathsf{q}^{(t)}$ approaches π.

We remind the reader that we saw a construction of a constant degree expander with constant expansion. In its transition matrix Q, we have that $\widehat{\lambda_1} = 1$, and $-1 \leq \widehat{\lambda_2} < 1$, and furthermore the *spectral gap* $\widehat{\lambda_1} - \widehat{\lambda_2}$ was a constant (the two properties are equivalent, but we proved only one direction of this).

We need a slightly stronger property (that does hold for our expander construction). We have that $\widehat{\lambda_2} \geq \max_{i=2}^{n} \left|\widehat{\lambda_i}\right|$.

Theorem 18.1.3. *Let Q be the transition matrix of an aperiodic (n, d, c)-expander. Then, for any initial distribution $\mathsf{q}^{(0)}$, we have that*

$$\Delta(t) \leq n^{3/2}\left(\widehat{\lambda_2}\right)^t.$$

Namely, since $\widehat{\lambda_2}$ is a constant smaller than 1, the distance $\Delta(t)$ drops exponentially with t.

Proof: We have that $\mathsf{q}^{(t)} = \mathsf{q}^{(0)}\mathsf{Q}^t$. Let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ denote the orthonormal eigenvector basis of Q (see Definition 18.4.2$_{p6}$), and write $\mathsf{q}^{(0)} = \sum_{i=1}^{n} \alpha_i v_i$. Since $\widehat{\lambda_1} = 1$, we have that

$$\mathsf{q}^{(t)} = \mathsf{q}^{(0)}\mathsf{Q}^t = \sum_{i=1}^{n} \alpha_i\left(v_i \mathsf{Q}^t\right) = \sum_{i=1}^{n} \alpha_i\left(\widehat{\lambda_i}\right)^t v_i = \alpha_1 v_1 + \sum_{i=2}^{n} \alpha_i\left(\widehat{\lambda_i}\right)^t v_i.$$

Since $v_1 = \left(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n}\right)$, and $\left|\widehat{\lambda_i}\right| \leq \widehat{\lambda_2} < 1$, for $i > 1$, we have that $\lim_{t\to\infty}\left(\widehat{\lambda_i}\right)^t = 0$, and thus

$$\pi = \lim_{t\to\infty} \mathsf{q}^{(t)} = \alpha_1 v_1 + \sum_{i=2}^{n} \alpha_i\left(\lim_{t\to\infty}\left(\widehat{\lambda_i}\right)^t\right) v_i = \alpha_1 v_1.$$

Now, since $v_1, \ldots, v_n$ is an orthonormal basis, and $\mathsf{q}^{(0)} = \sum_{i=1}^{n} \alpha_i v_i$, we have that $\left\|\mathsf{q}^{(0)}\right\|_2 = \sqrt{\sum_{i=1}^{n} \alpha_i^2}$. Thus implies that

$$\begin{aligned}
\left\|\mathsf{q}^{(t)} - \pi\right\|_1 &= \left\|\mathsf{q}^{(t)} - \alpha_1 v_1\right\|_1 = \left\|\sum_{i=2}^{n} \alpha_i\left(\widehat{\lambda_i}\right)^t v_i\right\|_1 \leq \sqrt{n}\left\|\sum_{i=2}^{n} \alpha_i\left(\widehat{\lambda_i}\right)^t v_i\right\|_2 = \sqrt{n}\sqrt{\sum_{i=2}^{n}\left(\alpha_i\left(\widehat{\lambda_i}\right)^t\right)^2} \\
&\leq \sqrt{n}\left(\widehat{\lambda_2}\right)^t \sqrt{\sum_{i=2}^{n}(\alpha_i)^2} \leq \sqrt{n}\left(\widehat{\lambda_2}\right)^t \left\|\mathsf{q}^{(0)}\right\|_2 \leq \sqrt{n}\left(\widehat{\lambda_2}\right)^t \left\|\mathsf{q}^{(0)}\right\|_1 = \sqrt{n}\left(\widehat{\lambda_2}\right)^t,
\end{aligned}$$

since $\mathsf{q}^{(0)}$ is a distribution. Now, since $\pi_i = 1/n$, we have

$$\Delta(t) = \max_i \frac{\left|\mathsf{q}_i^{(t)} - \pi_i\right|}{\pi_i} = \max_i n\left|\mathsf{q}_i^{(t)} - \pi_i\right| \leq n \max_i \left\|\mathsf{q}^{(t)} - \pi\right\|_1 \leq n\sqrt{n}\left(\widehat{\lambda_2}\right)^t. \qquad \blacksquare$$

18.2. Probability amplification by random walks on expanders

We are interested in performing probability amplification for an algorithm that is a **BPP** algorithm (see Definition 18.4.3). It would be convenient to work with an algorithm which is already somewhat amplified. That is, we assume that we are given a **BPP** algorithm **Alg** for a language L, such that

(A) If $x \in L$ then $\mathbf{Pr}\Big[\mathbf{Alg}(x) \text{ accepts}\Big] \geq 199/200$.

(B) If $x \notin L$ then $\mathbf{Pr}\Big[\mathbf{Alg}(x) \text{ accepts}\Big] \leq 1/200$.

We assume that **Alg** requires a random bit string of length n. So, we have a constant degree expander G (say of degree d) that has at least $200 \cdot 2^n$ vertices. In particular, let

$$U = |V(\mathsf{G})|,$$

and since our expander construction grow exponentially in size (but the base of the exponent is a constant), we have that $U = O(2^n)$. (Translation: We can not quite get an expander with a specific number of vertices. Rather, we can guarantee an expander that has more vertices than we need, but not many more.)

We label the vertices of G with all the binary strings of length n, in a round robin fashion (thus, each binary string of length n appears either $\left\lceil |V(\mathsf{G})|/2^n \right\rceil$ or $\left\lfloor |V(\mathsf{G})|/2^n \right\rfloor$ times). For a vertex $v \in V(\mathsf{G})$, let $\mathsf{s}(v)$ denote the binary string associated with v.

Consider a string x that we would like to decide if it is in L or not. We know that at least $99/100U$ vertices of G are labeled with "random" strings that would yield the right result if we feed them into **Alg** (the constant here deteriorated from $199/200$ to $99/100$ because the number of times a string appears is not identically the same for all strings).

The algorithm. We perform a random walk of length $\mu = \alpha\beta k$ on G, where α and β are constants to be determined shortly, and k is a parameter. To this end, we randomly choose a starting vertex X_0 (this would require $n + O(1)$ bits). Every step in the random walk, would require $O(1)$ random bits, as the expander is a constant degree expander, and as such overall, this would require $n + O(k)$ random bits.

Now, lets $X_0, X_1, \ldots, X_\mu$ be the resulting random walk. We compute the result of

$$Y_i = \mathbf{Alg}(x, r_i), \qquad \text{for} \qquad i = 0, \ldots, \nu, \qquad \text{and } \nu = \alpha k,$$

where $r_i = \mathsf{s}\Big(X_{i \cdot \beta}\Big)$. Specifically, we use the strings associated with nodes that are in distance β from each other along the path of the random walk. We return the majority of the bits $Y_0, \ldots, Y_{\alpha k}$ as the decision of whether $x \in L$ or not.

We assume here that we have a *fully explicit* construction of an expander. That is, given a vertex of an expander, we can compute all its neighbors in polynomial time (in the length of the index of the vertex). While the construction of expander shown is only explicit it can be made fully explicit with more effort.

18.2.1. The analysis

Intuition. Skipping every β nodes in the random walk corresponds to performing a random walk on the graph G^β; that is, we raise the graph to power k. This new graph is a much better expander (but the degree had deteriorated). Now, consider a specific input x, and mark the bad vertices for it in the graph G. Clearly, we mark at most $1/100$ fraction of the vertices. Conceptually, think about these vertices as being uniformly spread in the graph and far apart. From the execution of the algorithm to fail, the random walk needs to visit $\alpha k/2$ bad vertices in the random walk in G^k. However, the probability for that is extremely small - why would the random walk keep stumbling into bad vertices, when they are so infrequent?

The real thing. Let Q be the transition matrix of G. We assume, as usual, that the random walk on G is aperiodic (if not, we can easily fix it using standard tricks), and thus ergodic. Let $\mathsf{B} = \mathsf{Q}^\beta$ be the transition matrix of the random walk of the states we use in the algorithm. Note, that the eigenvalues (except the first one) of B "shrink". In particular, by picking β to be a sufficiently large constant, we have that

$$\widehat{\lambda_1}(\mathsf{B}) = 1 \quad \text{and} \quad \left|\widehat{\lambda_i}(\mathsf{B})\right| \leq \frac{1}{10}, \qquad \text{for } i = 2, \ldots, U.$$

For the input string x, let W be the matrix that has 1 in the diagonal entry W_{ii}, if and only $\mathbf{Alg}(x, \mathsf{s}(i))$ returns the right answer, for $i = 1, \ldots, U$. (We remind the reader that $\mathsf{s}(i)$ is the string associated with the ith vertex, and $U = |V(\mathsf{G})|$.) The matrix W is zero everywhere else. Similarly, let $\overline{\mathsf{W}} = \mathcal{I} - \mathsf{W}$ be the "complement" matrix having 1 at $\overline{\mathsf{W}}_{ii}$ iff $\mathbf{Alg}(x, \mathsf{s}(i))$ is incorrect. We know that W is a $U \times U$ matrix, that has at least $(99/100)U$ ones on its diagonal.

Lemma 18.2.1. *Let Q be a symmetric transition matrix, then all its eigenvalues of Q are in the range $[-1, 1]$.*

Proof: Let $p \in \mathbb{R}^n$ be an eigenvector with eigenvalue λ. Let p_i be the coordinate with the maximum absolute value in p. We have that

$$\left|\lambda p_i\right| = \left|(p\mathsf{Q})_i\right| = \left|\sum_{j=1}^{U} p_j \mathsf{Q}_{ji}\right| \leq \sum_{j=1}^{U} \left|p_j\right| \left|\mathsf{Q}_{ji}\right| \leq |p_i| \sum_{j=1}^{U} \left|\mathsf{Q}_{ji}\right| = \left|p_i\right|.$$

This implies that $|\lambda| \leq 1$.

(We used the symmetry of the matrix, in implying that Q eigenvalues are all real numbers.) ■

Lemma 18.2.2. *Let Q be a symmetric transition matrix, then for any $p \in \mathbb{R}^n$, we have that $\|p\mathsf{Q}\|_2 \leq \|p\|_2$.*

Proof: Let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ denote the orthonormal eigenvector basis of Q, with eigenvalues $1 = \lambda_1, \ldots, \lambda_n$. Write $p = \sum_i \alpha_i v_i$, and observe that

$$\left\|p\mathsf{Q}\right\|_2 = \left\|\sum_i \alpha_i v_i \mathsf{Q}\right\|_2 = \left\|\sum_i \alpha_i \lambda_i v_i\right\|_2 = \sqrt{\sum_i \alpha_i^2 \lambda_i^2} \leq \sqrt{\sum_i \alpha_i^2} = \left\|p\right\|_2,$$

since $|\lambda_i| \leq 1$, for $i = 1, \ldots, n$, by Lemma 18.2.1. ■

Lemma 18.2.3. *Let $\mathsf{B} = \mathsf{Q}^\beta$ be the transition matrix of the graph G^β. For all vectors $p \in \mathbb{R}^n$, we have: (i) $\|p\mathsf{B}\mathsf{W}\|_2 \leq \|p\|_2$, and (ii) $\left\|p\mathsf{B}\overline{\mathsf{W}}\right\| \leq \|p\| /5$.*

Proof: (i) Since multiplying a vector by W has the effect of zeroing out some coordinates, its clear that it can not enlarge the norm of a matrix. As such, $\|p\mathsf{B}\mathsf{W}\|_2 \leq \|p\mathsf{B}\|_2 \leq \|p\|_2$ by Lemma 18.2.2.

(ii) Write $p = \sum_i \alpha_i v_i$, where $v_1, \ldots, v_n$ is the orthonormal basis of Q (and thus also of B), with eigenvalues $1 = \widehat{\lambda_1}, \ldots, \widehat{\lambda_n}$. We remind the reader that $v_1 = (1, 1, \ldots, 1)/\sqrt{n}$. Since $\overline{\mathsf{W}}$ zeroes out at least 99/100 of the entries of a vectors it is multiplied by (and copy the rest as they are), we have that $\left\|v_1\overline{\mathsf{W}}\right\| \leq \sqrt{(n/100)(1/\sqrt{n})^2} \leq 1/10 = \|v_1\|/10$. Now, for any $x \in \mathbb{R}^U$, we have $\left\|x\overline{\mathsf{W}}\right\| \leq \|x\|$. As such, we have that

$$\left\|p\mathsf{B}\overline{\mathsf{W}}\right\|_2 = \left\|\sum_i \alpha_i v_i \mathsf{B}\overline{\mathsf{W}}\right\|_2 \leq \left\|\alpha_1 v_1 \mathsf{B}\overline{\mathsf{W}}\right\| + \left\|\sum_{i=2}^{U} \alpha_i v_i \mathsf{B}\overline{\mathsf{W}}\right\|$$

$$\leq \left\|\alpha_1 v_1 \overline{\mathsf{W}}\right\| + \left\|\left(\sum_{i=2}^{U} \alpha_i v_i \widehat{\lambda}_i^{\beta}\right)\overline{\mathsf{W}}\right\| \leq \frac{|\alpha_1|}{10} + \left\|\sum_{i=2}^{U} \alpha_i v_i \widehat{\lambda}_i^{\beta}\right\|$$
$$\leq \frac{|\alpha_1|}{10} + \sqrt{\sum_{i=2}^{U}\left(\alpha_i \widehat{\lambda}_i^{\beta}\right)^2} \leq \frac{|\alpha_1|}{10} + \frac{1}{10}\sqrt{\sum_{i=2}^{U}\alpha_i^2} \leq \frac{\|p\|}{10} + \frac{1}{10}\|p\| \leq \frac{\|p\|}{5},$$

since $\left|\lambda_i^{\beta}\right| \leq 1/10$, for $i = 2, \ldots, n$. ■

Consider the strings $r_0, \ldots, r_\nu$. For each one of these strings, we can write down whether its a "good" string (i.e., **Alg** return the correct result), or a bad string. This results in a binary pattern $b_0, \ldots, b_k$. Given a distribution $p \in \mathbb{R}^U$ on the states of the graph, its natural to ask what is the probability of being in a "good" state. Clearly, this is the quantity $\|p\mathsf{W}\|_1$. Thus, if we are interested in the probability of a specific pattern, then we should start with the initial distribution p^0, truncate away the coordinates that represent an invalid state, apply the transition matrix, again truncate away forbidden coordinates, and repeat in this fashion till we exhaust the pattern. Clearly, the ℓ_1-norm of the resulting vector is the probability of this pattern. To this end, given a pattern $b_0, \ldots, b_k$, let $\mathcal{S} = \langle S_0, \ldots, S_\nu \rangle$ denote the corresponding sequence of "truncating" matrices (i.e., S_i is either W or $\overline{\mathsf{W}}$). Formally, we set $S_i = \mathsf{W}$ if **Alg**(x, r_i) returns the correct answer, and set $S_i = \overline{\mathsf{W}}$ otherwise.

The above argument implies the following lemma.

Lemma 18.2.4. *For any fixed pattern $b_0, \ldots, b_\nu$ the probability of the random walk to generate this pattern of random strings is $\left\|p^{(0)} S_0 \mathsf{B} S_1 \ldots \mathsf{B} S_\nu\right\|_1$, where $\mathcal{S} = \langle S_0, \ldots, S_\nu \rangle$ is the sequence of W and $\overline{\mathsf{W}}$ encoded by this pattern.*

Theorem 18.2.5. *The probability that the majority of the outputs **Alg**(x, r_0), **Alg**(x, r_1), $\ldots$, **Alg**(x, r_k) is incorrect is at most $1/2^k$.*

Proof: The majority is wrong, only if (at least) half the elements of the sequence $\mathcal{S} = \langle S_0, \ldots, S_\nu \rangle$ belong to $\overline{\mathsf{W}}$. Fix such a "bad" sequence $\mathcal{S}$, and observe that the distributions we work with are vectors in $\mathbb{R}^U$. As such, if p^0 is the initial distribution, then we have that

$$\mathbf{Pr}\left[\mathcal{S}\right] = \left\|p^{(0)} S_0 \mathsf{B} S_1 \ldots \mathsf{B} S_\nu\right\|_1 \leq \sqrt{U}\left\|p^{(0)} S_0 \mathsf{B} S_1 \ldots \mathsf{B} S_\nu\right\|_2 \leq \sqrt{U}\frac{1}{5^{\nu/2}}\left\|p^{(0)}\right\|_2,$$

by Lemma 18.3.1 below (i.e., Cauchy-Schwarz inequality) and by repeatedly applying Lemma 18.2.3, since half of the sequence $\mathcal{S}$ are $\overline{\mathsf{W}}$, and the rest are W. The distribution $p^{(0)}$ was uniform, which implies that $\left\|p^{(0)}\right\|_2 = 1/\sqrt{U}$. As such, let $\mathcal{S}$ be the set of all bad patterns (there are $2^{\nu-1}$ such "bad" patterns). We have

$$\mathbf{Pr}\Big[\text{ majority is bad }\Big] \leq 2^k \sqrt{U}\frac{1}{5^{\nu/2}}\left\|p^{(0)}\right\|_2 = (4/5)^{\nu/2} = (4/5)^{\alpha k/2} \leq \frac{1}{2^k},$$

for $\alpha = 7$. ■

18.3. Some standard inequalities

Lemma 18.3.1. *For any vector $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$, we have that $\|\mathbf{v}\|_1 \leq \sqrt{d}\,\|\mathbf{v}\|_2$.*

Proof: We can safely assume all the coordinates of $\mathbf{v}$ are positive. Now,

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{d} v_i = \sum_{i=1}^{d} v_i \cdot 1 = |\mathbf{v} \cdot (1, 1, \ldots, 1)| \leq \sqrt{\sum_{i=1}^{d} v_i^2}\sqrt{\sum_{i=1}^{d} 1^2} = \sqrt{d}\,\|\mathbf{v}\|,$$

by the Cauchy-Schwarz inequality. ■

18.4. Tools from previous lecture

Theorem 18.4.1 (Fundamental theorem of Markov chains). *Any irreducible, finite, and aperiodic Markov chain has the following properties.*

(i) All states are ergodic.

(ii) There is a unique stationary distribution π such that, for $1 \leq i \leq n$, we have $\pi_i > 0$.

(iii) For $1 \leq i \leq n$, we have $\mathtt{f}_{ii} = 1$ and $\mathtt{h}_{ii} = 1/\pi_i$.

(iv) Let $N(i,t)$ be the number of times the Markov chain visits state i in t steps. Then

$$\lim_{t \to \infty} \frac{N(i,t)}{t} = \pi_i.$$

Namely, independent of the starting distribution, the process converges to the stationary distribution.

Definition 18.4.2. Given a random walk matrix Q associated with a d-regular graph, let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ denote the *orthonormal eigenvector basis* defined by Q. That is, $v_1, \ldots, v_n$ is an orthonormal basis for $\mathbb{R}^n$, where all these vectors are eigenvectors of Q and $v_1 = 1^n/\sqrt{n}$. Furthermore, let $\widehat{\lambda_i}$ denote the ith eigenvalue of Q, associated with the eigenvector v_i, such that $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \cdots \geq \widehat{\lambda_n}$.

Definition 18.4.3. The class **BPP** (for Bounded-error Probabilistic Polynomial time) is the class of languages that have a randomized algorithm **Alg** with worst case polynomial running time such that for any input $x \in \Sigma^*$, we have

(i) If $x \in L$ then $\mathbf{Pr}[\mathbf{Alg}(x) \text{ accepts}] \geq 3/4$.

(ii) If $x \notin L$ then $\mathbf{Pr}[\mathbf{Alg}(x) \text{ accepts}] \leq 1/4$.

Chapter 19

The Johnson-Lindenstrauss Lemma

By Sariel Har-Peled, December 30, 2015[①]

> Dixon was alive again. Consciousness was upon him before he could get out of the way; not for him the slow, gracious wandering from the halls of sleep, but a summary, forcible ejection. He lay sprawled, too wicked to move, spewed up like a broken spider-crab on the tarry shingle of the morning. The light did him harm, but not as much as looking at things did; he resolved, having done it once, never to move his eyeballs again. A dusty thudding in his head made the scene before him beat like a pulse. His mouth had been used as a latrine by some small creature of the night, and then as its mausoleum. During the night, too, he'd somehow been on a cross-country run and then been expertly beaten up by secret police. He felt bad.
>
> – Lucky Jim, Kingsley Amis.

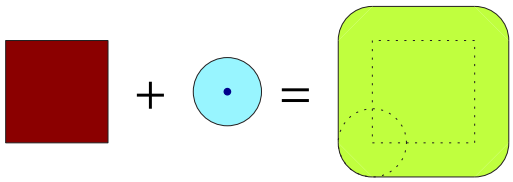
In this chapter, we will prove that given a set P of n points in $\mathbb{R}^d$, one can reduce the dimension of the points to $k = O(\varepsilon^{-2} \log n)$ such that distances are $1 \pm \varepsilon$ preserved. Surprisingly, this reduction is done by randomly picking a subspace of k dimensions and projecting the points into this random subspace. One way of thinking about this result is that we are "compressing" the input of size nd (i.e., n points with d coordinates) into size $O(n\varepsilon^{-2} \log n)$, while (approximately) preserving distances.

19.1. The Brunn-Minkowski inequality

For a set $A \subseteq \mathbb{R}^d$, an a point $\mathsf{p} \in \mathbb{R}^d$, let $A + \mathsf{p}$ denote the translation of A by p. Formally, $A + \mathsf{p} = \left\{ \mathsf{q} + \mathsf{p} \;\middle|\; \mathsf{q} \in A \right\}$.

Definition 19.1.1. For two sets A and B in $\mathbb{R}^n$, let $A + B$ denote the *Minkowski sum* of A and B. Formally,

$$A + B = \left\{ a + b \;\middle|\; a \in A, b \in B \right\} = \bigcup_{\mathsf{p} \in A} (\mathsf{p} + B).$$



Remark 19.1.2. It is easy to verify that if A' and B' are translated copies of A and B (that is, $A' = A + \mathsf{p}$ and $B = B + \mathsf{q}$, for some points $\mathsf{p}, \mathsf{q} \in \mathbb{R}^d$), respectively, then $A' + B'$ is a translated copy of $A + B$. In particular, since volume is preserved under translation, we have that $\mathrm{vol}\Big(A' + B'\Big) = \mathrm{vol}\Big((A + B) + \mathsf{p} + \mathsf{q}\Big) = \mathrm{vol}\Big(A + B\Big)$, where $\mathrm{vol}(X)$ is the *volume* (i.e., measure) of the set X.

Our purpose here is to prove the following theorem.

Theorem 19.1.3 (Brunn-Minkowski inequality). *Let A and B be two non-empty compact sets in $\mathbb{R}^n$. Then*

$$\operatorname{vol}(A+B)^{1/n} \geq \operatorname{vol}(A)^{1/n} + \operatorname{vol}(B)^{1/n}.$$

Definition 19.1.4. A set $A \subseteq \mathbb{R}^n$ is a *brick set* if it is the union of finitely many (close) axis parallel boxes with disjoint interiors.

It is intuitively clear, by limit arguments, that proving Theorem 19.1.3 for brick sets will imply it for the general case.

Lemma 19.1.5 (Brunn-Minkowski inequality for Brick Sets). *Let A and B be two non-empty brick sets in $\mathbb{R}^n$. Then*

$$\Big(\operatorname{vol}(A+B)\Big)^{1/n} \geq \operatorname{vol}(A)^{1/n} + \operatorname{vol}(B)^{1/n}.$$

Proof: By induction on the number k of bricks in A and B. If $k = 2$ then A and B are just bricks, with dimensions $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, respectively. In this case, the dimensions of $A + B$ are $a_1 + b_1, \ldots, a_n + b_n$, as can be easily verified. Thus, we need to prove that $(\prod_{i=1}^n a_i)^{1/n} + (\prod_{i=1}^n b_i)^{1/n} \leq \Big(\prod_{i=1}^n (a_i + b_i)\Big)^{1/n}$.Dividing the left side by the right side, we have

$$\left(\prod_{i=1}^n \frac{a_i}{a_i+b_i}\right)^{1/n} + \left(\prod_{i=1}^n \frac{b_i}{a_i+b_i}\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^n \frac{a_i}{a_i+b_i} + \frac{1}{n}\sum_{i=1}^n \frac{b_i}{a_i+b_i} = 1,$$

by the generalized arithmetic-geometric mean inequality[②], and the claim follows for this case.

Now let $k > 2$ and suppose that the Brunn-Minkowski inequality holds for any pair of brick sets with fewer than k bricks (together). Let A and B be a pair of sets having k bricks together, the A has at least two (disjoint) bricks. However, this implies that there is an axis parallel hyperplane h that separates the interior of one brick of A from the interior of another brick of A (the hyperplane h might intersect other bricks of A). Assume that h is the hyperplane $x_1 = 0$ (this can be achieved by translation and renaming of coordinates).

Let $\overline{A^+} = A \cap h^+$ and $\overline{A^-} = A \cap h^-$, where h^+ and h^- are the two open half spaces induced by h. Let A^+ and A^- be the closure of $\overline{A^+}$ and $\overline{A^-}$, respectively. Clearly, A^+ and A^- are both brick sets with (at least) one fewer brick than A.

Next, observe that the claim is translation invariant (see Remark 19.1.2), and as such, let us translate B so that its volume is split by h in the same ratio A's volume is being split. Denote the two parts of B by B^+ and B^-, respectively. Let $\rho = \operatorname{vol}(A^+)/\operatorname{vol}(A) = \operatorname{vol}(B^+)/\operatorname{vol}(B)$ (if $\operatorname{vol}(A) = 0$ or $\operatorname{vol}(B) = 0$ the claim trivially holds).

Observe, that $A^+ + B^+ \subseteq A + B$, and it lies on one side of h (since $h \equiv (x_1 = 0)$), and similarly $A^- + B^- \subseteq A + B$ and it lies on the other side of h. Thus, by induction and since $A^+ + B^+$ and $A^- + B^-$ are interior disjoint, we have

$$\begin{aligned}\operatorname{vol}(A+B) &\geq \operatorname{vol}(A^+ + B^+) + \operatorname{vol}(A^- + B^-) \\ &\geq \Big(\operatorname{vol}(A^+)^{1/n} + \operatorname{vol}(B^+)^{1/n}\Big)^n + \Big(\operatorname{vol}(A^-)^{1/n} + \operatorname{vol}(B^-)^{1/n}\Big)^n\end{aligned}$$

[②]Here is a proof of the generalized form: Let $x_1, \ldots, x_n$ be n positive real numbers. Consider the quantity $R = x_1 x_2 \cdots x_n$. If we fix the sum of the n numbers to be equal to α, then R is maximized when all the x_is are equal. Thus, $\sqrt[n]{x_1 x_2 \cdots x_n} \leq \sqrt[n]{(\alpha/n)^n} = \alpha/n = (x_1 + \cdots + x_n)/n$.

$$
\begin{aligned}
&= \Big[\rho^{1/n}\mathrm{vol}(A)^{1/n} + \rho^{1/n}\mathrm{vol}(B)^{1/n}\Big]^n \\
&\qquad \Big[(1-\rho)^{1/n}\mathrm{vol}(A)^{1/n} + (1-\rho)^{1/n}\mathrm{vol}(B)^{1/n}\Big]^n \\
&= (\rho + (1-\rho))\Big[\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}\Big]^n \\
&= \Big[\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}\Big]^n,
\end{aligned}
$$

establishing the claim. ■

Proof of ***Theorem 19.1.3***: Let $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_i \subseteq \cdots$ be a sequence of finite brick sets, such that $\bigcup_i A_i = A$, and similarly let $B_1 \subseteq B_2 \subseteq \cdots \subseteq B_i \subseteq \cdots$ be a sequence of finite brick sets, such that $\bigcup_i B_i = B$. By the definition of volume[3],we have that $\lim_{i\to\infty} \mathrm{vol}(A_i) = \mathrm{vol}(A)$ and $\lim_{i\to\infty} \mathrm{vol}(B_i) = \mathrm{vol}(B)$.

We claim that $\lim_{i\to\infty} \mathrm{vol}(A_i + B_i) = \mathrm{vol}(A + B)$. Indeed, consider any point $z \in A + B$, and let $u \in A$ and $v \in B$ be such that $u + v = z$. By definition, there exists an i, such that for all $j > i$ we have $u \in A_j$, $v \in B_j$, and as such $z \in A_j + B_j$. Thus, $A + B \subseteq \cup_j (A_j + B_j)$ and $\cup_j (A_j + B_j) \subseteq \cup_j (A + B) \subseteq A + B$; namely, $\cup_j (A_j + B_j) = A + B$.

Furthermore, for any $i > 0$, since A_i and B_i are brick sets, we have

$$
\mathrm{vol}(A_i + B_i)^{1/n} \geq \mathrm{vol}(A_i)^{1/n} + \mathrm{vol}(B_i)^{1/n},
$$

by Lemma 19.1.5. Thus,

$$
\begin{aligned}
\mathrm{vol}\Big(A + B\Big)^{1/n} &= \lim_{i\to\infty} \mathrm{vol}(A_i + B_i)^{1/n} \geq \lim_{i\to\infty} \Big(\mathrm{vol}(A_i)^{1/n} + \mathrm{vol}(B_i)^{1/n}\Big) \\
&= \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}.
\end{aligned}
$$

■

Theorem 19.1.6 (Brunn-Minkowski for slice volumes.). *Let $\mathcal{P}$ be a convex set in $\mathbb{R}^{n+1}$, and let $A = \mathcal{P} \cap (x_1 = a)$, $B = \mathcal{P} \cap (x_1 = b)$ and $C = \mathcal{P} \cap (x_1 = c)$ be three slices of A, for $a < b < c$. We have* $\mathrm{vol}(B) \geq \min(\mathrm{vol}(A), \mathrm{vol}(C))$. *Specifically, consider the function*

$$
v(t) = \left(\mathrm{vol}\Big(\mathcal{P} \cap (x_1 = t)\Big)\right)^{1/n},
$$

and let $\mathcal{I} = [t_{min}, t_{max}]$ be the interval where the hyperplane $x_1 = t$ intersects $\mathcal{P}$. Then, $v(t)$ is concave on $\mathcal{I}$.

Proof: If a or c are outside $\mathcal{I}$, then $\mathrm{vol}(A) = 0$ or $\mathrm{vol}(C) = 0$, respectively, and then the claim trivially holds.

Otherwise, let $\alpha = (b - a)/(c - a)$. We have that $b = (1 - \alpha) \cdot a + \alpha \cdot c$, and by the convexity of $\mathcal{P}$, we have $(1 - \alpha)A + \alpha C \subseteq B$. Thus, by Theorem 19.1.3 we have

$$
\begin{aligned}
v(b) = \mathrm{vol}(B)^{1/n} &\geq \mathrm{vol}((1-\alpha)A + \alpha C)^{1/n} \geq \mathrm{vol}((1-\alpha)A)^{1/n} + \mathrm{vol}(\alpha C)^{1/n} \\
&= ((1-\alpha)^n \mathrm{vol}(A))^{1/n} + (\alpha^n \mathrm{vol}(C))^{1/n} \\
&= (1-\alpha) \cdot \mathrm{vol}(A)^{1/n} + \alpha \cdot \mathrm{vol}(C)^{1/n} \\
&= (1-\alpha)v(a) + \alpha v(c).
\end{aligned}
$$

Namely, $v(\cdot)$ is concave on $\mathcal{I}$, and in particular $v(b) \geq \min(v(a), v(c))$, which in turn implies that $\mathrm{vol}(B) = v(b)^n \geq (\min(v(a), v(c)))^n = \min(\mathrm{vol}(A), \mathrm{vol}(C))$, as claimed. ■

[3] This is the standard definition in measure theory of volume. The reader unfamiliar with this fanfare can either consult a standard text on the topic, or take it for granted as this is intuitively clear.

Corollary 19.1.7. *For A and B compact sets in $\mathbb{R}^n$, the following holds* $\mathrm{vol}((A+B)/2) \geq \sqrt{\mathrm{vol}(A)\mathrm{vol}(B)}$.

Proof: We have that

$$\begin{aligned}\mathrm{vol}\Big((A+B)/2\Big)^{1/n} &= \mathrm{vol}\Big(A/2+B/2\Big)^{1/n} \geq \mathrm{vol}\Big(A/2\Big)^{1/n} + \mathrm{vol}\Big(B/2\Big)^{1/n} = \Big(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}\Big)/2 \\ &\geq \sqrt{\mathrm{vol}(A)^{1/n}\mathrm{vol}(B)^{1/n}}\end{aligned}$$

by Theorem 19.1.3, and since $(a+b)/2 \geq \sqrt{ab}$ for any $a, b \geq 0$. The claim now follows by raising this inequality to the power n. ■

19.1.1. The Isoperimetric Inequality

The following is not used anywhere else and is provided because of its mathematical elegance. The skip-able reader can thus employ their special gift and move on to Section 19.2.

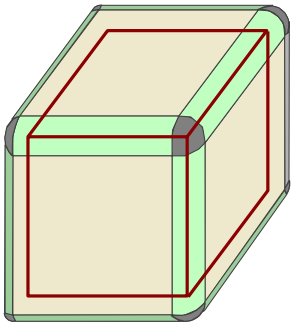
The *isoperimetric inequality* states that among all convex bodies of a fixed surface area, the ball has the largest volume (in particular, the unit circle is the largest area planar region with perimeter 2π). This problem can be traced back to antiquity, in particular Zenodorus (200–140 BC) wrote a monograph (which was lost) that seemed to have proved the claim in the plane for some special cases. The first formal proof for the planar case was done by Steiner in 1841. Interestingly, the more general claim is an easy consequence of the Brunn-Minkowski inequality.

Let K be a convex body in $\mathbb{R}^n$ and $\mathbf{b}$ be the n dimensional ball of radius one centered at the origin. Let $\mathsf{S}(X)$ denote the surface area of a compact set $X \subseteq \mathbb{R}^n$. The *isoperimetric inequality* states that

$$\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(\mathbf{b})}\right)^{1/n} \leq \left(\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})}\right)^{1/(n-1)}. \tag{19.1}$$

Namely, the left side is the radius of a ball having the same volume as K, and the right side is the radius of a sphere having the same surface area as K. In particular, if we scale K so that its surface area is the same as $\mathbf{b}$, then the above inequality implies that $\mathrm{vol}(K) \leq \mathrm{vol}(\mathbf{b})$.

To prove Eq. (19.1), observe that $\mathrm{vol}(\mathbf{b}) = \mathsf{S}(\mathbf{b})/n$[④]. Also, observe that $K + \varepsilon\mathbf{b}$ is the body K together with a small "atmosphere" around it of thickness ε. In particular, the volume of this "atmosphere" is (roughly) $\varepsilon\,\mathsf{S}(K)$ (in fact, Minkowski defined the surface area of a convex body to be the limit stated next).



Formally, we have

$$\begin{aligned}\mathsf{S}(K) &= \lim_{\varepsilon\to 0+} \frac{\mathrm{vol}(K+\varepsilon\mathbf{b}) - \mathrm{vol}(K)}{\varepsilon} \\ &\geq \lim_{\varepsilon\to 0+} \frac{\Big(\mathrm{vol}(K)^{1/n} + \mathrm{vol}(\varepsilon\mathbf{b})^{1/n}\Big)^n - \mathrm{vol}(K)}{\varepsilon},\end{aligned}$$

by the Brunn-Minkowski inequality. Now $\mathrm{vol}(\varepsilon\mathbf{b})^{1/n} = \varepsilon\mathrm{vol}(\mathbf{b})^{1/n}$, and as such

$$\begin{aligned}\mathsf{S}(K) &\geq \lim_{\varepsilon\to 0+} \frac{\mathrm{vol}(K) + \binom{n}{1}\varepsilon\mathrm{vol}(K)^{(n-1)/n}\mathrm{vol}(\mathbf{b})^{1/n} + \binom{n}{2}\varepsilon^2\langle\cdots\rangle + \cdots + \varepsilon^n\mathrm{vol}(\mathbf{b}) - \mathrm{vol}(K)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0+} \frac{n\varepsilon\mathrm{vol}(K)^{(n-1)/n}\mathrm{vol}(\mathbf{b})^{1/n}}{\varepsilon} = n\mathrm{vol}(K)^{(n-1)/n}\mathrm{vol}(\mathbf{b})^{1/n}.\end{aligned}$$

[④]Indeed, $\mathrm{vol}(\mathbf{b}) = \int_{r=0}^{1} \mathsf{S}(\mathbf{b}) r^{n-1} dr = \mathsf{S}(\mathbf{b})/n$.

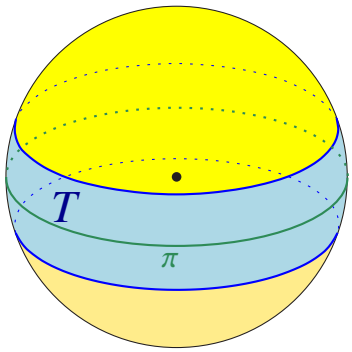
Dividing both sides by $\mathsf{S}(\mathbf{b}) = n\mathrm{vol}(\mathbf{b})$, we have

$$\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})} \geq \frac{\mathrm{vol}(K)^{(n-1)/n}}{\mathrm{vol}(\mathbf{b})^{(n-1)/n}} \qquad \Longrightarrow \qquad \left(\frac{\mathsf{S}(K)}{\mathsf{S}(\mathbf{b})}\right)^{1/(n-1)} \geq \left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(\mathbf{b})}\right)^{1/n},$$

establishing the isoperimetric inequality.

19.2. Measure Concentration on the Sphere

Let $\mathbb{S}^{(n-1)}$ be the unit sphere in $\mathbb{R}^n$. We assume there is a uniform probability measure defined over $\mathbb{S}^{(n-1)}$, such that its total measure is 1. Surprisingly, most of the mass of this measure is near the equator. Indeed, consider an arbitrary equator π on $\mathbb{S}^{(n-1)}$ (that it, it is the intersection of the sphere with a hyperplane passing through the center of ball inducing the sphere). Next, consider all the points that are in distance $\approx \ell(n) = c/n^{1/3}$ from π. The question we are interested in is what fraction of the sphere is covered by this strip T (depicted on the right).



Notice, that as the dimension increases the width $\ell(n)$ of this strip decreases. But surprisingly, despite its width becoming smaller, as the dimension increases, this strip contains a larger and larger fraction of the sphere. In particular, the total fraction of the sphere not covered by this (shrinking!) strip converges to zero.

Furthermore, counter intuitively, this is true for *any* equator. We are going to show that even a stronger result holds: The mass of the sphere is concentrated close to the boundary of any set $A \subseteq \mathbb{S}^{(n-1)}$ such that $\mathbf{Pr}[A] = 1/2$.

Before proving this somewhat surprising theorem, we will first try to get an intuition about the behavior of the hypersphere in high dimensions.

19.2.1. The strange and curious life of the hypersphere

Consider the ball of radius r in $\mathbb{R}^n$ denoted by $r\,\mathbf{b}^n$, where $\mathbf{b}^n$ is the unit radius ball centered at the origin. Clearly, $\mathrm{vol}(r\,\mathbf{b}^n) = r^n\mathrm{vol}(\mathbf{b}^n)$. Now, even if r is very close to 1, the quantity r^n might be very close to zero if n is sufficiently large. Indeed, if $r = 1-\delta$, then $r^n = (1-\delta)^n \leq \exp(-\delta n)$, which is very small if $\delta \gg 1/n$. (Here, we used the fact that $1 - x \leq e^x$, for $x \geq 0$.) Namely, for the ball in high dimensions, its mass is concentrated in a very thin shell close to its surface.

The volume of a ball and the surface area of hypersphere. Let $\mathrm{vol}(r\mathbf{b}^n)$ denote the volume of the ball of radius r in $\mathbb{R}^n$, and $\mathrm{Area}\left(r\mathbb{S}^{(n-1)}\right)$ denote the surface area of its bounding sphere (i.e., the surface area of $r\mathbb{S}^{(n-1)}$). It is known that

$$\mathrm{vol}(r\mathbf{b}^n) = \frac{\pi^{n/2}r^n}{\Gamma(n/2+1)} \quad \text{and} \quad \mathrm{Area}\left(r\mathbb{S}^{(n-1)}\right) = \frac{2\pi^{n/2}r^{n-1}}{\Gamma(n/2)},$$

where the gamma function, $\Gamma(\cdot)$, is an extension of the factorial function. Specifically, if n is even then $\Gamma(n/2+1) = (n/2)!$, and for n odd $\Gamma(n/2+1) = \sqrt{\pi}(n!!)/2^{(n+1)/2}$, where $n!! = 1\cdot 3\cdot 5\cdots n$ is the *double factorial*. The most surprising implication of these two formulas is that, as n increases, the volume of the unit ball first increases (till dimension 5 in fact) and then starts decreasing to zero.

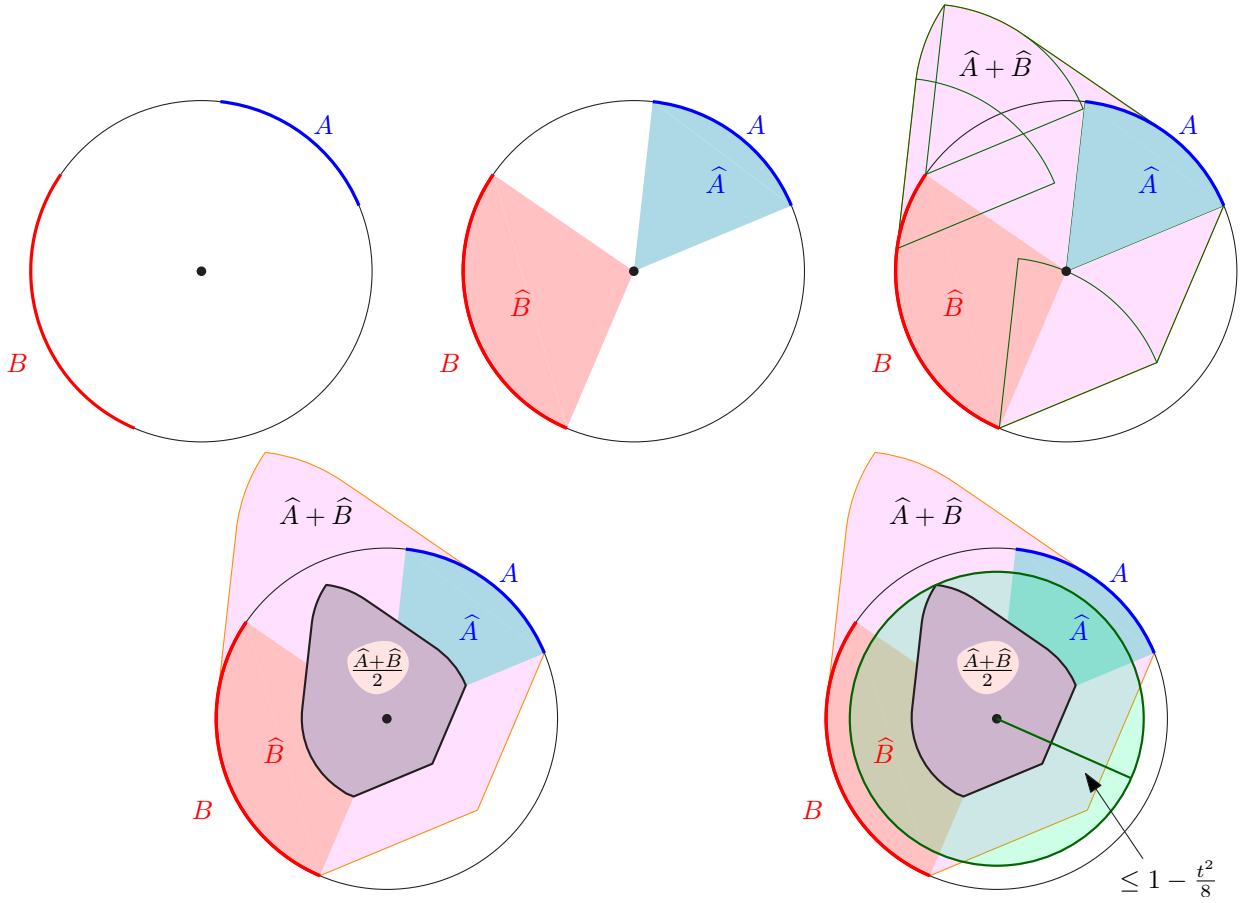
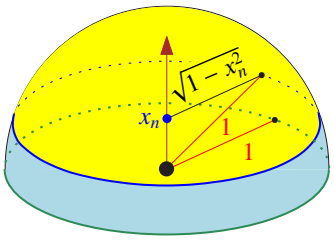


Figure 19.1: Illustration of the proof of Theorem 19.2.1.



Similarly, the surface area of the unit sphere $\mathbb{S}^{(n-1)}$ in $\mathbb{R}^n$ tends to zero as the dimension increases. To see this, compute the volume of the unit ball using an integral of its slice volume, when it is being sliced by a hyperplanes perpendicular to the nth coordinate.

We have, see figure on the right, that

$$\operatorname{vol}\left(\mathbf{b}^n\right) = \int_{x_n=-1}^{1} \operatorname{vol}\left(\sqrt{1-x_n^2}\,\mathbf{b}^{n-1}\right) dx_n = \operatorname{vol}\left(\mathbf{b}^{n-1}\right) \int_{x_n=-1}^{1} \left(1-x_n^2\right)^{(n-1)/2} dx_n,$$

Now, the integral on the right side tends to zero as n increases. In fact, for n very large, the term $\left(1-x_n^2\right)^{(n-1)/2}$ is very close to 0 everywhere except for a small interval around 0. This implies that the main contribution of the volume of the ball happens when we consider slices of the ball by hyperplanes of the form $x_n = \delta$, where δ is small.

If one has to visualize how such a ball in high dimensions looks like, it might be best to think about it as a star-like creature: It has very little mass close to the tips of any set of orthogonal directions we pick, and most of its mass somehow lies on hyperplanes close to its center.[5]

[5] In short, it looks like a Boojum [Car76].

19.2.2. Measure Concentration on the Sphere

Theorem 19.2.1 (Measure concentration on the sphere.). *Let $A \subseteq \mathbb{S}^{(n-1)}$ be a measurable set with $\mathbf{Pr}[A] \geq 1/2$, and let A_t denote the set of points of $\mathbb{S}^{(n-1)}$ in distance at most t from A, where $t \leq 2$. Then $1 - \mathbf{Pr}[A_t] \leq 2\exp\left(-nt^2/2\right)$.*

Proof: We will prove a slightly weaker bound, with $-nt^2/4$ in the exponent. Let $\widehat{A} = T(A)$, where

$$T(X) = \left\{\alpha x \;\middle|\; x \in X, \alpha \in [0,1]\right\} \subseteq \mathbf{b}^n,$$

and $\mathbf{b}^n$ is the unit ball in $\mathbb{R}^n$. We have that $\mathbf{Pr}[A] = \mu\left(\widehat{A}\right)$, where $\mu\left(\widehat{A}\right) = \mathrm{vol}\left(\widehat{A}\right)/\mathrm{vol}(\mathbf{b}^n)$[6].

Let $B = \mathbb{S}^{(n-1)} \setminus A_t$ and $\widehat{B} = T(B)$, see Figure 19.1. We have that $\|a - b\| \geq t$ for all $a \in A$ and $b \in B$. By Lemma 19.2.2 below, the set $\left(\widehat{A} + \widehat{B}\right)/2$ is contained in the ball $r\mathbf{b}^n$ centered at the origin, where $r = 1 - t^2/8$. Observe that $\mu(r\mathbf{b}^n) = \mathrm{vol}(r\mathbf{b}^n)/\mathrm{vol}(\mathbf{b}^n) = r^n = \left(1 - t^2/8\right)^n$. As such, applying the Brunn-Minkowski inequality in the form of Corollary 19.1.7, we have

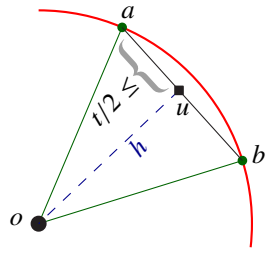
$$\left(1 - \frac{t^2}{8}\right)^n = \mu\left(r\mathbf{b}^n\right) \geq \mu\left(\frac{\widehat{A} + \widehat{B}}{2}\right) \geq \sqrt{\mu\left(\widehat{A}\right)\mu\left(\widehat{B}\right)} = \sqrt{\mathbf{Pr}[A]\,\mathbf{Pr}[B]} \geq \sqrt{\mathbf{Pr}[B]\,/2}\,.$$

Thus, $\mathbf{Pr}[B] \leq 2(1 - t^2/8)^{2n} \leq 2\exp(-2nt^2/8)$, since $1 - x \leq \exp(-x)$, for $x \geq 0$. ■

Lemma 19.2.2. *For any $\widehat{a} \in \widehat{A}$ and $\widehat{b} \in \widehat{B}$, we have* $\left\|\frac{\widehat{a} + \widehat{b}}{2}\right\| \leq 1 - \frac{t^2}{8}$.

Proof: Let $\widehat{a} = \alpha a$ and $\widehat{b} = \beta b$, where $a \in A$ and $b \in B$. We have

$$\|u\| = \left\|\frac{a+b}{2}\right\| = \sqrt{1^2 - \left\|\frac{a-b}{2}\right\|^2} \leq \sqrt{1 - \frac{t^2}{4}} \leq 1 - \frac{t^2}{8}, \tag{19.2}$$



since $\|a - b\| \geq t$. As for $\widehat{a}$ and $\widehat{b}$, assume that $\alpha \leq \beta$, and observe that the quantity $\left\|\widehat{a} + \widehat{b}\right\|$ is maximized when $\beta = 1$. As such, by the triangle inequality, we have

$$\begin{aligned}\left\|\frac{\widehat{a} + \widehat{b}}{2}\right\| &= \left\|\frac{\alpha a + b}{2}\right\| \leq \left\|\frac{\alpha(a+b)}{2}\right\| + \left\|(1-\alpha)\frac{b}{2}\right\| \\ &\leq \alpha\left(1 - \frac{t^2}{8}\right) + (1-\alpha)\frac{1}{2} = \tau,\end{aligned}$$

by Eq. (19.2) and since $\|b\| = 1$. Now, τ is a convex combination of the two numbers $1/2$ and $1 - t^2/8$. In particular, we conclude that $\tau \leq \max(1/2, 1 - t^2/8) \leq 1 - t^2/8$, since $t \leq 2$. ■

[6] This is one of these "trivial" claims that might give the reader a pause, so here is a formal proof. Pick a random point p uniformly inside the ball $\mathbf{b}^n$. Let ψ be the probability that $\mathsf{p} \in \widehat{A}$. Clearly, $\mathrm{vol}\left(\widehat{A}\right) = \psi\mathrm{vol}(\mathbf{b}^n)$. So, consider the normalized point $\mathsf{q} = \mathsf{p}/\|\mathsf{p}\|$. Clearly, $\mathsf{p} \in \widehat{A}$ if and only if $\mathsf{q} \in A$, by the definition of $\widehat{A}$. Thus, $\mu\left(\widehat{A}\right) = \mathrm{vol}\left(\widehat{A}\right)/\mathrm{vol}(\mathbf{b}^n) = \psi = \mathbf{Pr}\left[\mathsf{p} \in \widehat{A}\right] = \mathbf{Pr}[\mathsf{q} \in A] = \mathbf{Pr}[A]$, since q has a uniform distribution on the hypersphere by assumption.

19.3. Concentration of Lipschitz Functions

Consider a function $f : \mathbb{S}^{(n-1)} \to \mathbb{R}$, and imagine that we have a probability density function defined over the sphere. Let $\mathbf{Pr}[f \leq t] = \mathbf{Pr}\Big[\Big\{x \in S^{n-1} \;\Big|\; f(x) \leq t\Big\}\Big]$. We define the *median* of f, denoted by $\mathrm{med}(f)$, to be the $\sup t$, such that $\mathbf{Pr}[f \leq t] \leq 1/2$.

We define $\mathbf{Pr}[f < \mathrm{med}(f)] = \sup_{x<\mathrm{med}(f)} \mathbf{Pr}[f \leq x]$. The following is obvious but (in fact) requires a formal proof.

Lemma 19.3.1. *We have* $\mathbf{Pr}[f < \mathrm{med}(f)] \leq 1/2$ *and* $\mathbf{Pr}[f > \mathrm{med}(f)] \leq 1/2$.

Proof: Since $\bigcup_{k\geq 1}(-\infty, \mathrm{med}(f) - 1/k] = (-\infty, \mathrm{med}(f))$, we have

$$\mathbf{Pr}[f < \mathrm{med}(f)] \quad = \quad \sup_{k\geq 1} \mathbf{Pr}\left[f \leq \mathrm{med}(f) - \frac{1}{k}\right] \leq \sup_{k\geq 1} \frac{1}{2} = \frac{1}{2}.$$

The second claim follows by a symmetric argument. ■

Definition 19.3.2 (*c*-***Lipschitz***). A function $f : A \to B$ is *c*-***Lipschitz*** if, for any $x, y \in A$, we have $\|f(x) - f(y)\| \leq c\,\|x - y\|$.

Theorem 19.3.3 (Lévy's Lemma). *Let $f : \mathbb{S}^{(n-1)} \to \mathbb{R}$ be 1-Lipschitz. Then for all $t \in [0, 1]$,*

$$\mathbf{Pr}[f > \mathrm{med}(f) + t] \leq 2\exp\left(-t^2 n/2\right) \quad \textit{and} \quad \mathbf{Pr}[f < \mathrm{med}(f) - t] \leq 2\exp\left(-t^2 n/2\right).$$

Proof: We prove only the first inequality, the second follows by symmetry. Let

$$A = \left\{x \in \mathbb{S}^{(n-1)} \;\middle|\; f(x) \leq \mathrm{med}(f)\right\}.$$

By Lemma 19.3.1, we have $\mathbf{Pr}[A] \geq 1/2$. Consider a point $x \in A_t$, where A_t is as defined in Theorem 19.2.1. Let $\mathsf{nn}(x)$ be the nearest point in A to x. We have by definition that $\|x - \mathsf{nn}(x)\| \leq t$. As such, since f is 1-Lipschitz and $\mathsf{nn}(x) \in A$, we have that

$$f(x) \leq f(\mathsf{nn}(x)) + \|\mathsf{nn}(x) - x\| \leq \mathrm{med}(f) + t.$$

Thus, by Theorem 19.2.1, we get $\mathbf{Pr}[f > \mathrm{med}(f) + t] \leq 1 - \mathbf{Pr}[A_t] \leq 2\exp\left(-t^2 n/2\right)$. ■

19.4. The Johnson-Lindenstrauss Lemma

Lemma 19.4.1. *For a unit vector $x \in \mathbb{S}^{(n-1)}$, let*

$$f(x) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$$

be the length of the projection of x into the subspace formed by the first k coordinates. Let x be a vector randomly chosen with uniform distribution from $\mathbb{S}^{(n-1)}$. Then $f(x)$ is sharply concentrated. Namely, there exists $m = m(n, k)$ such that

$$\mathbf{Pr}[f(x) \geq m + t] \leq 2\exp(-t^2 n/2) \quad \textit{and} \quad \mathbf{Pr}[f(x) \leq m - t] \leq 2\exp(-t^2 n/2),$$

for any $t \in [0, 1]$. Furthermore, for $k \geq 10 \ln n$, we have $m \geq \frac{1}{2}\sqrt{k/n}$.

Proof: The orthogonal projection $p : \mathbb{R}^n \to \mathbb{R}^k$ given by $p(x_1, \ldots, x_n) = (x_1, \ldots, x_k)$ is 1-Lipschitz (since projections can only shrink distances, see Exercise 19.6.4). As such, $f(x) = \|p(x)\|$ is 1-Lipschitz, since for any x, y we have

$$|f(x) - f(y)| = |\|p(x)\| - \|p(y)\|| \leq \|p(x) - p(y)\| \leq \|x - y\|,$$

by the triangle inequality and since p is 1-Lipschitz. Theorem 19.3.3 (i.e., Lévy's lemma) gives the required tail estimate with $m = \operatorname{med}(f)$.

Thus, we only need to prove the lower bound on m. For a random $x = (x_1, \ldots, x_n) \in \mathbb{S}^{(n-1)}$, we have $\mathbf{E}\left[\|x\|^2\right] = 1$. By linearity of expectations, and symmetry, we have $1 = \mathbf{E}\left[\|x\|^2\right] = \mathbf{E}\left[\sum_{i=1}^n x_i^2\right] = \sum_{i=1}^n \mathbf{E}\left[x_i^2\right] = n\,\mathbf{E}\left[x_j^2\right]$, for any $1 \leq j \leq n$. Thus, $\mathbf{E}\left[x_j^2\right] = 1/n$, for $j = 1, \ldots, n$. Thus,

$$\mathbf{E}\left[(f(x))^2\right] = \mathbf{E}\left[\sum_{i=1}^k x_i^2\right] = \sum_{i=1}^k \mathbf{E}[x_i] = \frac{k}{n},$$

by linearity of expectation.

We next use that f is concentrated, to show that f^2 is also relatively concentrated. For any $t \geq 0$, we have

$$\frac{k}{n} = \mathbf{E}\left[f^2\right] \leq \mathbf{Pr}[f \leq m + t]\,(m + t)^2 + \mathbf{Pr}[f \geq m + t] \cdot 1 \leq 1 \cdot (m + t)^2 + 2\exp(-t^2 n/2),$$

since $f(x) \leq 1$, for any $x \in \mathbb{S}^{(n-1)}$. Let $t = \sqrt{k/5n}$. Since $k \geq 10 \ln n$, we have that $2\exp(-t^2 n/2) \leq 2/n$. We get that

$$\frac{k}{n} \leq \left(m + \sqrt{k/5n}\right)^2 + 2/n.$$

Implying that $\sqrt{(k-2)/n} \leq m + \sqrt{k/5n}$, which in turn implies that $m \geq \sqrt{(k-2)/n} - \sqrt{k/5n} \geq \frac{1}{2}\sqrt{k/n}$. ■

Next, we would like to argue that given a fixed vector, projecting it down into a random k-dimensional subspace results in a random vector such that its length is highly concentrated. This would imply that we can do dimension reduction and still preserve distances between points that we care about.

To this end, we would like to flip Lemma 19.4.1 around. Instead of randomly picking a point and projecting it down to the first k-dimensional space, we would like x to be fixed, and randomly pick the k-dimensional subspace we project into. However, we need to pick this random k-dimensional space carefully. Indeed, if we rotate this random subspace, by a transformation T, so that it occupies the first k dimensions, then the point $T(x)$ needs to be uniformly distributed on the hypersphere if we want to use Lemma 19.4.1.

As such, we would like to randomly pick a rotation of $\mathbb{R}^n$. This maps the standard orthonormal basis into a randomly rotated orthonormal space. Taking the subspace spanned by the first k vectors of the rotated basis results in a k-dimensional random subspace. Such a rotation is an orthonormal matrix with determinant 1. We can generate such a matrix, by randomly picking a vector $e_1 \in \mathbb{S}^{(n-1)}$. Next, we set e_1 as the first column of our rotation matrix, and generate the other $n - 1$ columns, by generating recursively $n - 1$ orthonormal vectors in the space orthogonal to e_1.

Remark 19.4.2 (Generating a random point on the sphere.). At this point, the reader might wonder how do we pick a point uniformly from the unit hypersphere. The idea is to pick a point from the multi-dimensional normal distribution $N^n(0, 1)$, and normalizing it to have length 1. Since the multi-dimensional normal distribution has the density function

$$(2\pi)^{-n/2} \exp\left(-(x_1^2 + x_2^2 + \cdots + x_n^2)/2\right),$$

which is symmetric (i.e., all the points in distance r from the origin have the same distribution), it follows that this indeed generates a point randomly and uniformly on $\mathbb{S}^{(n-1)}$.

Generating a vector with multi-dimensional normal distribution, is no more than picking each coordinate according to the normal distribution, see Lemma 19.7.1$_{\text{p13}}$. Given a source of random numbers according to the uniform distribution, this can be done using a $O(1)$ computations per coordinate, using the Box-Muller transformation [BM58]. Overall, each random vector can be generated in $O(n)$ time.

Since projecting down n-dimensional normal distribution to the lower dimensional space yields a normal distribution, it follows that generating a random projection, is no more than randomly picking n vectors according to the multidimensional normal distribution $v_1, \ldots, v_n$. Then, we orthonormalize them, using Graham-Schmidt, where $\widehat{v_1} = v_1 / \|v_1\|$, and $\widehat{v_i}$ is the normalized vector of $v_i - w_i$, where w_i is the projection of v_i to the space spanned by $v_1, \ldots, v_{i-1}$.

Taking those vectors as columns of a matrix, generates a matrix A, with determinant either 1 or -1. We multiply one of the vectors by -1 if the determinant is -1. The resulting matrix is a random rotation matrix.

We can now restate Lemma 19.4.1 in the setting where the vector is fixed and the projection is into a random subspace.

Lemma 19.4.3. *Let $x \in \mathbb{S}^{(n-1)}$ be an arbitrary unit vector, and consider a random k dimensional subspace $\mathcal{F}$, and let $f(x)$ be the length of the projection of x into $\mathcal{F}$. Then, there exists $m = m(n,k)$ such that*

$$\mathbf{Pr}[f(x) \geq m + t] \leq 2\exp(-t^2 n/2) \quad \text{and} \quad \mathbf{Pr}[f(x) \leq m - t] \leq 2\exp(-t^2 n/2),$$

for any $t \in [0,1]$. Furthermore, for $k \geq 10 \ln n$, we have $m \geq \frac{1}{2}\sqrt{k/n}$.

Proof: Let v_i be the ith orthonormal vector having 1 at the ith coordinate. Let M be a random translation of space generated as described above. Clearly, for arbitrary fixed unit vector x, the vector $\mathsf{M}x$ is distributed uniformly on the sphere. Now, the ith column of the matrix M is the random vector e_i, and $\mathsf{M}^T v_i = e_i$. As such, we have

$$\langle \mathsf{M}x, v_i \rangle = (\mathsf{M}x)^T v_i = x^T \mathsf{M}^T v_i = x^T e_i = \langle x, e_i \rangle .$$

In particular, treating $\mathsf{M}x$ as a random vector, and projecting it on the first k coordinates, we have that

$$f(x) = \sqrt{\sum_{i=1}^{k} \langle \mathsf{M}x, v_i \rangle^2} = \sqrt{\sum_{i=1}^{k} \langle x, e_i \rangle^2}.$$

But $e_1, \ldots, e_k$ is just an orthonormal basis of a random k-dimensional subspace. As such, the expression on the right is the length of the projection of x into a k-dimensional random subspace. As such, the length of the projection of x into a random k-dimensional subspace has exactly the same distribution as the length of the projection of a random vector into the first k coordinates. The claim now follows by Lemma 19.4.1. ■

Definition 19.4.4. The mapping $f : \mathbb{R}^n \to \mathbb{R}^k$ is called ***K-bi-Lipschitz*** for a subset $X \subseteq \mathbb{R}^n$ if there exists a constant $c > 0$ such that

$$cK^{-1} \cdot \|\mathsf{p} - \mathsf{q}\| \leq \|f(\mathsf{p}) - f(\mathsf{q})\| \leq c \cdot \|\mathsf{p} - \mathsf{q}\| ,$$

for all $\mathsf{p}, \mathsf{q} \in X$.

The least K for which f is K-bi-Lipschitz is called the *distortion* of f, and is denoted dist(f). We will refer to f as a ***K-embedding*** of X.

Remark 19.4.5. Let $X \subseteq \mathbb{R}^m$ be a set of n points, where m potentially might be much larger than n. Observe, that in this case, since we only care about the inter-point distances of points in X, we can consider X to be a set of points lying in the affine subspace $\mathcal{F}$ spanned by the points of X. Note, that this subspace has dimension $n-1$. As such, each point of X be interpreted as $n-1$ dimensional point in $\mathcal{F}$. Namely, we can assume, for our purposes, that the set of n points in Euclidean space we care about lies in $\mathbb{R}^n$ (in fact, $\mathbb{R}^{n-1}$).

Note, that if $m < n$ we can always pad all the coordinates of the points of X by zeros, such that the resulting point set lies in $\mathbb{R}^n$.

Theorem 19.4.6 (Johnson-Lindenstrauss lemma.). *Let X be an n-point set in a Euclidean space, and let $\varepsilon \in (0,1]$ be given. Then there exists a $(1+\varepsilon)$-embedding of X into $\mathbb{R}^k$, where $k = O(\varepsilon^{-2}\log n)$.*

Proof: By Remark 19.4.5, we can assume that $X \subseteq \mathbb{R}^n$. Let $k = 200\varepsilon^{-2}\ln n$. Assume $k < n$, and let $\mathcal{F}$ be a random k-dimensional linear subspace of $\mathbb{R}^n$. Let $P_{\mathcal{F}} : \mathbb{R}^n \to \mathcal{F}$ be the orthogonal projection operator of $\mathbb{R}^n$ into $\mathcal{F}$. Let m be the number around which $\|P_{\mathcal{F}}(x)\|$ is concentrated, for $x \in \mathbb{S}^{(n-1)}$, as in Lemma 19.4.3.

Fix two points $x, y \in \mathbb{R}^n$, we prove that

$$\left(1-\frac{\varepsilon}{3}\right)m\|x-y\| \le \|P_{\mathcal{F}}(x)-P_{\mathcal{F}}(y)\| \le \left(1+\frac{\varepsilon}{3}\right)m\|x-y\|$$

holds with probability $\ge 1-n^{-2}$. Since there are $\binom{n}{2}$ pairs of points in X, it follows that with constant probability (say $> 1/3$) this holds for all pairs of points of X. In such a case, the mapping p is D-embedding of X into $\mathbb{R}^k$ with $D \le \frac{1+\varepsilon/3}{1-\varepsilon/3} \le 1+\varepsilon$, for $\varepsilon \le 1$.

Let $u = x - y$, we have $P_{\mathcal{F}}(u) = P_{\mathcal{F}}(x) - P_{\mathcal{F}}(y)$ since $P_{\mathcal{F}}(\cdot)$ is a linear operator. Thus, the condition becomes $\left(1-\frac{\varepsilon}{3}\right)m\|u\| \le \|P_{\mathcal{F}}(u)\| \le \left(1+\frac{\varepsilon}{3}\right)m\|u\|$. Again, since projection is a linear operator, for any $\alpha > 0$, the condition is equivalent to

$$\left(1-\tfrac{\varepsilon}{3}\right)m\|\alpha u\| \le \|P_{\mathcal{F}}(\alpha u)\| \le \left(1+\tfrac{\varepsilon}{3}\right)m\|\alpha u\|.$$

As such, we can assume that $\|u\| = 1$ by picking $\alpha = 1/\|u\|$. Namely, we need to show that

$$\left|\|P_{\mathcal{F}}(u)\| - m\right| \le \frac{\varepsilon}{3}m.$$

Let $f(u) = \|P_{\mathcal{F}}(u)\|$. By Lemma 19.4.1 (exchanging the random space with the random vector), for $t = \varepsilon m/3$, we have that the probability that this does not hold is bounded by

$$\mathbf{Pr}[|f(u)-m| \ge t] \le 4\exp\left(-\frac{t^2 n}{2}\right) = 4\exp\left(\frac{-\varepsilon^2 m^2 n}{18}\right) \le 4\exp\left(-\frac{\varepsilon^2 k}{72}\right) < n^{-2},$$

since $m \ge \frac{1}{2}\sqrt{k/n}$ and $k = 200\varepsilon^{-2}\ln n$. ∎

19.5. Bibliographical notes

Our presentation follows Matoušek [Mat02]. The Brunn-Minkowski inequality is a powerful inequality which is widely used in mathematics. A nice survey of this inequality and its applications is provided by Gardner [Gar02]. Gardner says: "In a sea of mathematics, the Brunn-Minkowski inequality appears like an octopus, tentacles reaching far and wide, its shape and color changing as it roams from one area to the next." However, Gardner is careful in claiming that the Brunn-Minkowski inequality is one of the most powerful inequalities

in mathematics since as a wit put it "the most powerful inequality is $x^2 \geq 0$, since all inequalities are in some sense equivalent to it."

A striking application of the Brunn-Minkowski inequality is the proof that in any partial ordering of n elements, there is a single comparison that knowing its result, reduces the number of linear extensions that are consistent with the partial ordering, by a constant fraction. This immediately implies (the uninteresting result) that one can sort n elements in $O(n \log n)$ comparisons. More interestingly, it implies that if there are m linear extensions of the current partial ordering, we can *always* sort it using $O(\log m)$ comparisons. A nice exposition of this surprising result is provided by Matoušek [Mat02, Section 12.3].

There are several alternative proofs of the JL lemma, see [IM98] and [DG03]. Interestingly, it is enough to pick each entry in the dimension reducing matrix randomly out of $-1, 0, 1$. This requires a more involved proof [Ach01]. This is useful when one cares about storing this dimension reduction transformation efficiently.

Magen [Mag07] observed that the JL lemma preserves angles, and in fact can be used to preserve any "k dimensional angle", by projecting down to dimension $O(k\varepsilon^{-2} \log n)$. In particular, Exercise 19.6.5 is taken from there.

In fact, the random embedding preserves much more structure than just distances between points. It preserves the structure and distances of surfaces as long as they are low dimensional and "well behaved", see [AHY07] for some results in this direction.

Dimension reduction is crucial in learning, AI, databases, etc. One common technique that is being used in practice is to do PCA (i.e., principal component analysis) and take the first few main axes. Other techniques include independent component analysis, and MDS (multidimensional scaling). MDS tries to embed points from high dimensions into low dimension ($d = 2$ or 3), while preserving some properties. Theoretically, dimension reduction into really low dimensions is hopeless, as the distortion in the worst case is $\Omega(n^{1/(k-1)})$, if k is the target dimension [Mat90].

19.6. Exercises

Exercise 19.6.1 (Boxes can be separated.). (Easy.) Let A and B be two axis-parallel boxes that are interior disjoint. Prove that there is always an axis-parallel hyperplane that separates the interior of the two boxes.

Exercise 19.6.2 (Brunn-Minkowski inequality slight extension.). Prove the following.

> **Corollary 19.6.3.** *For A and B compact sets in $\mathbb{R}^n$, we have for any $\lambda \in [0, 1]$ that* $\mathrm{vol}(\lambda A + (1-\lambda)B) \geq \mathrm{vol}(A)^{\lambda}\mathrm{vol}(B)^{1-\lambda}$.

Exercise 19.6.4 (Projections are contractions.). (Easy.) Let $\mathcal{F}$ be a k-dimensional affine subspace, and let $P_{\mathcal{F}} : \mathbb{R}^d \to \mathcal{F}$ be the projection that maps every point $x \in \mathbb{R}^d$ to its nearest neighbor on $\mathcal{F}$. Prove that p is a contraction (i.e., 1-Lipschitz). Namely, for any $\mathsf{p}, \mathsf{q} \in \mathbb{R}^d$, it holds that $\|P_{\mathcal{F}}(\mathsf{p}) - P_{\mathcal{F}}(\mathsf{q})\| \leq \|\mathsf{p} - \mathsf{q}\|$.

Exercise 19.6.5 (JL Lemma works for angles.). Show that the Johnson-Lindenstrauss lemma also $(1 \pm \varepsilon)$-preserves angles among triples of points of P (you might need to increase the target dimension however by a constant factor). *[For every angle, construct a equilateral triangle that its edges are being preserved by the projection (add the vertices of those triangles [conceptually] to the point set being embedded). Argue, that this implies that the angle is being preserved.]*

19.7. Miscellaneous

Lemma 19.7.1. *(A) The multidimensional normal distribution is symmetric; that is, for any two points* $\mathsf{p}, \mathsf{q} \in \mathbb{R}^d$ *such that* $\|\mathsf{p}\| = \|\mathsf{q}\|$ *we have that* $g(\mathsf{p}) = g(\mathsf{q})$*, where* $g(\cdot)$ *is the density function of the multidimensional normal distribution* $\mathbf{N}^d$*.*

(B) The projection of the normal distribution on any direction is a one dimensional normal distribution.

(C) Picking d variables $X_1, \ldots, X_d$ *using one dimensional normal distribution* $\mathbf{N}$ *results in a point* $(X_1, \ldots, X_d)$ *that has multidimensional normal distribution* $\mathbf{N}^d$*.*

Bibliography

[Ach01] D. Achlioptas. Database-friendly random projections. In *Proc. 20th ACM Sympos. Principles Database Syst.* (PODS), pages 274–281, 2001.

[AHY07] P. Agarwal, S. Har-Peled, and H. Yu. Embeddings of surfaces, curves, and moving points in Euclidean space. In *Proc. 23rd Annu. Sympos. Comput. Geom.* (SoCG), pages 381–389, 2007.

[BM58] G. E.P. Box and M. E. Muller. A note on the generation of random normal deviates. *Ann. Math. Stat.*, 28:610–611, 1958.

[Car76] L. Carroll. The hunting of the snark, 1876.

[DG03] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Rand. Struct. Alg.*, 22(3):60–65, 2003.

[Gar02] R. J. Gardner. The Brunn-Minkowski inequality. *Bull. Amer. Math. Soc.*, 39:355–405, 2002.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annu. ACM Sympos. Theory Comput.* (STOC), pages 604–613, 1998.

[Mag07] A. Magen. Dimensionality reductions in ℓ_2 that preserve volumes and distance to affine spaces. *Discrete Comput. Geom.*, 38(1):139–153, 2007.

[Mat90] J. Matoušek. Bi-Lipschitz embeddings into low-dimensional Euclidean spaces. *Comment. Math. Univ. Carolinae*, 31:589–600, 1990.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Grad. Text in Math.* Springer, 2002.

Chapter 20

On Complexity, Sampling, and ε-Nets and ε-Samples

By Sariel Har-Peled, December 30, 2015[①]

> "I've never touched the hard stuff, only smoked grass a few times with the boys to be polite, and that's all, though ten is the age when the big guys come around teaching you all sorts to things. But happiness doesn't mean much to me, I still think life is better. Happiness is a mean son of a bitch and needs to be put in his place. Him and me aren't on the same team, and I'm cutting him dead. I've never gone in for politics, because somebody always stand to gain by it, but happiness is an even crummier racket, and their ought to be laws to put it out of business."
>
> – Momo, Emile Ajar.

In this chapter we will try to quantify the notion of geometric complexity. It is intuitively clear that a ● (i.e., disk) is a simpler shape than an ⬬ (i.e., ellipse), which is in turn simpler than a ☻ (i.e., smiley). This becomes even more important when we consider several such shapes and how they interact with each other. As these examples might demonstrate, this notion of complexity is somewhat elusive.

To this end, we show that one can capture the structure of a distribution/point set by a small subset. The size here would depend on the complexity of the shapes/ranges we care about, but surprisingly it would be independent of the size of the point set.

20.1. VC dimension

Definition 20.1.1. A *range space* S is a pair $(\mathsf{X}, \mathcal{R})$, where X is a *ground set* (finite or infinite) and $\mathcal{R}$ is a (finite or infinite) family of subsets of X. The elements of X are ***points*** and the elements of $\mathcal{R}$ are ***ranges***.

Our interest is in the size/weight of the ranges in the range space. For technical reasons, it will be easier to consider a finite subset x as the underlining ground set.

Definition 20.1.2. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space, and let x be a finite (fixed) subset of X. For a range $\mathbf{r} \in \mathcal{R}$, its *measure* is the quantity

$$\overline{m}(\mathbf{r}) = \frac{|\mathbf{r} \cap \mathsf{x}|}{|\mathsf{x}|}.$$

While x is finite, it might be very large. As such, we are interested in getting a good estimate to $\overline{m}(\mathbf{r})$ by using a more compact set to represent the range space.

Definition 20.1.3. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space. For a subset N (which might be a multi-set) of x, its *estimate* of the measure of $\overline{m}(\mathbf{r})$, for $\mathbf{r} \in \mathcal{R}$, is the quantity

$$\overline{s}(\mathbf{r}) = \frac{|\mathbf{r} \cap N|}{|N|}.$$

The main purpose of this chapter is to come up with methods to generate a sample N, such that $\overline{m}(\mathbf{r}) \approx \overline{s}(\mathbf{r})$, for all the ranges $\mathbf{r} \in \mathcal{R}$.

It is easy to see that in the worst case, no sample can capture the measure of all ranges. Indeed, given a sample N, consider the range $\mathsf{x} \setminus N$ that is being completely missed by N. As such, we need to concentrate on range spaces that are "low dimensional", where not all subsets are allowable ranges. The notion of VC dimension (named after Vapnik and Chervonenkis [VC71]) is one way to limit the complexity of a range space.

Definition 20.1.4. Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space. For $Y \subseteq \mathsf{X}$, let

$$\mathcal{R}_{|Y} = \left\{ \mathbf{r} \cap Y \;\middle|\; \mathbf{r} \in \mathcal{R} \right\} \tag{20.1}$$

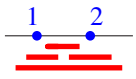
denote the ***projection*** of $\mathcal{R}$ on Y. The range space S projected to Y is $\mathsf{S}_{|Y} = (Y, \mathcal{R}_{|Y})$.

If $\mathcal{R}_{|Y}$ contains all subsets of Y (i.e., if Y is finite, we have $\left|\mathcal{R}_{|Y}\right| = 2^{|Y|}$), then Y is *shattered* by $\mathcal{R}$ (or equivalently Y is shattered by S).

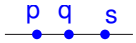
The ***Vapnik-Chervonenkis*** dimension (or VC ***dimension***) of S, denoted by $\dim_{\mathsf{VC}}(S)$, is the maximum cardinality of a shattered subset of X. If there are arbitrarily large shattered subsets, then $\dim_{\mathsf{VC}}(S) = \infty$.

20.1.1. Examples

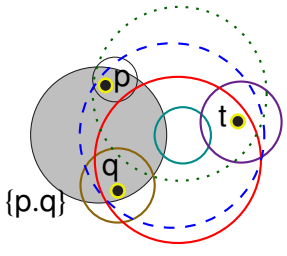
Intervals. Consider the set X to be the real line, and consider $\mathcal{R}$ to be the set of all intervals on the real line. Consider the set $Y = \{1, 2\}$. Clearly, one can find four intervals that contain all possible subsets of Y. Formally, the projection $\mathcal{R}_{|Y} = \{\{\}, \{1\}, \{2\}, \{1, 2\}\}$. The intervals realizing each of these subsets are depicted on the right.



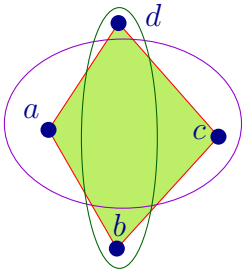
However, this is false for a set of three points $B = \{\mathsf{p}, \mathsf{q}, \mathsf{r}\}$, since there is no interval that can contain the two extreme points p and r without also containing q. Namely, the subset $\{\mathsf{p}, \mathsf{r}\}$ is not realizable for intervals, implying that the largest shattered set by the range space (real line, intervals) is of size two. We conclude that the VC dimension of this space is two.



Disks. Let $\mathsf{X} = \mathbb{R}^2$, and let $\mathcal{R}$ be the set of disks in the plane. Clearly, for any three points in the plane (in general position), denoted by p, q, and r, one can find eight disks that realize all possible 2^3 different subsets. See the figure on the right.

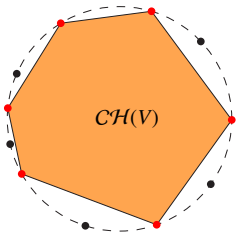


But can disks shatter a set with four points? Consider such a set P of four points. If the convex hull of P has only three points on its boundary, then the subset X having only those three vertices (i.e., it does not include the middle point) is impossible, by convexity. Namely, there is no disk that contains only the points of X without the middle point.



Alternatively, if all four points are vertices of the convex hull and they are a, b, c, d along the boundary of the convex hull, either the set $\{a, c\}$ or the set $\{b, d\}$ is not realizable. Indeed, if both options are realizable, then consider the two disks D_1 and D_2 that realize those assignments. Clearly, ∂D_1 and ∂D_2 must intersect in four points, but this is not possible, since two circles have at most two intersection points. See the figure on the left. Hence the VC dimension of this range space is 3.

Convex sets. Consider the range space $\mathsf{S} = (\mathbb{R}^2, \mathcal{R})$, where $\mathcal{R}$ is the set of all (closed) convex sets in the plane. We claim that $\dim_{\mathsf{VC}}(\mathsf{S}) = \infty$. Indeed, consider a set U of n points $p_1, \ldots, p_n$ all lying on the boundary of the unit circle in the plane. Let V be any subset of U, and consider the convex hull $\mathcal{CH}(V)$. Clearly, $\mathcal{CH}(V) \in \mathcal{R}$, and furthermore, $\mathcal{CH}(V) \cap U = V$. Namely, any subset of U is realizable by S. Thus, S can shatter sets of arbitrary size, and its VC dimension is unbounded.



Complement. Consider the range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ with $\delta = \dim_{\mathsf{VC}}(\mathsf{S})$. Next, consider the complement space, $\overline{\mathsf{S}} = \left(\mathsf{X}, \overline{\mathcal{R}}\right)$, where

$$\overline{\mathcal{R}} = \left\{ \mathsf{X} \setminus \mathbf{r} \;\middle|\; \mathbf{r} \in \mathcal{R} \right\};$$

namely, the ranges of $\overline{\mathsf{S}}$ are the complement of the ranges in S. What is the VC dimension of $\overline{\mathsf{S}}$? Well, a set $B \subseteq \mathsf{X}$ is shattered by $\overline{\mathsf{S}}$ if and only if it is shattered by S. Indeed, if S shatters B, then for any $Z \subseteq B$, we have that $(B \setminus Z) \in \mathcal{R}_{|B}$, which implies that $Z = B \setminus (B \setminus Z) \in \overline{\mathcal{R}}_{|B}$. Namely, $\overline{\mathcal{R}}_{|B}$ contains all the subsets of B, and $\overline{\mathsf{S}}$ shatters B. Thus, $\dim_{\mathsf{VC}}\left(\overline{\mathsf{S}}\right) = \dim_{\mathsf{VC}}(\mathsf{S})$.

Lemma 20.1.5. *For a range space $\mathsf{S} = (X, \mathcal{R})$ we have that $\dim_{\mathsf{VC}}(\mathsf{S}) = \dim_{\mathsf{VC}}\left(\overline{\mathsf{S}}\right)$, where $\overline{\mathsf{S}}$ is the complement range space.*

20.1.1.1. Halfspaces

Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, where $\mathsf{X} = \mathbb{R}^d$ and $\mathcal{R}$ is the set of all (closed) halfspaces in $\mathbb{R}^d$. We need the following technical claim.

Claim 20.1.6. *Let $\mathsf{P} = \{\mathsf{p}_1, \ldots, \mathsf{p}_{d+2}\}$ be a set of $d+2$ points in $\mathbb{R}^d$. There are real numbers $\beta_1, \ldots, \beta_{d+2}$, not all of them zero, such that $\sum_i \beta_i \mathsf{p}_i = 0$ and $\sum_i \beta_i = 0$.*

Proof: Indeed, set $\mathsf{q}_i = (\mathsf{p}_i, 1)$, for $i = 1, \ldots, d+2$. Now, the points $\mathsf{q}_1, \ldots, \mathsf{q}_{d+2} \in \mathbb{R}^{d+1}$ are linearly dependent, and there are coefficients $\beta_1, \ldots, \beta_{d+2}$, not all of them zero, such that $\sum_{i=1}^{d+2} \beta_i \mathsf{q}_i = 0$. Considering only the first d coordinates of these points implies that $\sum_{i=1}^{d+2} \beta_i \mathsf{p}_i = 0$. Similarly, by considering only the $(d+1)$st coordinate of these points, we have that $\sum_{i=1}^{d+2} \beta_i = 0$. ■

To see what the VC dimension of halfspaces in $\mathbb{R}^d$ is, we need the following result of Radon. (For a reminder of the formal definition of convex hulls, see Definition 20.9.1p25.)

Theorem 20.1.7 (Radon's theorem). *Let $\mathsf{P} = \{\mathsf{p}_1, \ldots, \mathsf{p}_{d+2}\}$ be a set of $d+2$ points in $\mathbb{R}^d$. Then, there exist two disjoint subsets C and D of P, such that $\mathcal{CH}(C) \cap \mathcal{CH}(D) \neq \emptyset$ and $C \cup D = \mathsf{P}$.*

Proof: By Claim 20.1.6 there are real numbers $\beta_1, \ldots, \beta_{d+2}$, not all of them zero, such that $\sum_i \beta_i \mathsf{p}_i = 0$ and $\sum_i \beta_i = 0$.

Assume, for the sake of simplicity of exposition, that $\beta_1, \ldots, \beta_k \geq 0$ and $\beta_{k+1}, \ldots, \beta_{d+2} < 0$. Furthermore, let $\mu = \sum_{i=1}^{k} \beta_i = -\sum_{i=k+1}^{d+2} \beta_i$. We have that

$$\sum_{i=1}^{k} \beta_i \mathsf{p}_i = -\sum_{i=k+1}^{d+2} \beta_i \mathsf{p}_i.$$

In particular, $v = \sum_{i=1}^{k} (\beta_i/\mu)\mathsf{p}_i$ is a point in $\mathcal{CH}(\{\mathsf{p}_1, \ldots, \mathsf{p}_k\})$. Furthermore, for the same point v we have $v = \sum_{i=k+1}^{d+2} -(\beta_i/\mu)\mathsf{p}_i \in \mathcal{CH}(\{\mathsf{p}_{k+1}, \ldots, \mathsf{p}_{d+2}\})$. We conclude that v is in the intersection of the two convex hulls, as required. ■

The following is a trivial observation, and yet we provide a proof to demonstrate it is true.

Lemma 20.1.8. *Let $\mathsf{P} \subseteq \mathbb{R}^d$ be a finite set, let r be any point in $\mathcal{CH}(\mathsf{P})$, and let h^+ be a halfspace of $\mathbb{R}^d$ containing r. Then there exists a point of P contained inside h^+.*

Proof: The halfspace h^+ can be written as $h^+ = \left\{\mathsf{t} \in \mathbb{R}^d \;\middle|\; \langle \mathsf{t}, v\rangle \leq c\right\}$. Now $\mathsf{r} \in \mathcal{CH}(\mathsf{P}) \cap h^+$, and as such there are numbers $\alpha_1, \ldots, \alpha_m \geq 0$ and points $\mathsf{p}_1, \ldots, \mathsf{p}_m \in \mathsf{P}$, such that $\sum_i \alpha_i = 1$ and $\sum_i \alpha_i \mathsf{p}_i = \mathsf{r}$. By the linearity of the dot product, we have that

$$\mathsf{r} \in h^+ \implies \langle \mathsf{r}, v\rangle \leq c \implies \left\langle \sum_{i=1}^{m} \alpha_i \mathsf{p}_i, v \right\rangle \leq c \implies \beta = \sum_{i=1}^{m} \alpha_i \langle \mathsf{p}_i, v\rangle \leq c.$$

Setting $\beta_i = \langle \mathsf{p}_i, v\rangle$, for $i = 1, \ldots, m$, the above implies that β is a weighted average of $\beta_1, \ldots, \beta_m$. In particular, there must be a β_i that is no larger than the average. That is $\beta_i \leq c$. This implies that $\langle \mathsf{p}_i, v\rangle \leq c$. Namely, $\mathsf{p}_i \in h^+$ as claimed. ■

Let S be the range space having $\mathbb{R}^d$ as the ground set and all the close halfspaces as ranges. Radon's theorem implies that if a set Q of $d+2$ points is being shattered by S, then we can partition this set Q into two disjoint sets Y and Z such that $\mathcal{CH}(Y) \cap \mathcal{CH}(Z) \neq \emptyset$. In particular, let r be a point in $\mathcal{CH}(Y) \cap \mathcal{CH}(Z)$. If a halfspace h^+ contains all the points of Y, then $\mathcal{CH}(Y) \subseteq h^+$, since a halfspace is a convex set. Thus, any halfspace h^+ containing all the points of Y will contain the point $\mathsf{r} \in \mathcal{CH}(Y)$. But $\mathsf{r} \in \mathcal{CH}(Z) \cap h^+$, and this implies that a point of Z must lie in h^+, by Lemma 20.1.8. Namely, the subset $Y \subseteq \mathsf{Q}$ cannot be realized by a halfspace, which implies that Q cannot be shattered. Thus $\dim_{\mathsf{VC}}(S) < d+2$. It is also easy to verify that the regular simplex with $d+1$ vertices is shattered by S. Thus, $\dim_{\mathsf{VC}}(S) = d+1$.

20.2. Shattering dimension and the dual shattering dimension

The main property of a range space with bounded VC dimension is that the number of ranges for a set of n elements grows polynomially in n (with the power being the dimension) instead of exponentially. Formally, let the *growth function* be

$$\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i} \leq \sum_{i=0}^{\delta} \frac{n^i}{i!} \leq n^\delta, \tag{20.2}$$

for $\delta > 1$ (the cases where $\delta = 0$ or $\delta = 1$ are not interesting and we will just ignore them). Note that for all $n, \delta \geq 1$, we have $\mathcal{G}_\delta(n) = \mathcal{G}_\delta(n-1) + \mathcal{G}_{\delta-1}(n-1)$[2].

[2] Here is a cute (and standard) counting argument: $\mathcal{G}_\delta(n)$ is just the number of different subsets of size at most δ out of n elements. Now, we either decide to not include the first element in these subsets (i.e., $\mathcal{G}_\delta(n-1)$) or, alternatively, we include the first element in these subsets, but then there are only $\delta - 1$ elements left to pick (i.e., $\mathcal{G}_{\delta-1}(n-1)$).

Lemma 20.2.1 (Sauer's lemma). *If $(X, \mathcal{R})$ is a range space of* VC *dimension δ with $|X| = n$, then $|\mathcal{R}| \leq \mathcal{G}_\delta(n)$.*

Proof: The claim trivially holds for $\delta = 0$ or $n = 0$.

Let x be any element of X, and consider the sets

$$\mathcal{R}_x = \left\{ \mathbf{r} \setminus \{x\} \;\middle|\; \mathbf{r} \cup \{x\} \in \mathcal{R} \text{ and } \mathbf{r} \setminus \{x\} \in \mathcal{R} \right\} \quad \text{and} \quad \mathcal{R} \setminus x = \left\{ \mathbf{r} \setminus \{x\} \;\middle|\; \mathbf{r} \in \mathcal{R} \right\}.$$

Observe that $|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus x|$. Indeed, we charge the elements of $\mathcal{R}$ to their corresponding element in $\mathcal{R} \setminus x$. The only bad case is when there is a range $\mathbf{r}$ such that both $\mathbf{r} \cup \{x\} \in \mathcal{R}$ and $\mathbf{r} \setminus \{x\} \in \mathcal{R}$, because then these two distinct ranges get mapped to the same range in $\mathcal{R} \setminus x$. But such ranges contribute exactly one element to $\mathcal{R}_x$. Similarly, every element of $\mathcal{R}_x$ corresponds to two such "twin" ranges in $\mathcal{R}$.

Observe that $(X \setminus \{x\}, \mathcal{R}_x)$ has VC dimension $\delta - 1$, as the largest set that can be shattered is of size $\delta - 1$. Indeed, any set $B \subset X \setminus \{x\}$ shattered by $\mathcal{R}_x$ implies that $B \cup \{x\}$ is shattered in $\mathcal{R}$.

Thus, we have

$$|\mathcal{R}| = |\mathcal{R}_x| + |\mathcal{R} \setminus x| \leq \mathcal{G}_{\delta-1}(n-1) + \mathcal{G}_\delta(n-1) = \mathcal{G}_\delta(n),$$

by induction. ■

Interestingly, Lemma 20.2.1 is tight. See Exercise 20.8.4.

Next, we show pretty tight bounds on $\mathcal{G}_\delta(n)$. The proof is technical and not very interesting, and it is delegated to Section 20.6.

Lemma 20.2.2. *For $n \geq 2\delta$ and $\delta \geq 1$, we have* $\left(\frac{n}{\delta}\right)^\delta \leq \mathcal{G}_\delta(n) \leq 2\left(\frac{ne}{\delta}\right)^\delta$, *where* $\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i}$.

Definition 20.2.3 (Shatter function). Given a range space $S = (X, \mathcal{R})$, its *shatter function* $\pi_S(m)$ is the maximum number of sets that might be created by S when restricted to subsets of size m. Formally,

$$\pi_S(m) = \max_{\substack{B \subset X \\ |B| = m}} \left| \mathcal{R}_{|B} \right|;$$

see Eq. (20.1).

The *shattering dimension* of S is the smallest d such that $\pi_S(m) = O(m^d)$, for all m.

By applying Lemma 20.2.1 to a finite subset of X, we get:

Corollary 20.2.4. *If $S = (X, \mathcal{R})$ is a range space of* VC *dimension δ, then for every finite subset B of X, we have $\left|\mathcal{R}_{|B}\right| \leq \pi_S(|B|) \leq \mathcal{G}_\delta(|B|)$. That is, the* VC *dimension of a range space always bounds its shattering dimension.*

Proof: Let $n = |B|$, and observe that $\left|\mathcal{R}_{|B}\right| \leq \mathcal{G}_\delta(n) \leq n^\delta$, by Eq. (20.2). As such, $\left|\mathcal{R}_{|B}\right| \leq n^\delta$, and, by definition, the shattering dimension of S is at most δ; namely, the shattering dimension is bounded by the VC dimension. ■

Our arch-nemesis in the following is the function $x/\ln x$. The following lemma states some properties of this function, and its proof is delegated to Exercise 20.8.2.

Lemma 20.2.5. *For the function $f(x) = x/\ln x$ the following hold.*

(A) $f(x)$ is monotonically increasing for $x \geq e$.

(B) $f(x) \geq e$, for $x > 1$.

(C) For $u \geq \sqrt{e}$, if $f(x) \leq u$, then $x \leq 2u \ln u$.

(D) For $u \geq \sqrt{e}$, if $x > 2u \ln u$, then $f(x) > u$.
(E) For $u \geq e$, if $f(x) \geq u$, then $x \geq u \ln u$.

The next lemma introduces a standard argument which is useful in bounding the VC dimension of a range space by its shattering dimension. It is easy to see that the bound is tight in the worst case.

Lemma 20.2.6. *If $\mathsf{S} = (X, \mathcal{R})$ is a range space with shattering dimension d, then its* VC *dimension is bounded by $O(d \log d)$.*

Proof: Let $N \subseteq \mathsf{X}$ be the largest set shattered by S, and let δ denote its cardinality. We have that $2^\delta = \left|\mathcal{R}_{|N}\right| \leq \pi_{\mathsf{S}}(|N|) \leq c\delta^d$, where c is a fixed constant. As such, we have that $\delta \leq \lg c + d \lg \delta$, which in turn implies that $\frac{\delta - \lg c}{\lg \delta} \leq d$.[3] Assuming $\delta \geq \max(2, 2 \lg c)$, we have that

$$\frac{\delta}{2 \lg \delta} \leq d \implies \frac{\delta}{\ln \delta} \leq \frac{2d}{\ln 2} \leq 6d \implies \delta \leq 2(6d) \ln(6d),$$

by Lemma 20.2.5(C). ■

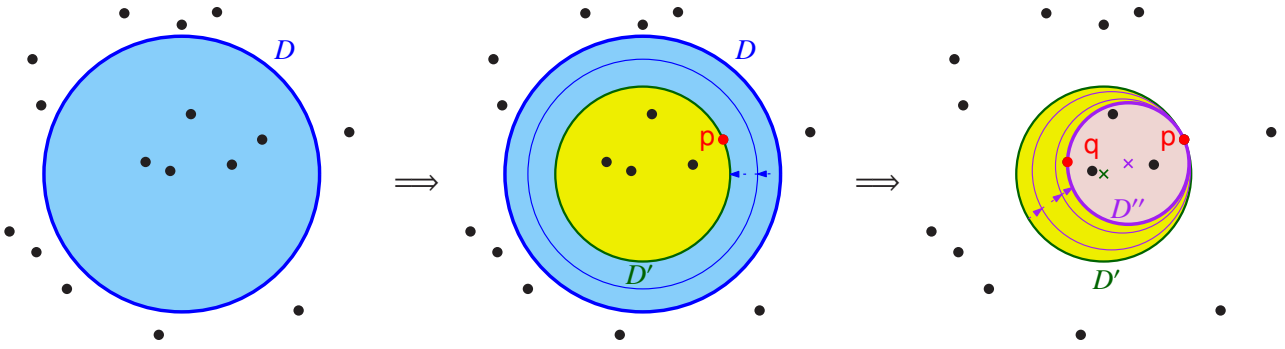
Disks revisited. To see why the shattering dimension is more convenient to work with than the VC dimension, consider the range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, where $\mathsf{X} = \mathbb{R}^2$ and $\mathcal{R}$ is the set of disks in the plane. We know that the VC dimension of S is 3 (see Section 20.1.1).

We next use a standard continuous deformation argument to argue that the shattering dimension of this range space is also 3.

Lemma 20.2.7. *Consider the range space $\mathsf{S} = (X, \mathcal{R})$, where $X = \mathbb{R}^2$ and $\mathcal{R}$ is the set of disks in the plane. The shattering dimension of S is 3.*

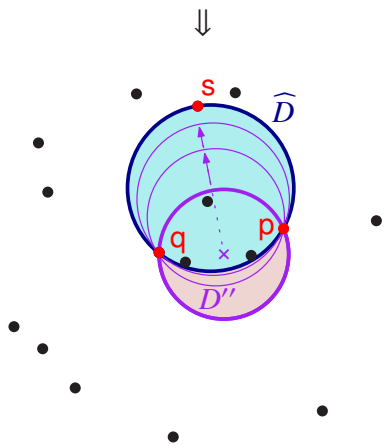
Proof: Consider any set P of n points in the plane, and consider the set $\mathcal{F} = \mathcal{R}_{|\mathsf{P}}$. We claim that $|\mathcal{F}| \leq 4n^3$.

The set $\mathcal{F}$ contains only n sets with a single point in them and only $\binom{n}{2}$ sets with two points in them. So, fix $\mathsf{Q} \in \mathcal{F}$ such that $|\mathsf{Q}| \geq 3$.



[3] We remind the reader that $\lg = \log_2$.

There is a disk D that realizes this subset; that is, $\mathsf{P} \cap D = \mathsf{Q}$. For the sake of simplicity of exposition, assume that P is in general position. Shrink D till its boundary passes through a point p of P.



Now, continue shrinking the new disk D' in such a way that its boundary passes through the point p (this can be done by moving the center of D' towards p). Continue in this continuous deformation till the new boundary hits another point q of P. Let D'' denote this disk.

Next, we continuously deform D'' so that it has both $\mathsf{p} \in \mathsf{Q}$ and $\mathsf{q} \in \mathsf{Q}$ on its boundary. This can be done by moving the center of D'' along the bisector linear between p and q. Stop as soon as the boundary of the disk hits a third point $\mathsf{r} \in \mathsf{P}$. (We have freedom in choosing in which direction to move the center. As such, move in the direction that causes the disk boundary to hit a new point r.) Let $\widehat{D}$ be the resulting disk. The boundary of $\widehat{D}$ is the unique circle passing through p, q, and r. Furthermore, observe that

$$D \cap (\mathsf{P} \setminus \{\mathsf{r}\}) = \widehat{D} \cap (\mathsf{P} \setminus \{\mathsf{r}\}).$$

That is, we can specify the point set $\mathsf{P} \cap D$ by specifying the three points $\mathsf{p}, \mathsf{q}, \mathsf{r}$ (and thus specifying the disk $\widehat{D}$) and the status of the three special points; that is, we specify for each point $\mathsf{p}, \mathsf{q}, \mathsf{r}$ whether or not it is inside the generated subset.

As such, there are at most $8\binom{n}{3}$ different subsets in $\mathcal{F}$ containing more than three points, as each such subset maps to a "canonical" disk, there are at most $\binom{n}{3}$ different such disks, and each such disk defines at most eight different subsets.

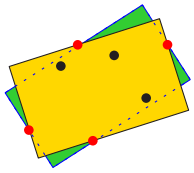
Similar argumentation implies that there are at most $4\binom{n}{2}$ subsets that are defined by a pair of points that realizes the diameter of the resulting disk. Overall, we have that

$$|\mathcal{F}| = 1 + n + 4\binom{n}{2} + 8\binom{n}{3} \leq 4n^3,$$

since there is one empty set in $\mathcal{F}$, n sets of size 1, and the rest of the sets are counted as described above. ■

The proof of Lemma 20.2.7 might not seem like a great simplification over the same bound we got by arguing about the VC dimension. However, the above argumentation gives us a very powerful tool – the shattering dimension of a range space defined by a family of shapes is always bounded by the number of points that determine a shape in the family.

Thus, the shattering dimension of, say, arbitrarily oriented rectangles in the plane is bounded by (and in this case, equal to) five, since such a rectangle is uniquely determined by five points. To see that, observe that if a rectangle has only four points on its boundary, then there is one degree of freedom left, since we can rotate the rectangle "around" these points; see the figure on the right.



20.2.1. The dual shattering dimension

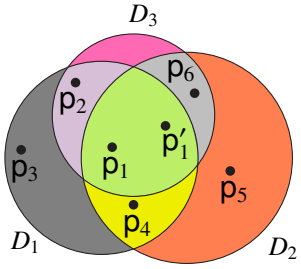
Given a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$, consider a point $\mathsf{p} \in \mathsf{X}$. There is a set of ranges of $\mathcal{R}$ associated with p, namely, the set of all ranges of $\mathcal{R}$ that contains p which we denote by

$$\mathcal{R}_{\mathsf{p}} = \left\{ \mathbf{r} \;\middle|\; \mathbf{r} \in \mathcal{R},\ \text{the range } \mathbf{r} \text{ contains } \mathsf{p} \right\}.$$

This gives rise to a natural dual range space to S.

Definition 20.2.8. The ***dual range space*** to a range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is the space $\mathsf{S}^\star = (\mathcal{R}, \mathsf{X}^\star)$, where $\mathsf{X}^\star = \left\{ \mathcal{R}_\mathsf{p} \;\middle|\; \mathsf{p} \in \mathsf{X} \right\}$.

(A)



(B)

	p_1	p_1'	p_2	p_3	p_4	p_5	p_6
D_1	1	1	1	1	1	0	0
D_2	1	1	0	0	1	1	1
D_3	1	1	1	0	0	0	1

(C)

	D_1	D_2	D_3
p_1	1	1	1
p_1'	1	1	1
p_2	1	0	1
p_3	1	0	0
p_4	1	1	0
p_5	0	1	0
p_6	0	1	1

Figure 20.1: (A) $\mathcal{R}_{\mathsf{p}_1} = \mathcal{R}_{\mathsf{p}_1'}$. (B) Writing the set system as an incidence matrix where a point is a column and a set is a row. For example, D_2 contains p_4, and as such the column of p_4 has a 1 in the row corresponding to D_2. (C) The dual set system is represented by a matrix which is the transpose of the original incidence matrix.

Naturally, the dual range space to $\mathsf{S}^\star$ is the original S, which is thus sometimes referred to as the ***primal range space***. (In other words, the dual to the dual is the primal.) The easiest way to see this, is to think about it as an abstract set system realized as an incidence matrix, where each point is a column and a set is a row in the matrix having 1 in an entry if and only if it contains the corresponding point; see Figure 20.1. Now, it is easy to verify that the dual range space is the transposed matrix.

To understand what the dual space is, consider X to be the plane and $\mathcal{R}$ to be a set of m disks. Then, in the dual range space $\mathsf{S}^\star = (\mathcal{R}, \mathsf{X}^\star)$, every point p in the plane has a set associated with it in $\mathsf{X}^\star$, which is the set of disks of $\mathcal{R}$ that contains p. In particular, if we consider the arrangement formed by the m disks of $\mathcal{R}$, then all the points lying inside a single face of this arrangement correspond to the same set of $\mathsf{X}^\star$. The number of ranges in $\mathsf{X}^\star$ is bounded by the complexity of the arrangement of these disks, which is $O(m^2)$; see Figure 20.1.

Let the ***dual shatter function*** of the range space S be $\pi^\star_\mathsf{S}(m) = \pi_{\mathsf{S}^\star}(m)$, where $\mathsf{S}^\star$ is the dual range space to S.

Definition 20.2.9. The ***dual shattering dimension*** of S is the shattering dimension of the dual range space $\mathsf{S}^\star$.

Note that the dual shattering dimension might be smaller than the shattering dimension and hence also smaller than the VC dimension of the range space. Indeed, in the case of disks in the plane, the dual shattering dimension is just 2, while the VC dimension and the shattering dimension of this range space is 3. Note, also, that in geometric settings bounding the dual shattering dimension is relatively easy, as all you have to do is bound the complexity of the arrangement of m ranges of this space.

The following lemma shows a connection between the VC dimension of a space and its dual. The interested reader[④] might find the proof amusing.

Lemma 20.2.10. *Consider a range space $\mathsf{S} = (X, \mathcal{R})$ with* VC *dimension δ. The dual range space $\mathsf{S}^\star = (\mathcal{R}, X^\star)$ has* VC *dimension bounded by $2^{\delta+1}$.*

[④]The author is quite aware that the interest of the reader in this issue might not be the result of free choice. Nevertheless, one might draw some comfort from the realization that the existence of the interested reader is as much an illusion as the existence of free choice. Both are convenient to assume, and both are probably false. Or maybe not.

Proof: Assume that $\mathsf{S}^\star$ shatters a set $\mathcal{F} = \{\mathbf{r}_1, \ldots, \mathbf{r}_k\} \subseteq \mathcal{R}$ of k ranges. Then, there is a set $\mathsf{P} \subseteq \mathsf{X}$ of $m = 2^k$ points that shatters $\mathcal{F}$. Formally, for every subset $V \subseteq \mathcal{F}$, there exists a point $\mathsf{p} \in \mathsf{P}$, such that $\mathcal{F}_{\mathsf{p}} = V$.

So, consider the matrix M (of dimensions $k \times 2^k$) having the points $\mathsf{p}_1, \ldots, \mathsf{p}_{2^k}$ of P as the columns, and every row is a set of $\mathcal{F}$, where the entry in the matrix corresponding to a point $\mathsf{p} \in \mathsf{P}$ and a range $\mathbf{r} \in \mathcal{F}$ is 1 if and only if $\mathsf{p} \in \mathbf{r}$ and zero otherwise. Since P shatters $\mathcal{F}$, we know that this matrix has all possible 2^k binary vectors as columns.

Next, let $\kappa' = 2^{\lfloor \lg k \rfloor} \leq k$, and consider the matrix M' of size $\kappa' \times \lg \kappa'$, where the ith row is the binary representation of the number $i-1$ (formally, the jth entry in the ith row is 1 if the jth bit in the binary representation of $i-1$ is 1), where $i = 1, \ldots, \kappa'$. See the figure on the right. Clearly, the $\lg \kappa'$ columns of M' are all different, and we can find $\lg \kappa'$ columns of M that are identical to the columns of M' (in the first κ' entries starting from the top of the columns).

$\mathbf{M}$:

	p_1	p_2	$\ldots$	p_{2^k}
$\mathbf{r}_1$	0	1		0
$\mathbf{r}_2$	1	1		1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{r}_{k-2}$	1	1	$\ldots$	0
$\mathbf{r}_{k-1}$	0	0	$\ldots$	1
$\mathbf{r}_k$	1	0	$\ldots$	1

$\mathbf{M}'$:

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Each such column corresponds to a point $\mathsf{p} \in \mathsf{P}$, and let $\mathsf{Q} \subset \mathsf{P}$ be this set of $\lg \kappa'$ points. Note that for any subset $Z \subseteq \mathsf{Q}$, there is a row t in M' that encodes this subset. Consider the corresponding row in M; that is, the range $\mathbf{r}_t \in \mathcal{F}$. Since M and M' are identical (in the relevant $\lg \kappa'$ columns of M) on the first κ', we have that $\mathbf{r}_t \cap \mathsf{Q} = Z$. Namely, the set of ranges $\mathcal{F}$ shatters Q. But since the original range space has VC dimension δ, it follows that $|\mathsf{Q}| \leq \delta$. Namely, $|\mathsf{Q}| = \lg \kappa' = \lfloor \lg k \rfloor \leq \delta$, which implies that $\lg k \leq \delta + 1$, which in turn implies that $k \leq 2^{\delta+1}$. ■

Lemma 20.2.11. *If a range space $\mathsf{S} = (X, \mathcal{R})$ has dual shattering dimension δ, then its VC dimension is bounded by $\delta^{O(\delta)}$.*

Proof: The shattering dimension of the dual range space $\mathsf{S}^\star$ is bounded by δ, and as such, by Lemma 20.2.6, its VC dimension is bounded by $\delta' = O(\delta \log \delta)$. Since the dual range space to $\mathsf{S}^\star$ is S, we have by Lemma 20.2.10 that the VC dimension of S is bounded by $2^{\delta'+1} = \delta^{O(\delta)}$. ■

The bound of Lemma 20.2.11 might not be pretty, but it is sufficient in a lot of cases to bound the VC dimension when the shapes involved are simple.

Example 20.2.12. Consider the range space $\mathsf{S} = \left(\mathbb{R}^2, \mathcal{R}\right)$, where $\mathcal{R}$ is a set of shapes in the plane, so that the boundary of any pair of them intersects at most s times. Then, the VC dimension of S is $O(1)$. Indeed, the dual shattering dimension of S is $O(1)$, since the complexity of the arrangement of n such shapes is $O(sn^2)$. As such, by Lemma 20.2.11, the VC dimension of S is $O(1)$.

20.2.1.1. Mixing range spaces

Lemma 20.2.13. *Let $\mathsf{S} = (X, \mathcal{R})$ and $\mathsf{T} = (X, \mathcal{R}')$ be two range spaces of VC dimension δ and δ', respectively, where $\delta, \delta' > 1$. Let $\widehat{\mathcal{R}} = \left\{ \mathbf{r} \cup \mathbf{r}' \;\middle|\; \mathbf{r} \in \mathcal{R}, \mathbf{r}' \in \mathcal{R}' \right\}$. Then, for the range space $\widehat{\mathsf{S}} = \left(X, \widehat{\mathcal{R}}\right)$, we have that $\dim_{\mathsf{VC}}\left(\widehat{\mathsf{S}}\right) = O(\delta + \delta')$.*

Proof: As a warm-up exercise, we prove a somewhat weaker bound here of $O((\delta+\delta')\log(\delta+\delta'))$. The stronger bound follows from Theorem 20.2.14 below. Let B be a set of n points in X that are shattered by $\widehat{\mathsf{S}}$. There are at most $\mathcal{G}_\delta(n)$ and $\mathcal{G}_{\delta'}(n)$ different ranges of B in the range sets $\mathcal{R}_{|B}$ and $\mathcal{R}'_{|B}$, respectively, by Lemma 20.2.1. Every subset C of B realized by $\widehat{r} \in \widehat{\mathcal{R}}$ is a union of two subsets $B \cap \mathbf{r}$ and $B \cap \mathbf{r}'$, where $\mathbf{r} \in R$ and $\mathbf{r}' \in \mathcal{R}'$, respectively. Thus, the number of different subsets of B realized by $\widehat{\mathsf{S}}$ is bounded by $\mathcal{G}_\delta(n)\mathcal{G}_{\delta'}(n)$. Thus, $2^n \leq n^\delta n^{\delta'}$, for $\delta, \delta' > 1$. We conclude that $n \leq (\delta + \delta') \lg n$, which implies that $n = O((\delta + \delta')\log(\delta + \delta'))$, by Lemma 20.2.5(C). ■

Interestingly, one can prove a considerably more general result with tighter bounds. The required computations are somewhat more painful.

Theorem 20.2.14. *Let* $\mathsf{S}_1 = \left(X, \mathcal{R}^1\right), \ldots, \mathsf{S}_k = \left(X, \mathcal{R}^k\right)$ *be range spaces with* VC *dimension* $\delta_1, \ldots, \delta_k$*, respectively. Next, let* $f(\mathbf{r}_1, \ldots, \mathbf{r}_k)$ *be a function that maps any* k*-tuple of sets* $\mathbf{r}_1 \in \mathcal{R}^1, \ldots, \mathbf{r}_k \in \mathcal{R}^k$ *into a subset of* X*. Consider the range set*

$$\mathcal{R}' = \left\{ f(\mathbf{r}_1, \ldots, \mathbf{r}_k) \;\middle|\; \mathbf{r}_1 \in \mathcal{R}_1, \ldots, \mathbf{r}_k \in \mathcal{R}_k \right\}$$

and the associated range space $\mathsf{T} = (X, \mathcal{R}')$*. Then, the* VC *dimension of* T *is bounded by* $O(k\delta \lg k)$*, where* $\delta = \max_i \delta_i$*.*

Proof: Assume a set $Y \subseteq \mathsf{X}$ of size t is being shattered by $\mathcal{R}'$, and observe that

$$\begin{aligned}\left|\mathcal{R}'_{|Y}\right| &\leq \left|\left\{(\mathbf{r}_1, \ldots, \mathbf{r}_k) \;\middle|\; \mathbf{r}_1 \in \mathcal{R}^1_{|Y}, \ldots, \mathbf{r}_k \in \mathcal{R}^k_{|Y}\right\}\right| \leq \left|\mathcal{R}^1_{|Y}\right| \cdots \left|\mathcal{R}^k_{|Y}\right| \leq \mathcal{G}_{\delta_1}(t) \cdot \mathcal{G}_{\delta_2}(t) \cdots \mathcal{G}_{\delta_k}(t) \\ &\leq (\mathcal{G}_\delta(t))^k \leq \left(2\left(\frac{te}{\delta}\right)^\delta\right)^k,\end{aligned}$$

by Lemma 20.2.1 and Lemma 20.2.2. On the other hand, since Y is being shattered by $\mathcal{R}'$, this implies that $\left|\mathcal{R}'_{|Y}\right| = 2^t$. Thus, we have the inequality $2^t \leq \left(2(te/\delta)^\delta\right)^k$, which implies $t \leq k(1 + \delta \lg(te/\delta))$. Assume that $t \geq e$ and $\delta \lg(te/\delta) \geq 1$ since otherwise the claim is trivial, and observe that $t \leq k(1 + \delta \lg(te/\delta)) \leq 3k\delta \lg(t/\delta)$. Setting $x = t/\delta$, we have

$$\frac{t}{\delta} \leq 3k \frac{\ln(t/\delta)}{\ln 2} \leq 6k \ln \frac{t}{\delta} \implies \frac{x}{\ln x} \leq 6k \implies x \leq 2 \cdot 6k \ln(6k) \implies x \leq 12k \ln(6k),$$

by Lemma 20.2.5(C). We conclude that $t \leq 12\delta k \ln(6k)$, as claimed. ■

Corollary 20.2.15. *Let* $\mathsf{S} = (X, \mathcal{R})$ *and* $\mathsf{T} = (X, \mathcal{R}')$ *be two range spaces of* VC *dimension* δ *and* δ'*, respectively, where* $\delta, \delta' > 1$*. Let* $\widehat{\mathcal{R}} = \left\{ \mathbf{r} \cap \mathbf{r}' \;\middle|\; \mathbf{r} \in \mathcal{R}, \mathbf{r}' \in \mathcal{R}' \right\}$*. Then, for the range space* $\widehat{\mathsf{S}} = (X, \widehat{\mathcal{R}})$*, we have that* $\dim_{\mathsf{VC}}(\widehat{\mathsf{S}}) = O(\delta + \delta')$*.*

Corollary 20.2.16. *Any finite sequence of combining range spaces with finite* VC *dimension (by intersecting, complementing, or taking their union) results in a range space with a finite* VC *dimension.*

20.3. On ε-nets and ε-sampling

20.3.1. ε-nets and ε-samples

Definition 20.3.1 (ε-sample). Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space, and let x be a finite subset of X. For $0 \leq \varepsilon \leq 1$, a subset $C \subseteq \mathsf{x}$ is an ***ε-sample*** for x if for any range $\mathbf{r} \in \mathcal{R}$, we have

$$|\, \overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| \leq \varepsilon,$$

where $\overline{m}(\mathbf{r}) = |\mathsf{x} \cap \mathbf{r}| \,/\, |\mathsf{x}|$ is the measure of $\mathbf{r}$ (see Definition 20.1.2) and $\overline{s}(\mathbf{r}) = |C \cap \mathbf{r}| \,/\, |C|$ is the estimate of $\mathbf{r}$ (see Definition 20.1.3). (Here C might be a multi-set, and as such $|C \cap \mathbf{r}|$ is counted with multiplicity.)

As such, an ε-sample is a subset of the ground set x that "captures" the range space up to an error of ε. Specifically, to estimate the fraction of the ground set covered by a range $\mathbf{r}$, it is sufficient to count the points of C that fall inside $\mathbf{r}$.

If X is a finite set, we will abuse notation slightly and refer to C as an ***ε-sample*** for S.

To see the usage of such a sample, consider $\mathsf{x} = \mathsf{X}$ to be, say, the population of a country (i.e., an element of X is a citizen). A range in $\mathcal{R}$ is the set of all people in the country that answer yes to a question (i.e., would you vote for party Y?, would you buy a bridge from me?, questions like that). An ε-sample of this range space enables us to estimate reliably (up to an error of ε) the answers for all these questions, by just asking the people in the sample.

The natural question of course is how to find such a subset of small (or minimal) size.

Theorem 20.3.2 (ε-sample theorem, [VC71]). *There is a positive constant c such that if $(\mathsf{X}, \mathcal{R})$ is any range space with VC dimension at most δ, $\mathsf{x} \subseteq \mathsf{X}$ is a finite subset and $\varepsilon, \varphi > 0$, then a random subset $C \subseteq \mathsf{x}$ of cardinality*

$$s = \frac{c}{\varepsilon^2}\left(\delta \log \frac{\delta}{\varepsilon} + \log \frac{1}{\varphi}\right)$$

is an ε-sample for x with probability at least $1 - \varphi$.

(In the above theorem, if $s > |\mathsf{x}|$, then we can just take all of x to be the ε-sample.)

For a strengthened version of the above theorem with slightly better bounds is known [Har11].

Sometimes it is sufficient to have (hopefully smaller) samples with a weaker property – if a range is "heavy", then there is an element in our sample that is in this range.

Definition 20.3.3 (ε-net). A set $N \subseteq \mathsf{x}$ is an ***ε-net*** for x if for any range $\mathbf{r} \in \mathcal{R}$, if $\overline{m}(\mathbf{r}) \geq \varepsilon$ (i.e., $|\mathbf{r} \cap \mathsf{x}| \geq \varepsilon |\mathsf{x}|$), then $\mathbf{r}$ contains at least one point of N (i.e., $\mathbf{r} \cap N \neq \emptyset$).

Theorem 20.3.4 (ε-net theorem, [HW87]). *Let $(\mathsf{X}, \mathcal{R})$ be a range space of VC dimension δ, let x be a finite subset of X, and suppose that $0 < \varepsilon \leq 1$ and $\varphi < 1$. Let N be a set obtained by m random independent draws from x, where*

$$m \geq \max\left(\frac{4}{\varepsilon} \lg \frac{4}{\varphi}, \frac{8\delta}{\varepsilon} \lg \frac{16}{\varepsilon}\right). \tag{20.3}$$

Then N is an ε-net for x with probability at least $1 - \varphi$.

(We remind the reader that $\lg = \log_2$.)

The proofs of the above theorems are somewhat involved and we first turn our attention to some applications before presenting the proofs.

Remark 20.3.5. The above two theorems also hold for spaces with shattering dimension at most δ, in which case the sample size is slightly larger. Specifically, for Theorem 20.3.4, the sample size needed is $O\left(\frac{1}{\varepsilon} \lg \frac{1}{\varphi} + \frac{\delta}{\varepsilon} \lg \frac{\delta}{\varepsilon}\right)$.

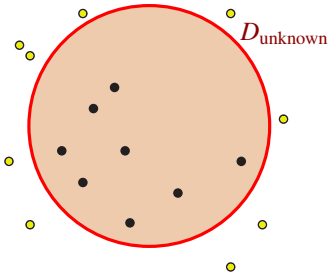
20.3.2. Some applications

We mention two (easy) applications of these theorems, which (hopefully) demonstrate their power.

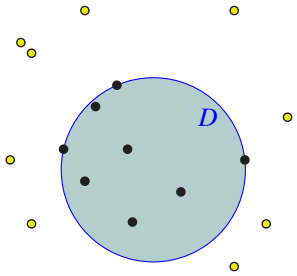
20.3.2.1. Range searching

So, consider a (very large) set of points P in the plane. We would like to be able to quickly decide how many points are included inside a query rectangle. Let us assume that we allow ourselves 1% error. What Theorem 20.3.2 tells us is that there is a subset of *constant size* (that depends only on ε) that can be used to perform this estimation, and it works for *all* query rectangles (we used here the fact that rectangles in the plane have finite VC dimension). In fact, a random sample of this size works with constant probability.

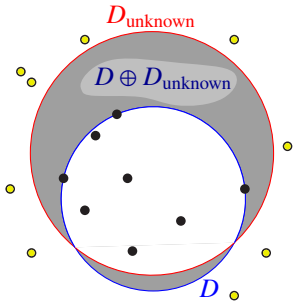
20.3.2.2. Learning a concept



Assume that we have a function f defined in the plane that returns '1' inside an (unknown) disk D_{unknown} and '0' outside it. There is some distribution $\mathcal{D}$ defined over the plane, and we pick points from this distribution. Furthermore, we can compute the function for these labels (i.e., we can compute f for certain values, but it is expensive). For a mystery value $\varepsilon > 0$, to be explained shortly, Theorem 20.3.4 tells us to pick (roughly) $O((1/\varepsilon)\log(1/\varepsilon))$ random points in a sample R from this distribution and to compute the labels for the samples. This is demonstrated in the figure on the right, where black dots are the sample points for which $f(\cdot)$ returned 1.



So, now we have positive examples and negative examples. We would like to find a hypothesis that agrees with all the samples we have and that hopefully is close to the true unknown disk underlying the function f. To this end, compute the smallest disk D that contains the sample labeled by '1' and does not contain any of the '0' points, and let $g : \mathbb{R}^2 \to \{0, 1\}$ be the function g that returns '1' inside the disk and '0' otherwise. We claim that g classifies correctly all but an ε-fraction of the points (i.e., the probability of misclassifying a point picked according to the given distribution is smaller than ε); that is, $\mathbf{Pr}_{\mathsf{p}\in\mathcal{D}}\left[f(\mathsf{p}) \neq g(\mathsf{p})\right] \leq \varepsilon$.



Geometrically, the region where g and f disagree is all the points in the symmetric difference between the two disks. That is, $\mathcal{E} = D \oplus D_{\text{unknown}}$; see the figure on the right.

Thus, consider the range space S having the plane as the ground set and the symmetric difference between any two disks as its ranges. By Corollary 20.2.16, this range space has finite VC dimension. Now, consider the (unknown) disk D' that induces f and the region $\mathbf{r} = D_{\text{unknown}} \oplus D$. Clearly, the learned classifier g returns incorrect answers only for points picked inside $\mathbf{r}$.

Thus, the probability of a mistake in the classification is the measure of $\mathbf{r}$ under the distribution $\mathcal{D}$. So, if $\mathbf{Pr}_{\mathcal{D}}[\mathbf{r}] > \varepsilon$ (i.e., the probability that a sample point falls inside $\mathbf{r}$), then by the ε-net theorem (i.e., Theorem 20.3.4) the set R is an ε-net for S (ignore for the time being the possibility that the random sample fails to be an ε-net) and as such, R contains a point q inside $\mathbf{r}$. But, it is not possible for g (which classifies correctly all the sampled points of R) to make a mistake on q, a contradiction, because by construction, the range $\mathbf{r}$ is where g misclassifies points. We conclude that $\mathbf{Pr}_{\mathcal{D}}[\mathbf{r}] \leq \varepsilon$, as desired.

Little lies. The careful reader might be tearing his or her hair out because of the above description. First, Theorem 20.3.4 might fail, and the above conclusion might not hold. This is of course true, and in real applications one might use a much larger sample to guarantee that the probability of failure is so small that it can be practically ignored. A more serious issue is that Theorem 20.3.4 is defined only for finite sets. Nowhere does it speak about a continuous distribution. Intuitively, one can approximate a continuous distribution to an arbitrary precision using a huge sample and apply the theorem to this sample as our ground set. A formal proof is more

tedious and requires extending the proof of Theorem 20.3.4 to continuous distributions. This is straightforward and we will ignore this topic altogether.

20.3.2.3. A naive proof of the ε-sample theorem.

To demonstrate why the ε-sample/net theorems are interesting, let us try to prove the ε-sample theorem in the natural naive way. Thus, consider a finite range space $\mathsf{S} = (\mathsf{x}, \mathcal{R})$ with shattering dimension δ. Also, consider a range $\mathbf{r}$ that contains, say, a p fraction of the points of x, where $p \geq \varepsilon$. Consider a random sample R of r points from x, picked with replacement.

Let p_i be the ith sample point, and let X_i be an indicator variable which is one if and only if $\mathsf{p}_i \in \mathbf{r}$. Clearly, $(\sum_i X_i)/r$ is an estimate for $p = |\mathbf{r} \cap \mathsf{x}| \,/\, |\mathsf{x}|$. We would like this estimate to be within $\pm\varepsilon$ of p and with confidence $\geq 1 - \varphi$.

As such, the sample failed if $\left|\sum_{i=1}^{r} X_i - pr\right| \geq \varepsilon r = (\varepsilon/p)pr$. Set $\phi = \varepsilon/p$ and $\mu = \mathbf{E}[\sum_i X_i] = pr$. Using Chernoff's inequality (Theorem 20.9.2$_{\text{p26}}$ and Theorem 20.9.2$_{\text{p26}}$), we have

$$\begin{aligned}\mathbf{Pr}\left[\left|\sum_{i=1}^{r} X_i - pr\right| \geq (\varepsilon/p)pr\right] &= \mathbf{Pr}\left[\left|\sum_{i=1}^{r} X_i - \mu\right| \geq \phi\mu\right] \leq \exp\left(-\mu\phi^2/2\right) + \exp\left(-\mu\phi^2/4\right) \\ &\leq 2\exp\left(-\mu\phi^2/4\right) = 2\exp\left(-\frac{\varepsilon^2}{4p}r\right) \leq \varphi,\end{aligned}$$

for $r \geq \left\lceil \frac{4}{\varepsilon^2} \ln \frac{2}{\varphi} \right\rceil \geq \left\lceil \frac{4p}{\varepsilon^2} \ln \frac{2}{\varphi} \right\rceil$.

Viola! We proved the ε-sample theorem. Well, not quite. We proved that the sample works correctly for a single range. Namely, we proved that for a specific range $\mathbf{r} \in \mathcal{R}$, we have that $\mathbf{Pr}[|\,\overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| > \varepsilon] \leq \varphi$. However, we need to prove that $\forall \mathbf{r} \in \mathcal{R}$, $\mathbf{Pr}[|\,\overline{m}(\mathbf{r}) - \overline{s}(\mathbf{r})| > \varepsilon] \leq \varphi$.

Now, naively, we can overcome this by using a union bound on the bad probability. Indeed, if there are k different ranges under consideration, then we can use a sample that is large enough such that the probability of it to fail for each range is at most φ/k. In particular, let $\mathcal{E}_i$ be the bad event that the sample fails for the ith range. We have that $\mathbf{Pr}[\mathcal{E}_i] \leq \varphi/k$, which implies that

$$\mathbf{Pr}[\text{sample fails for any range}] \leq \mathbf{Pr}\left[\bigcup_{i=1}^{k} \mathcal{E}_i\right] \leq \sum_{i=1}^{k} \mathbf{Pr}[\mathcal{E}_i] \leq k(\varphi/k) \leq \varphi,$$

by the union bound; that is, the sample works for all ranges with good probability.

However, the number of ranges that we need to prove the theorem for is $\pi_{\mathsf{S}}(|\mathsf{x}|)$ (see Definition 20.2.3). In particular, if we plug in confidence $\varphi/\pi_{\mathsf{S}}(|\mathsf{x}|)$ to the above analysis and use the union bound, we get that for

$$r \geq \left\lceil \frac{4}{\varepsilon^2} \ln \frac{\pi_{\mathsf{S}}(|\mathsf{x}|)}{\varphi} \right\rceil$$

the sample estimates correctly (up to $\pm\varepsilon$) the size of all ranges with confidence $\geq 1 - \varphi$. Bounding $\pi_{\mathsf{S}}(|\mathsf{x}|)$ by $O\left(|\mathsf{x}|^{\delta}\right)$ (using Eq. (20.2)$_{\text{p4}}$ for a space with VC dimension δ), we can bound the required size of r by $O\left(\delta\varepsilon^{-2}\log(|\mathsf{x}|/\varphi)\right)$. We summarize the result.

Lemma 20.3.6. *Let $(\mathsf{x}, \mathcal{R})$ be a finite range space with VC dimension at most δ, and let $\varepsilon, \varphi > 0$ be parameters. Then a random subset $C \subseteq \mathsf{x}$ of cardinality $O\left(\delta\varepsilon^{-2}\log(|\mathsf{x}|/\varphi)\right)$ is an ε-sample for x with probability at least $1 - \varphi$.*

Namely, the "naive" argumentation gives us a sample bound which depends on the underlying size of the ground set. However, the sample size in the ε-sample theorem (Theorem 20.3.2) is independent of the size of the ground set. This is the magical property of the ε-sample theorem[5].

Interestingly, using a chaining argument on Lemma 20.3.6, one can prove the ε-sample theorem for the finite case; see Exercise 20.8.3. We provide a similar proof when using discrepancy, in Section 20.4. However, the original proof uses a clever double sampling idea that is both interesting and insightful that makes the proof work for the infinite case also.

20.3.3. A quicky proof of the ε-net theorem (Theorem 20.3.4)

Here we provide a sketchy proof of Theorem 20.3.4, which conveys the main ideas. The full proof in all its glory and details is provided in Section 20.5.

Let $N = (x_1, \ldots, x_m)$ be the sample obtained by m independent samples from x (observe that N might contain the same element several times, and as such it is a multi-set). Let $\mathcal{E}_1$ be the probability that N fails to be an ε-net. Namely, for $n = |\mathsf{x}|$, let

$$\mathcal{E}_1 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n \quad \text{and} \quad \mathbf{r} \cap N = \emptyset \right\}.$$

To complete the proof, we must show that $\mathbf{Pr}[\mathcal{E}_1] \leq \varphi$.

Let $T = (y_1, \ldots, y_m)$ be another random sample generated in a similar fashion to N. It might be that N fails for a certain range $\mathbf{r}$, but then since T is an independent sample, we still expect that $|\mathbf{r} \cap T| = \varepsilon m$. In particular, the probability that $\mathbf{Pr}\left[|\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2}\right]$ is a large constant close to 1, regardless of how N performs. Indeed, if m is sufficiently large, we expect the random variable $|\mathbf{r} \cap T|$ to concentrate around εm, and one can argue this formally using Chernoff's inequality. Namely, intuitively, for a heavy range $\mathbf{r}$ we have that

$$\mathbf{Pr}[\mathbf{r} \cap N = \emptyset] \approx \mathbf{Pr}\left[\mathbf{r} \cap N = \emptyset \text{ and } \left(|\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2}\right)\right].$$

Inspired by this, let $\mathcal{E}_2$ be the event that N fails for some range $\mathbf{r}$ but T "works" for $\mathbf{r}$; formally

$$\mathcal{E}_2 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n,\ \mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Intuitively, since $\mathbf{E}[|\mathbf{r} \cap T|] \geq \varepsilon m$, then for the range $\mathbf{r}$ that N fails for, we have with "good" probability that $|\mathbf{r} \cap T| \geq \varepsilon m/2$. Namely, $\mathbf{Pr}[\mathcal{E}_1] \approx \mathbf{Pr}[\mathcal{E}_2]$.

Next, let

$$\mathcal{E}_2' = \left\{ \exists \mathbf{r} \in R \;\middle|\; \mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_2'$ and as such $\mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}\left[\mathcal{E}_2'\right]$. Now, fix $Z = N \cup T$, and observe that $|Z| = 2m$. Next, fix a range $\mathbf{r}$, and observe that the bad probability of $\mathcal{E}_2'$ is maximized if $|\mathbf{r} \cap Z| = \varepsilon m/2$. Now, the probability that all the elements of $\mathbf{r} \cap Z$ fall only into the second half of the sample is at most $2^{-\varepsilon m/2}$ as a careful calculation shows. Now, there are at most $\left|Z_{|R}\right| \leq \mathcal{G}_d(2m)$ different ranges that one has to consider. As such, $\mathbf{Pr}[\mathcal{E}_1] \approx \mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}\left[\mathcal{E}_2'\right] \leq \mathcal{G}_d(2m)2^{-\varepsilon m/2}$ and this is smaller than φ, as a careful calculation shows by just plugging the value of m into the right-hand side; see Eq. $(20.3)_{\text{p11}}$. ■

[5] The notion of magic is used here in the sense of Arthur C. Clarke's statement that "any sufficiently advanced technology is indistinguishable from magic."

20.4. Discrepancy

The proof of the ε-sample/net theorem is somewhat complicated. It turns out that one can get a somewhat similar result by attacking the problem from the other direction; namely, let us assume that we would like to take a truly large sample of a finite range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ defined over n elements with m ranges. We would like this sample to be as representative as possible as far as S is concerned. In fact, let us decide that we would like to pick exactly half of the points of X in our sample (assume that $n = |\mathsf{X}|$ is even).

To this end, let us color half of the points of X by -1 (i.e., black) and the other half by 1 (i.e., white). If for every range, $\mathbf{r} \in \mathcal{R}$, the number of black points inside it is equal to the number of white points, then doubling the number of black points inside a range gives us the exact number of points inside the range. Of course, such a perfect coloring is unachievable in almost all situations. To see this, consider the complete graph K_3 – clearly, in any coloring (by two colors) of its vertices, there must be an edge with two endpoints having the same color (i.e., the edges are the ranges).

Formally, let $\chi : \mathsf{X} \to \{-1, 1\}$ be a coloring. The *discrepancy* of χ over a range $\mathbf{r}$ is the amount of imbalance in the coloring inside χ. Namely,

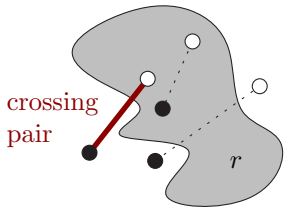
$$|\chi(\mathbf{r})| = \left| \sum_{\mathsf{p} \in \mathbf{r}} \chi(\mathsf{p}) \right|.$$

The overall *discrepancy* of χ is $\operatorname{disc}(\chi) = \max_{\mathbf{r} \in \mathcal{R}} |\chi(\mathbf{r})|$. The *discrepancy* of a (finite) range space $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ is the discrepancy of the best possible coloring; namely,

$$\operatorname{disc}(\mathsf{S}) = \min_{\chi : \mathsf{X} \to \{-1,+1\}} \operatorname{disc}(\chi).$$

The natural question is, of course, how to compute the coloring χ of minimum discrepancy. This seems like a very challenging question, but when you do not know what to do, you might as well do something random. So, let us pick a random coloring χ of X. To this end, let Π be an arbitrary partition of X into pairs (i.e., a perfect matching). For a pair $\{\mathsf{p}, \mathsf{q}\} \in \Pi$, we will either color $\chi(\mathsf{p}) = -1$ and $\chi(\mathsf{q}) = 1$ or the other way around; namely, $\chi(\mathsf{p}) = 1$ and $\chi(\mathsf{q}) = -1$. We will decide how to color this pair using a single coin flip. Thus, our coloring would be induced by making such a decision for every pair of Π, and let χ be the resulting coloring. We will refer to χ as ***compatible*** with the partition Π if, for all $\{\mathsf{p}, \mathsf{q}\} \in \Pi$, we have that $\chi(\{\mathsf{p}, \mathsf{q}\}) = 0$; namely,

$$\begin{aligned} \forall \{\mathsf{p}, \mathsf{q}\} \in \Pi \quad & (\chi(\mathsf{p}) = +1 \text{ and } \chi(\mathsf{q}) = -1) \\ & \text{or} \quad (\chi(\mathsf{p}) = -1 \text{ and } \chi(\mathsf{q}) = +1). \end{aligned}$$



.

Consider a range $\mathbf{r}$ and a coloring χ compatible with Π. If a pair $\{\mathsf{p}, \mathsf{q}\} \in \Pi$ falls completely inside $\mathbf{r}$ or completely outside $\mathbf{r}$, then it does not contribute anything to the discrepancy of $\mathbf{r}$. Thus, the only pairs that contribute to the discrepancy of $\mathbf{r}$ are the ones that ***cross*** it. Namely, $\{\mathsf{p}, \mathsf{q}\} \cap \mathbf{r} \neq \emptyset$ and $\{\mathsf{p}, \mathsf{q}\} \cap (\mathsf{X} \setminus \mathbf{r}) \neq \emptyset$.

As such, let $\#_{\mathbf{r}}$ denote the *crossing number* of $\mathbf{r}$, that is, the number of pairs that cross $\mathbf{r}$. Next, let $X_i \in \{-1, +1\}$ be the indicator variable which is the contribution of the ith crossing pair to the discrepancy of $\mathbf{r}$. For $\Delta_{\mathbf{r}} = \sqrt{2\#_{\mathbf{r}} \ln(4m)}$, we have by Chernoff's inequality (Theorem **??**$_{\text{p??}}$), that

$$\begin{aligned} \mathbf{Pr}[|\chi(\mathbf{r})| \geq \Delta_{\mathbf{r}}] &= \mathbf{Pr}[\chi(\mathbf{r}) \geq \Delta_{\mathbf{r}}] + \mathbf{Pr}[\chi(\mathbf{r}) \leq -\Delta_{\mathbf{r}}] = 2\,\mathbf{Pr}\left[\sum_i X_i \geq \Delta_{\mathbf{r}}\right] \\ &\leq 2\exp\left(-\frac{\Delta_{\mathbf{r}}^2}{2\#_{\mathbf{r}}}\right) = \frac{1}{2m}. \end{aligned}$$

Since there are m ranges in $\mathcal{R}$, it follows that with good probability (i.e., at least half) for all $\mathbf{r} \in \mathcal{R}$ the discrepancy of $\mathbf{r}$ is at most $\Delta_{\mathbf{r}}$.

Theorem 20.4.1. *Let $\mathsf{S} = (X, \mathcal{R})$ be a range space defined over $n = |X|$ elements with $m = |\mathcal{R}|$ ranges. Consider any partition Π of the elements of X into pairs. Then, with probability $\geq 1/2$, for any range $\mathbf{r} \in \mathcal{R}$, a random coloring $\chi : X \to \{-1, +1\}$ that is compatible with the partition Π has discrepancy at most*

$$|\chi(\mathbf{r})| < \Delta_{\mathbf{r}} = \sqrt{2 \#_{\mathbf{r}} \ln(4m)},$$

where $\#_{\mathbf{r}}$ denotes the number of pairs of Π that cross $\mathbf{r}$. In particular, since $\#_{\mathbf{r}} \leq |\mathbf{r}|$, we have $|\chi(\mathbf{r})| \leq \sqrt{2\,|\mathbf{r}| \ln(4m)}$.

Observe that for every range $\mathbf{r}$ we have that $\#_{\mathbf{r}} \leq n/2$, since $2\#_{\mathbf{r}} \leq |X|$. As such, we have:

Corollary 20.4.2. *Let $\mathsf{S} = (X, \mathcal{R})$ be a range space defined over n elements with m ranges. Let Π be an arbitrary partition of X into pairs. Then a random coloring which is compatible with Π has $\operatorname{disc}(\chi) < \sqrt{n \ln(4m)}$, with probability $\geq 1/2$.*

One can easily amplify the probability of success of the coloring by increasing the threshold. In particular, for any constant $c \geq 1$, one has that

$$\forall \mathbf{r} \in \mathcal{R} \quad |\chi(\mathbf{r})| \leq \sqrt{2c \#_{\mathbf{r}} \ln(4m)},$$

with probability $\geq 1 - \dfrac{2}{(4m)^c}$.

20.4.1. Building ε-sample via discrepancy

Let $\mathsf{S} = (\mathsf{X}, \mathcal{R})$ be a range space with shattering dimension δ. Let $\mathsf{P} \subseteq \mathsf{X}$ be a set of n points, and consider the induced range space $\mathsf{S}_{|\mathsf{P}} = (\mathsf{P}, \mathcal{R}_{|\mathsf{P}})$; see Definition 20.1.4$_{\text{p2}}$. Here, by the definition of shattering dimension, we have that $m = \left|\mathcal{R}_{|\mathsf{P}}\right| = O\left(n^{\delta}\right)$. Without loss of generality, we assume that n is a power of 2. Consider a coloring χ of P with discrepancy bounded by Corollary 20.4.2. In particular, let Q be the points of P colored by, say, -1. We know that $|\mathsf{Q}| = n/2$, and for any range $\mathbf{r} \in \mathcal{R}$, we have that

$$\chi(\mathbf{r}) = \|(\mathsf{P} \setminus \mathsf{Q}) \cap \mathbf{r}| - |\mathsf{Q} \cap \mathbf{r}\| < \sqrt{n \ln(4m)} = \sqrt{n \ln O(n^{\delta})} \leq c \sqrt{n \ln(n^{\delta})},$$

for some absolute constant c. Observe that $|(\mathsf{P} \setminus \mathsf{Q}) \cap \mathbf{r}| = |\mathsf{P} \cap \mathbf{r}| - |\mathsf{Q} \cap \mathbf{r}|$. In particular, we have that for any range $\mathbf{r}$,

$$\|\mathsf{P} \cap \mathbf{r}| - 2\,|\mathsf{Q} \cap \mathbf{r}\| \leq c \sqrt{n \ln(n^{\delta})}. \tag{20.4}$$

Dividing both sides by $n = |\mathsf{P}| = 2\,|\mathsf{Q}|$, we have that

$$\left| \frac{|\mathsf{P} \cap \mathbf{r}|}{|\mathsf{P}|} - \frac{|\mathsf{Q} \cap \mathbf{r}|}{|\mathsf{Q}|} \right| \leq \tau(n) \quad \text{for} \quad \tau(n) = c \sqrt{\frac{\delta \ln n}{n}}. \tag{20.5}$$

Namely, a coloring with discrepancy bounded by Corollary 20.4.2 yields a $\tau(n)$-sample. Intuitively, if n is very large, then Q provides a good approximation to P. However, we want an ε-sample for a prespecified $\varepsilon > 0$. Conceptually, ε is a fixed constant while $\tau(n)$ is considerably smaller. Namely, Q is a sample which is too tight for our purposes (and thus too big). As such, we will coarsen (and shrink) Q till we get the desired ε-sample by repeated application of Corollary 20.4.2. Specifically, we can "chain" together several approximations generated by Corollary 20.4.2. This is sometime refereed to as the *sketch* property of samples. Informally, as testified by the following lemma, a sketch of a sketch is a sketch[6].

[6] Try saying this quickly 100 times.

Lemma 20.4.3. *Let* $\mathsf{Q} \subseteq \mathsf{P}$ *be a* ρ*-sample for* P *(in some underlying range space* S*), and let* $\mathsf{R} \subseteq \mathsf{Q}$ *be a* ρ'*-sample for* Q*. Then* R *is a* $(\rho + \rho')$*-sample for* P*.*

Proof: By definition, we have that, for every $\mathbf{r} \in \mathcal{R}$,

$$\left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|}\right| \leq \rho \qquad \text{and} \qquad \left|\frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| \leq \rho'.$$

By adding the two inequalities together, we get

$$\left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| = \left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} + \frac{|\mathbf{r} \cap \mathsf{Q}|}{|\mathsf{Q}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| \leq \rho + \rho'. \qquad \blacksquare$$

Thus, let $\mathsf{P}_0 = \mathsf{P}$ and $\mathsf{P}_1 = \mathsf{Q}$. Now, in the ith iteration, we will compute a coloring χ_{i-1} of P_{i-1} with low discrepancy, as guaranteed by Corollary 20.4.2, and let P_i be the points of P_{i-1} colored white by χ_{i-1}. Let $\delta_i = \tau(n_{i-1})$, where $n_{i-1} = |\mathsf{P}_{i-1}| = n/2^{i-1}$. By Lemma 20.4.3, we have that P_k is a $(\sum_{i=1}^{k} \delta_i)$-sample for P. Since we would like the smallest set in the sequence $\mathsf{P}_1, \mathsf{P}_2, \ldots$ that is still an ε-sample, we would like to find the maximal k, such that $(\sum_{i=1}^{k} \delta_i) \leq \varepsilon$. Plugging in the value of δ_i and $\tau(\cdot)$, see Eq. (20.5), it is sufficient for our purposes that

$$\sum_{i=1}^{k} \delta_i = \sum_{i=1}^{k} \tau(n_{i-1}) = \sum_{i=1}^{k} c\sqrt{\frac{\delta \ln(n/2^{i-1})}{n/2^{i-1}}} \leq c_1 \sqrt{\frac{\delta \ln(n/2^{k-1})}{n/2^{k-1}}} = c_1 \sqrt{\frac{\delta \ln n_{k-1}}{n_{k-1}}} \leq \varepsilon,$$

since the above series behaves like a geometric series, and as such its total sum is proportional to its largest element[7], where c_1 is a sufficiently large constant. This holds for

$$c_1 \sqrt{\frac{\delta \ln n_{k-1}}{n_{k-1}}} \leq \varepsilon \iff c_1^2 \frac{\delta \ln n_{k-1}}{n_{k-1}} \leq \varepsilon^2 \iff \frac{c_1^2 \delta}{\varepsilon^2} \leq \frac{n_{k-1}}{\ln n_{k-1}}.$$

The last inequality holds for $n_{k-1} \geq 2\frac{c_1^2\delta}{\varepsilon^2} \ln \frac{c_1^2\delta}{\varepsilon^2}$, by Lemma 20.2.5(D). In particular, taking the largest k for which this holds results in a set P_k of size $O\Big((\delta/\varepsilon^2)\ln(\delta/\varepsilon)\Big)$ which is an ε-sample for P.

Theorem 20.4.4 (ε-sample via discrepancy). *For a range space* $(X, \mathcal{R})$ *with shattering dimension at most* δ *and* $B \subseteq X$ *a finite subset and* $\varepsilon > 0$*, there exists a subset* $C \subseteq B$*, of cardinality* $O\Big((\delta/\varepsilon^2)\ln(\delta/\varepsilon)\Big)$*, such that* C *is an* ε*-sample for* B*.*

Note that it is not obvious how to turn Theorem 20.4.4 into an efficient construction algorithm of such an ε-sample. Nevertheless, this theorem can be turned into a relatively efficient deterministic algorithm using conditional probabilities. In particular, there is a deterministic $O\Big(n^{\delta+1}\Big)$ time algorithm for computing an ε-sample for a range space of VC dimension δ and with n points in its ground set using the above approach (see the bibliographical notes in Section 20.7 for details). Inherently, however, it is a far cry from the simplicity of Theorem 20.3.2 that just requires us to take a random sample. Interestingly, there are cases where using discrepancy leads to smaller ε-samples; again see bibliographical notes for details.

[7] Formally, one needs to show that the ratio between two consecutive elements in the series is larger than some constant, say 1.1. This is easy but tedious, but the well-motivated reader (of little faith) might want to do this calculation.

20.4.1.1. Faster deterministic construction of ε-samples.

One can speed up the deterministic construction mentioned above by using a sketch-and-merge approach. To this end, we need the following *merge* property of ε-samples. (The proof of the following lemma is quite easy. Nevertheless, we provide the proof in excruciating detail for the sake of completeness.)

Lemma 20.4.5. *Consider the sets* $\mathsf{R} \subseteq \mathsf{P}$ *and* $\mathsf{R}' \subseteq \mathsf{P}'$. *Assume that* P *and* P' *are disjoint,* $|\mathsf{P}| = |\mathsf{P}'|$, *and* $|\mathsf{R}| = |\mathsf{R}'|$. *Then, if* R *is an* ε*-sample of* P *and* R' *is an* ε*-sample of* P', *then* $\mathsf{R} \cup \mathsf{R}'$ *is an* ε*-sample of* $\mathsf{P} \cup \mathsf{P}'$.

Proof: We have for any range $\mathbf{r}$ that

$$\begin{aligned}
\left|\frac{|\mathbf{r} \cap (\mathsf{P} \cup \mathsf{P}')|}{|\mathsf{P} \cup \mathsf{P}'|} - \frac{|\mathbf{r} \cap (\mathsf{R} \cup \mathsf{R}')|}{|\mathsf{R} \cup \mathsf{R}'|}\right| &= \left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P} \cup \mathsf{P}'|} + \frac{|\mathbf{r} \cap \mathsf{P}'|}{|\mathsf{P} \cup \mathsf{P}'|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R} \cup \mathsf{R}'|} - \frac{|\mathbf{r} \cap \mathsf{R}'|}{|\mathsf{R} \cup \mathsf{R}'|}\right| \\
&= \left|\frac{|\mathbf{r} \cap \mathsf{P}|}{2\,|\mathsf{P}|} + \frac{|\mathbf{r} \cap \mathsf{P}'|}{2\,|\mathsf{P}'|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{2\,|\mathsf{R}|} - \frac{|\mathbf{r} \cap \mathsf{R}'|}{2\,|\mathsf{R}'|}\right| \\
&= \frac{1}{2}\left|\left(\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right) + \left(\frac{|\mathbf{r} \cap \mathsf{P}'|}{|\mathsf{P}'|} - \frac{|\mathbf{r} \cap \mathsf{R}'|}{|\mathsf{R}'|}\right)\right| \\
&\leq \frac{1}{2}\left|\frac{|\mathbf{r} \cap \mathsf{P}|}{|\mathsf{P}|} - \frac{|\mathbf{r} \cap \mathsf{R}|}{|\mathsf{R}|}\right| + \frac{1}{2}\left|\frac{|\mathbf{r} \cap \mathsf{P}'|}{|\mathsf{P}'|} - \frac{|\mathbf{r} \cap \mathsf{R}'|}{|\mathsf{R}'|}\right| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

■

Interestingly, by breaking the given ground sets into sets of equal size and building a balanced binary tree over these sets, one can speed up the deterministic algorithm for building ε-samples. The idea is to compute the sample bottom-up, where at every node we merge the samples provided by the children (i.e., using Lemma 20.4.5), and then we sketch the resulting set using Lemma 20.4.3. By carefully fine-tuning this construction, one can get an algorithm for computing ε-samples in time which is near linear in n (assuming ε and δ are small constants). We delegate the details of this construction to Exercise 20.8.6.

This algorithmic idea is quite useful and we will refer to it as ***sketch-and-merge***.

20.4.2. Building ε-net via discrepancy

We are given range space $(\mathsf{X}, \mathcal{R})$ with shattering dimension d and $\varepsilon > 0$ and the target is to compute an ε-net for this range space.

We need to be slightly more careful if we want to use discrepancy to build ε-nets, and we will use Theorem 20.4.1 instead of Corollary 20.4.2 in the analysis.

The construction is as before – we set $\mathsf{P}_0 = \mathsf{P}$, and P_i is all the points colored $+1$ in the coloring of P_{i-1} by Theorem 20.4.1. We repeat this till we get a set that is the required net.

To analyze this construction (and decide when it should stop), let $\mathbf{r}$ be a range in a given range space $(\mathsf{X}, \mathcal{R})$ with shattering dimension d, and let

$$\nu_i = |\mathsf{P}_i \cap \mathbf{r}|$$

denote the size of the range $\mathbf{r}$ in the ith set P_i and let $n_i = |\mathsf{P}_i|$, for $i \geq 0$. Observer that the number of points in $\mathbf{r}$ colored by $+1$ and -1 when coloring P_{i-1} is

$$\alpha_i = |\mathsf{P}_i \cap \mathbf{r}| = \nu_i \quad \text{and} \quad \beta_i = |\mathsf{P}_{i-1} \cap \mathbf{r}| - |\mathsf{P}_i \cap \mathbf{r}| = \nu_{i-1} - \nu_i,$$

respectively. As such, setting $m_i = \left|\mathcal{R}_{|\mathsf{P}_i}\right| = O\left(n_i^d\right)$, we have, by Theorem 20.4.1, that the discrepancy of $\mathbf{r}$ in this coloring of P_{i-1} is

$$|\alpha_i - \beta_i| = |\nu_i - 2\nu_{i-1}| \leq \sqrt{2\nu_{i-1} \ln 4m_{i-1}} \leq c\sqrt{d\nu_{i-1} \ln n_{i-1}}$$

for some constant c, since the crossing number $\#_{\mathbf{r}}$ of a range $\mathbf{r} \cap \mathsf{P}_{i-1}$ is always bounded by its size. This is equivalent to

$$\left|2^{i-1}\nu_{i-1} - 2^i\nu_i\right| \leq c2^{i-1}\sqrt{d\nu_{i-1} \ln n_{i-1}}. \tag{20.6}$$

We need the following technical claim that states that the size of ν_k behaves as we expect; as long as the set P_k is large enough, the size of ν_k is roughly $\nu_0/2^k$.

Claim 20.4.6. *There is a constant c_4 (independent of d), such that for all k with $\nu_0/2^k \geq c_4 d \ln n_k$, $(\nu_0/2^k)/2 \leq \nu_k \leq 2(\nu_0/2^k)$.*

Proof: The proof is by induction. For $k = 0$ the claim trivially holds. Assume that it holds for $i < k$. Adding up the inequalities of Eq. (20.6), for $i = 1, \ldots, k$, we have that

$$\left|\nu_0 - 2^k\nu_k\right| \leq \sum_{i=1}^{k} c2^{i-1}\sqrt{d\nu_{i-1} \ln n_{i-1}} \leq \sum_{i=1}^{k} c2^{i-1}\sqrt{2d\frac{\nu_0}{2^{i-1}} \ln n_{i-1}} \leq c_3\, 2^k\sqrt{d\frac{\nu_0}{2^k} \ln n_k},$$

for some constant c_3 since this summation behaves like an increasing geometric series and the last term dominates the summation. Thus,

$$\frac{\nu_0}{2^k} - c_3\sqrt{d\frac{\nu_0}{2^k} \ln n_k} \;\leq\; \nu_k \;\leq\; \frac{\nu_0}{2^k} + c_3\sqrt{d\frac{\nu_0}{2^k} \ln n_k}.$$

By assumption, we have that $\sqrt{\dfrac{\nu_0}{c_4 2^k}} \geq \sqrt{d \ln n_k}$. This implies that

$$\nu_k \leq \frac{\nu_0}{2^k} + c_3\sqrt{\frac{\nu_0}{2^k} \cdot \frac{\nu_0}{c_4 2^k}} = \frac{\nu_0}{2^k}\left(1 + \frac{c_3}{\sqrt{c_4}}\right) \leq 2\frac{\nu_0}{2^k},$$

by selecting $c_4 \geq 4c_3^2$. Similarly, we have

$$\nu_k \geq \frac{\nu_0}{2^k}\left(1 - \frac{c_3\sqrt{d \ln n_k}}{\sqrt{\nu_0/2^k}}\right) \geq \frac{\nu_0}{2^k}\left(1 - \frac{c_3\sqrt{\nu_0/c_4 2^k}}{\sqrt{\nu_0/2^k}}\right) = \frac{\nu_0}{2^k}\left(1 - \frac{c_3}{\sqrt{c_4}}\right) \geq \frac{\nu_0}{2^k}/2. \qquad \blacksquare$$

So consider a "heavy" range $\mathbf{r}$ that contains at least $\nu_0 \geq \varepsilon n$ points of P. To show that P_k is an ε-net, we need to show that $\mathsf{P}_k \cap \mathbf{r} \neq \emptyset$. To apply Claim 20.4.6, we need a k such that $\varepsilon n/2^k \geq c_4 d \ln n_{k-1}$, or equivalently, such that

$$\frac{2n_k}{\ln(2n_k)} \geq \frac{2c_4 d}{\varepsilon},$$

which holds for $n_k = \Omega\left(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon}\right)$, by Lemma 20.2.5(D). But then, by Claim 20.4.6, we have that

$$\nu_k = |\mathsf{P}_k \cap \mathbf{r}| \geq \frac{|\mathsf{P} \cap \mathbf{r}|}{2 \cdot 2^k} \geq \frac{1}{2} \cdot \frac{\varepsilon n}{2^k} = \frac{\varepsilon}{2} n_k = \Omega\left(d \ln \frac{d}{\varepsilon}\right) > 0.$$

We conclude that the set P_k, which is of size $\Omega\left(\frac{d}{\varepsilon} \ln \frac{d}{\varepsilon}\right)$, is an ε-net for P.

Theorem 20.4.7 (ε-net via discrepancy). *For any range space $(X, \mathcal{R})$ with shattering dimension at most d, a finite subset $B \subseteq X$, and $\varepsilon > 0$, there exists a subset $C \subseteq B$, of cardinality $O((d/\varepsilon)\ln(d/\varepsilon))$, such that C is an ε-net for B.*

20.5. Proof of the ε-net theorem

In this section, we finally prove Theorem 20.3.4.

Let $(\mathsf{X}, \mathcal{R})$ be a range space of VC dimension δ, and let x be a subset of X of cardinality n. Suppose that m satisfies Eq. (20.3)$_{\text{p11}}$. Let $N = (x_1, \ldots, x_m)$ be the sample obtained by m independent samples from x (the elements of N are not necessarily distinct, and we treat N as an ordered set). Let $\mathcal{E}_1$ be the probability that N fails to be an ε-net. Namely,

$$\mathcal{E}_1 = \left\{ \exists \mathbf{r} \in \mathcal{R} \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n \quad \text{and} \quad \mathbf{r} \cap N = \emptyset \right\}.$$

(Namely, there exists a "heavy" range $\mathbf{r}$ that does not contain any point of N.) To complete the proof, we must show that $\mathbf{Pr}[\mathcal{E}_1] \leq \varphi$. Let $T = (y_1, \ldots, y_m)$ be another random sample generated in a similar fashion to N. Let $\mathcal{E}_2$ be the event that N fails but T "works"; formally

$$\mathcal{E}_2 = \left\{ \exists \mathbf{r} \in R \;\middle|\; |\mathbf{r} \cap \mathsf{x}| \geq \varepsilon n,\ \mathbf{r} \cap N = \emptyset,\ \text{and}\ |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Intuitively, since $\mathbf{E}\Big[|\mathbf{r} \cap T|\Big] \geq \varepsilon m$, we have that for the range $\mathbf{r}$ that N fails for, it follows with "good" probability that $|\mathbf{r} \cap T| \geq \varepsilon m/2$. Namely, $\mathcal{E}_1$ and $\mathcal{E}_2$ have more or less the same probability.

Claim 20.5.1. $\mathbf{Pr}\Big[\mathcal{E}_2\Big] \leq \mathbf{Pr}\Big[\mathcal{E}_1\Big] \leq 2\,\mathbf{Pr}\Big[\mathcal{E}_2\Big]$.

Proof: Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_1$, and thus $\mathbf{Pr}\Big[\mathcal{E}_2\Big] \leq \mathbf{Pr}\Big[\mathcal{E}_1\Big]$. As for the other part, note that by the definition of conditional probability, we have

$$\mathbf{Pr}\Big[\mathcal{E}_2 \;\Big|\; \mathcal{E}_1\Big] = \mathbf{Pr}\Big[\mathcal{E}_2 \cap \mathcal{E}_1\Big] / \mathbf{Pr}\Big[\mathcal{E}_1\Big] = \mathbf{Pr}\Big[\mathcal{E}_2\Big] / \mathbf{Pr}\Big[\mathcal{E}_1\Big].$$

It is thus enough to show that $\mathbf{Pr}\Big[\mathcal{E}_2 \;\Big|\; \mathcal{E}_1\Big] \geq 1/2$.

Assume that $\mathcal{E}_1$ occurs. There is $\mathbf{r} \in R$, such that $|\mathbf{r} \cap \mathsf{x}| > \varepsilon n$ and $\mathbf{r} \cap N = \emptyset$. The required probability is at least the probability that for this specific $\mathbf{r}$, we have $|\mathbf{r} \cap T| \geq \frac{\varepsilon n}{2}$. However, $X = |\mathbf{r} \cap T|$ is a binomial variable with expectation $\mathbf{E}[X] = pm$, and variance $\mathbf{V}[X] = p(1-p)m \leq pm$, where $p = |\mathbf{r} \cap \mathsf{x}|/n \geq \varepsilon$. Thus, by Chebychev's inequality (Theorem **??**$_{\text{p??}}$),

$$\begin{aligned}
\mathbf{Pr}\left[X < \frac{\varepsilon m}{2}\right] &\leq \mathbf{Pr}\left[X < \frac{pm}{2}\right] \leq \mathbf{Pr}\left[|X - pm| > \frac{pm}{2}\right] \\
&= \mathbf{Pr}\left[|X - pm| > \frac{\sqrt{pm}}{2}\sqrt{pm}\right] \leq \mathbf{Pr}\left[\Big|X - \mathbf{E}[X]\Big| > \frac{\sqrt{pm}}{2}\sqrt{\mathbf{V}[X]}\right] \\
&\leq \left(\frac{2}{\sqrt{pm}}\right)^2 \leq \frac{1}{2},
\end{aligned}$$

since $m \geq 8/\varepsilon \geq 8/p$; see Eq. (20.3)$_{\text{p11}}$. Thus, for $\mathbf{r} \in \mathcal{E}_1$, we have

$$\frac{\mathbf{Pr}[\mathcal{E}_2]}{\mathbf{Pr}[\mathcal{E}_1]} \geq \mathbf{Pr}\Big[|\mathbf{r} \cap T| \geq \tfrac{\varepsilon m}{2}\Big] = 1 - \mathbf{Pr}\left[|\mathbf{r} \cap T| < \frac{\varepsilon m}{2}\right] \geq \frac{1}{2}. \qquad \blacksquare$$

Claim 20.5.1 implies that to bound the probability of $\mathcal{E}_1$, it is enough to bound the probability of $\mathcal{E}_2$. Let

$$\mathcal{E}_2' = \left\{ \exists \mathbf{r} \in R \;\middle|\; \mathbf{r} \cap N = \emptyset, |\mathbf{r} \cap T| \geq \frac{\varepsilon m}{2} \right\}.$$

Clearly, $\mathcal{E}_2 \subseteq \mathcal{E}_2'$. Thus, bounding the probability of $\mathcal{E}_2'$ is enough to prove Theorem 20.3.4. Note, however, that a shocking thing happened! We no longer have x participating in our event. Namely, we turned bounding an event that depends on a global quantity (i.e., the ground set x) into bounding a quantity that depends only on a local quantity/experiment (involving only N and T). This is the crucial idea in this proof.

Claim 20.5.2. $\mathbf{Pr}[\mathcal{E}_2] \leq \mathbf{Pr}\left[\mathcal{E}'_2\right] \leq \mathcal{G}_d(2m)2^{-\varepsilon m/2}$.

Proof: We imagine that we sample the elements of $N \cup T$ together, by picking $Z = (z_1, \ldots, z_{2m})$ independently from x. Next, we randomly decide the m elements of Z that go into N, and the remaining elements go into T. Clearly,

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}'_2] &= \sum_{z \in \mathsf{x}^{2m}} \mathbf{Pr}[\mathcal{E}'_2 \cap (Z = z)] = \sum_{z \in \mathsf{x}^{2m}} \frac{\mathbf{Pr}\left[\mathcal{E}'_2 \cap (Z = z)\right]}{\mathbf{Pr}[Z = z]} \cdot \mathbf{Pr}[Z = z] \\ &= \sum_{z} \mathbf{Pr}\left[\mathcal{E}'_2 \,\middle|\, Z = z\right] \mathbf{Pr}[Z = z] = \mathbf{E}\left[\mathbf{Pr}\left[\mathcal{E}'_2 \,\middle|\, Z = z\right]\right].\end{aligned}$$

Thus, from this point on, we fix the set Z, and we bound $\mathbf{Pr}\left[\mathcal{E}'_2 \,\middle|\, Z\right]$. Note that $\mathbf{Pr}[\mathcal{E}'_2]$ is a weighted average of $\mathbf{Pr}[\mathcal{E}'_2 | Z = z]$, and as such a bound on this quantity would imply the same bound on $\mathbf{Pr}[\mathcal{E}'_2]$.

It is now enough to consider the ranges in the projection space $(Z, \mathcal{R}_{|Z})$ (which has VC dimension δ). By Lemma 20.2.1, we have $\left|\mathcal{R}_{|Z}\right| \leq \mathcal{G}_\delta(2m)$.

Let us fix any $\mathbf{r} \in \mathcal{R}_{|Z}$, and consider the event

$$\mathcal{E}_{\mathbf{r}} = \left\{\mathbf{r} \cap N = \emptyset \text{ and } |\mathbf{r} \cap T| > \frac{\varepsilon m}{2}\right\}.$$

We claim that $\mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] \leq 2^{-\varepsilon m/2}$. Observe that if $k = |\mathbf{r} \cap (N \cup T)| \leq \varepsilon m/2$, then the event is empty, and this claim trivially holds. Otherwise, $\mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] = \mathbf{Pr}[\mathbf{r} \cap N = \emptyset]$. To bound this probability, observe that we have the $2m$ elements of Z, and we can choose any m of them to be N, as long as none of them is one of the k "forbidden" elements of $\mathbf{r} \cap (N \cup T)$. The probability of that is $\binom{2m-k}{m}/\binom{2m}{m}$. We thus have

$$\begin{aligned}\mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] \leq \mathbf{Pr}[\mathbf{r} \cap N = \emptyset] &= \frac{\binom{2m-k}{m}}{\binom{2m}{m}} = \frac{(2m-k)(2m-k-1)\cdots(m-k+1)}{2m(2m-1)\cdots(m+1)} \\ &= \frac{m(m-1)\cdots(m-k+1)}{2m(2m-1)\cdots(2m-k+1)} \leq 2^{-k} \leq 2^{-\varepsilon m/2}.\end{aligned}$$

Thus,

$$\mathbf{Pr}\left[\mathcal{E}'_2 \,\middle|\, Z\right] = \mathbf{Pr}\left[\bigcup_{\mathbf{r} \in \mathcal{R}_{|Z}} \mathcal{E}_{\mathbf{r}}\right] \leq \sum_{\mathbf{r} \in \mathcal{R}_{|Z}} \mathbf{Pr}[\mathcal{E}_{\mathbf{r}}] \leq \left|\mathcal{R}_{|Z}\right| 2^{-\varepsilon m/2} \leq \mathcal{G}_\delta(2m)2^{-\varepsilon m/2},$$

implying that $\mathbf{Pr}\left[\mathcal{E}'_2\right] \leq \mathcal{G}_\delta(2m)2^{-\varepsilon m/2}$. ∎

Proof of Theorem 20.3.4. By Claim 20.5.1 and Claim 20.5.2, we have that $\mathbf{Pr}[\mathcal{E}_1] \leq 2\mathcal{G}_\delta(2m)2^{-\varepsilon m/2}$. It thus remains to verify that if m satisfies Eq. (20.3), then $2\mathcal{G}_\delta(2m)2^{-\varepsilon m/2} \leq \varphi$.

Indeed, we know that $2m \geq 8\delta$ (by Eq. $(20.3)_{\text{p11}}$) and by Lemma 20.2.2, $\mathcal{G}_\delta(2m) \leq 2(2em/\delta)^\delta$, for $\delta \geq 1$. Thus, it is sufficient to show that the inequality $4(2em/\delta)^\delta 2^{-\varepsilon m/2} \leq \varphi$ holds. By rearranging and taking lg of both sides, we have that this is equivalent to

$$2^{\varepsilon m/2} \geq \frac{4}{\varphi}\left(\frac{2em}{\delta}\right)^\delta \qquad \Longrightarrow \qquad \frac{\varepsilon m}{2} \geq \delta \lg \frac{2em}{\delta} + \lg \frac{4}{\varphi}.$$

By our choice of m (see Eq. (20.3)), we have that $\varepsilon m/4 \geq \lg(4/\varphi)$. Thus, we need to show that

$$\frac{\varepsilon m}{4} \geq \delta \lg \frac{2em}{\delta}.$$

We verify this inequality for $m = \dfrac{8\delta}{\varepsilon} \lg \dfrac{16}{\varepsilon}$ (this would also hold for bigger values, as can be easily verified). Indeed

$$2\delta \lg \frac{16}{\varepsilon} \geq \delta \lg\left(\frac{16e}{\varepsilon} \lg \frac{16}{\varepsilon}\right).$$

This is equivalent to $\left(\dfrac{16}{\varepsilon}\right)^2 \geq \dfrac{16e}{\varepsilon} \lg \dfrac{16}{\varepsilon}$, which is equivalent to $\dfrac{16}{e\,\varepsilon} \geq \lg \dfrac{16}{\varepsilon}$, which is certainly true for $0 < \varepsilon \leq 1$.

This completes the proof of the theorem. ■

20.6. A better bound on the growth function

In this section, we prove Lemma 20.2.2$_{\text{p5}}$. Since the proof is straightforward but tedious, the reader can safely skip reading this section.

Lemma 20.6.1. *For any positive integer n, the following hold.*

(i) $(1 + 1/n)^n \leq e$. *(ii)* $(1 - 1/n)^{n-1} \geq e^{-1}$.

(iii) $n! \geq (n/e)^n$. *(iv)* *For any $k \leq n$, we have* $\left(\dfrac{n}{k}\right)^k \leq \dbinom{n}{k} \leq \left(\dfrac{ne}{k}\right)^k$.

Proof: (i) Indeed, $1 + 1/n \leq \exp(1/n)$, since $1 + x \leq e^x$, for $x \geq 0$. As such $(1 + 1/n)^n \leq \exp(n(1/n)) = e$.

(ii) Rewriting the inequality, we have that we need to prove $\left(\frac{n-1}{n}\right)^{n-1} \geq \frac{1}{e}$. This is equivalent to proving $e \geq \left(\frac{n}{n-1}\right)^{n-1} = \left(1 + \frac{1}{n-1}\right)^{n-1}$, which is our friend from (i).

(iii) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

(iv) Indeed, for any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$, as can be easily verified. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k-1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{1} = \binom{n}{k}.$$

As for the other direction, by (iii), we have $\dbinom{n}{k} \leq \dfrac{n^k}{k!} \leq \dfrac{n^k}{\left(\frac{k}{e}\right)^k} = \left(\dfrac{ne}{k}\right)^k$. ■

Lemma 20.2.2 restated. *For $n \geq 2\delta$ and $\delta \geq 1$, we have* $\left(\dfrac{n}{\delta}\right)^\delta \leq \mathcal{G}_\delta(n) \leq 2\left(\dfrac{ne}{\delta}\right)^\delta$, *where* $\mathcal{G}_\delta(n) = \displaystyle\sum_{i=0}^{\delta} \binom{n}{i}$.

Proof: Note that by Lemma 20.6.1(iv), we have $\mathcal{G}_\delta(n) = \sum_{i=0}^{\delta} \binom{n}{i} \leq 1 + \sum_{i=1}^{\delta} \left(\frac{ne}{i}\right)^i$. This series behaves like a geometric series with constant larger than 2, since

$$\left(\frac{ne}{i}\right)^i \Big/ \left(\frac{ne}{i-1}\right)^{i-1} = \frac{ne}{i}\left(\frac{i-1}{i}\right)^{i-1} = \frac{ne}{i}\left(1-\frac{1}{i}\right)^{i-1} \geq \frac{ne}{i}\frac{1}{e} = \frac{n}{i} \geq \frac{n}{\delta} \geq 2,$$

by Lemma 20.6.1. As such, this series is bounded by twice the largest element in the series, implying the claim. ■

20.7. Bibliographical notes

The exposition of the ε-net and ε-sample theorems is roughly based on Alon and Spencer [AS00] and Komlós *et al.* [KPW92]. In fact, Komlós *et al.* proved a somewhat stronger bound; that is, a random sample of size $(\delta/\varepsilon)\ln(1/\varepsilon)$ is an ε-net with constant probability. For a proof that shows that in general ε-nets cannot be much smaller in the worst case, see [PA95]. The original proof of the ε-net theorem is due to Haussler and Welzl [HW87]. The proof of the ε-sample theorem is due to Vapnik and Chervonenkis [VC71]. The bound in Theorem 20.3.2 can be improved to $O\left(\frac{\delta}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\frac{1}{\varphi}\right)$ [AB99].

An alternative proof of the ε-net theorem proceeds by first computing an $(\varepsilon/4)$-sample of sufficient size, using the ε-sample theorem (Theorem $20.3.2_{\text{p11}}$), and then computing and $\varepsilon/4$-net for this sample using a direct sample of the right size. It is easy to verify the resulting set is an ε-net. Furthermore, using the "naive" argument (see Section 20.3.2.3) then implies that this holds with the right probability, thus implying the ε-net theorem (the resulting constants might be slightly worse). Exercise 20.8.3 deploys similar ideas.

The beautiful alternative proof of both theorems via the usage of discrepancy is due to Chazelle and Matoušek [CM96]. The discrepancy method is a beautiful topic which is quite deep mathematically, and we have just skimmed the thin layer of melted water on top of the tip of the iceberg[8]. Two nice books on the topic are the books by Chazelle [Cha01] and Matoušek [Mat99]. The book by Chazelle [Cha01] is currently available online for free from Chazelle's webpage.

We will revisit discrepancy since in some geometric cases it yields better results than the ε-sample theorem. In particular, the random coloring of Theorem 20.4.1 can be derandomized using conditional probabilities. One can then use it to get an ε-sample/net by applying it repeatedly. A faster algorithm results from a careful implementation of the sketch-and-merge approach. The disappointing feature of all the deterministic constructions of ε-samples/nets is that their running time is exponential in the dimension δ, since the number of ranges is usually exponential in δ.

A similar result to the one derived by Haussler and Welzl [HW87], using a more geometric approach, was done independently by Clarkson at the same time [Cla87], exposing the fact that VC dimension is not necessary if we are interested only in geometric applications. This was later refined by Clarkson [Cla88], leading to a general technique that, in geometric settings, yields stronger results than the ε-net theorem. This technique has numerous applications in discrete and computational geometry and leads to several "proofs from the book" in discrete geometry.

Exercise 20.8.5 is from Anthony and Bartlett [AB99].

20.7.1. Variants and extensions

A natural application of the ε-sample theorem is to use it to estimate the weights of ranges. In particular, given a finite range space $(\mathsf{X}, \mathcal{R})$, we would like to build a data-structure such that we can decide quickly, given a

[8] The iceberg is melting because of global warming; so sorry, climate change.

query range $\mathbf{r}$, what the number of points of X inside $\mathbf{r}$ is. We could always use a sample of size (roughly) $O(\varepsilon^{-2})$ to get an estimate of the weight of a range, using the ε-sample theorem. The error of the estimate of the size $|\mathbf{r} \cap \mathsf{X}|$ is $\leq \varepsilon n$, where $n = |\mathsf{X}|$; namely, the error is additive. The natural question is whether one can get a multiplicative estimate ρ, such that $|\mathbf{r} \cap \mathsf{X}| \leq \rho \leq (1+\varepsilon)|\mathbf{r} \cap \mathsf{X}|$, where $|\mathbf{r} \cap \mathsf{X}|$.

In particular, a subset $A \subset \mathsf{X}$ is a (relative) (ε, p)-***sample*** if for each $\mathbf{r} \in \mathcal{R}$ of weight $\geq pn$,

$$\left| \frac{|\mathbf{r} \cap A|}{|A|} - \frac{|\mathbf{r} \cap \mathsf{X}|}{|\mathsf{X}|} \right| \leq \varepsilon \frac{|\mathbf{r} \cap \mathsf{X}|}{|\mathsf{X}|}.$$

Of course, one can simply generate an εp-sample of size (roughly) $O(1/(\varepsilon p)^2)$ by the ε-sample theorem. This is not very interesting when $p = 1/\sqrt{n}$. Interestingly, the dependency on p can be improved.

Theorem 20.7.1 ([LLS01]). *Let $(X, \mathcal{R})$ be a range space with shattering dimension d, where $|X| = n$, and let $0 < \varepsilon < 1$ and $0 < p < 1$ be given parameters. Then, consider a random sample $A \subseteq X$ of size $\frac{c}{\varepsilon^2 p}\left(d \log \frac{1}{p} + \log \frac{1}{\varphi}\right)$, where c is a constant. Then, it holds that for each range $\mathbf{r} \in \mathcal{R}$ of at least pn points, we have*

$$\left| \frac{|\mathbf{r} \cap A|}{|A|} - \frac{|\mathbf{r} \cap X|}{|X|} \right| \leq \varepsilon \frac{|\mathbf{r} \cap X|}{|X|}.$$

In other words, A is a (p, ε)-sample for $(X, \mathcal{R})$. The probability of success is $\geq 1 - \varphi$.

A similar result is achievable by using discrepancy; see Exercise 20.8.7.

20.8. Exercises

Exercise 20.8.1 (Compute clustering radius). Let C and P be two given sets of points in the plane, such that $k = |C|$ and $n = |P|$. Let $r = \max_{p \in \mathsf{P}} \min_{c \in C} \|c - p\|$ be the *covering radius* of P by C (i.e., if we place a disk of radius r centered at each point of C, all those disks cover the points of P).

(A) Give an $O(n + k \log n)$ expected time algorithm that outputs a number α, such that $r \leq \alpha \leq 10r$.

(B) For $\varepsilon > 0$ a prescribed parameter, give an $O(n + k\varepsilon^{-2} \log n)$ expected time algorithm that outputs a number α, such that $r \leq \alpha \leq (1+\varepsilon)r$.

Exercise 20.8.2 (Some calculus required). Prove Lemma 20.2.5.

Exercise 20.8.3 (A direct proof of the ε-sample theorem). For the case that the given range space is finite, one can prove the ε-sample theorem (Theorem $20.3.2_{\text{p11}}$) directly. So, we are given a range space $\mathsf{S} = (\mathsf{x}, \mathcal{R})$ with VC dimension δ, where x is a finite set.

(A) Show that there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{\log|\mathsf{x}|}{\varepsilon}\right)$ by extracting an $\varepsilon/3$-sample from an $\varepsilon/9$-sample of the original space (i.e., apply Lemma 20.3.6 twice and use Lemma 20.4.3).

(B) Show that for any k, there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{\log^{(k)}|\mathsf{x}|}{\varepsilon}\right)$.

(C) Show that there exists an ε-sample of S of size $O\left(\delta\varepsilon^{-2} \log \frac{1}{\varepsilon}\right)$.

Exercise 20.8.4 (Sauer's lemma is tight). Show that Sauer's lemma (Lemma 20.2.1) is tight. Specifically, provide a finite range space that has the number of ranges as claimed by Lemma 20.2.1.

Exercise 20.8.5 (Flip and flop). (A) Let $b_1, \ldots, b_{2m}$ be m binary bits. Let Ψ be the set of all permutations of $1, \ldots, 2m$, such that for any $\sigma \in \Psi$, we have $\sigma(i) = i$ or $\sigma(i) = m + i$, for $1 \le i \le m$, and similarly, $\sigma(m + i) = i$ or $\sigma(m + i) = m + i$. Namely, $\sigma \in \Psi$ either leaves the pair $i, i + m$ in their positions or it exchanges them, for $1 \le i \le m$. As such $|\Psi| = 2^m$.
Prove that for a random $\sigma \in \Psi$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^{m} b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^{m} b_{\sigma(i+m)}}{m}\right| \ge \varepsilon\right] \le 2e^{-\varepsilon^2 m/2}.$$

(B) Let Ψ' be the set of all permutations of $1, \ldots, 2m$. Prove that for a random $\sigma \in \Psi'$, we have

$$\mathbf{Pr}\left[\left|\frac{\sum_{i=1}^{m} b_{\sigma(i)}}{m} - \frac{\sum_{i=1}^{m} b_{\sigma(i+m)}}{m}\right| \ge \varepsilon\right] \le 2e^{-C\varepsilon^2 m/2},$$

where C is an appropriate constant. *[Use (A), but be careful.]*
(C) Prove Theorem 20.3.2 using (B).

Exercise 20.8.6 (Sketch and merge). Assume that you are given a deterministic algorithm that can compute the discrepancy of Theorem 20.4.1 in $O(nm)$ time, where n is the size of the ground set and m is the number of induced ranges. We are assuming that the VC dimension δ of the given range space is small and that the algorithm input is only the ground set X (i.e., the algorithm can figure out on its own what the relevant ranges are).
(A) For a prespecified $\varepsilon > 0$, using the ideas described in Section 20.4.1.1, show how to compute a small ε-sample of X quickly. The running time of your algorithm should be (roughly) $O\left(n/\varepsilon^{O(\delta)}\text{polylog}\right)$. What is the exact bound on the running time of your algorithm?
(B) One can slightly improve the running of the above algorithm by more aggressively sketching the sets used. That is, one can add additional sketch layers in the tree. Show how by using such an approach one can improve the running time of the above algorithm by a logarithmic factor.

Exercise 20.8.7 (Building relative approximations). Prove the following theorem using discrepancy.

> **Theorem 20.8.8.** *Let $(X, \mathcal{R})$ be a range space with shattering dimension δ, where $|X| = n$, and let $0 < \varepsilon < 1$ and $0 < p < 1$ be given parameters. Then one can construct a set $N \subseteq X$ of size $O\left(\frac{\delta}{\varepsilon^2 p} \ln \frac{\delta}{\varepsilon p}\right)$, such that, for each range $\mathbf{r} \in \mathcal{R}$ of at least pn points, we have*
>
> $$\left|\frac{|\mathbf{r} \cap N|}{|N|} - \frac{|\mathbf{r} \cap X|}{|X|}\right| \le \varepsilon \frac{|\mathbf{r} \cap X|}{|X|}.$$
>
> *In other words, N is a relative (p, ε)-approximation for $(X, \mathcal{R})$.*

20.9. From previous lectures

Definition 20.9.1 (Convex hull). The *convex hull* of a set $\mathsf{R} \subseteq \mathbb{R}^d$ is the set of all convex combinations of points of R; that is,

$$\mathcal{CH}(\mathsf{R}) = \left\{ \sum_{i=0}^{m} \alpha_i \mathsf{r}_i \;\middle|\; \forall i \;\; \mathsf{r}_i \in \mathsf{R}, \alpha_i \ge 0, \text{ and } \sum_{j=1}^{m} \alpha_i = 1 \right\}.$$

Theorem 20.9.2. *For any $\delta > 0$, we have* $\mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \left(\dfrac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$.

Or in a more simplified form, we have:

$$\delta \le 2e-1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\Big(-\mu\delta^2/4\Big), \tag{20.7}$$

$$\delta > 2e-1 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < 2^{-\mu(1+\delta)}, \tag{20.8}$$

and $$\delta \ge e^2 \qquad \mathbf{Pr}\Big[X > (1+\delta)\mu\Big] < \exp\left(-\frac{\mu\delta\ln\delta}{2}\right). \tag{20.9}$$

Bibliography

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge, 1999.

[AS00] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley InterScience, 2nd edition, 2000.

[Cha01] B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, New York, 2001.

[Cla87] K. L. Clarkson. New applications of random sampling in computational geometry. *Discrete Comput. Geom.*, 2:195–222, 1987.

[Cla88] K. L. Clarkson. Applications of random sampling in computational geometry, II. In *Proc. 4th Annu. Sympos. Comput. Geom.* (SoCG), pages 1–11, New York, NY, USA, 1988. ACM.

[CM96] B. Chazelle and J. Matoušek. On linear-time deterministic algorithms for optimization problems in fixed dimension. *J. Algorithms*, 21:579–597, 1996.

[Har11] S. Har-Peled. *Geometric Approximation Algorithms*, volume 173 of *Mathematical Surveys and Monographs*. Amer. Math. Soc., Boston, MA, USA, 2011.

[HW87] D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[KPW92] J. Komlós, J. Pach, and G. Woeginger. Almost tight bounds for ε-nets. *Discrete Comput. Geom.*, 7:163–173, 1992.

[LLS01] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *J. Comput. Syst. Sci.*, 62(3):516–527, 2001.

[Mat99] J. Matoušek. *Geometric Discrepancy*. Springer, 1999.

[PA95] J. Pach and P. K. Agarwal. *Combinatorial Geometry*. John Wiley & Sons, 1995.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

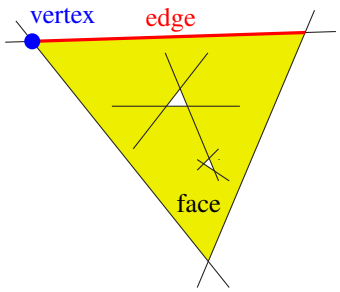
Chapter 21

Sampling and the Moments Technique

By Sariel Har-Peled, December 30, 2015[①]

> Sun and rain and bush had made the site look old, like the site of a dead civilization. The ruins, spreading over so many acres, seemed to speak of a final catastrophe. But the civilization wasn't dead. It was the civilization I existed in and in fact was still working towards. And that could make for an odd feeling: to be among the ruins was to have your time-sense unsettled. You felt like a ghost, not from the past, but from the future. You felt that your life and ambition had already been lived out for you and you were looking at the relics of that life. You were in a place where the future had come and gone.
>
> – A bend in the river, V. S. Naipaul.

21.1. Vertical decomposition



Given a set S of n segments in the plane, its *arrangement*, denoted by $\mathcal{A}(\mathsf{S})$, is the decomposition of the plane into faces, edges, and vertices. The *vertices* of $\mathcal{A}(\mathsf{S})$ are the endpoints and the intersection points of the segments of S, the ***edges*** are the maximal connected portions of the segments not containing any vertex, and the ***faces*** are the connected components of the complement of the union of the segments of S. These definitions are depicted on the right.

For numerical reasons (and also conceptually), a symbolic representation would be better than a numerical one. Thus, an intersection vertex would be represented by two pointers to the segments that their intersection is this vertex. Similarly, an edge would be represented as a pointer to the segment that contains it, and two pointers to the vertices forming its endpoints.

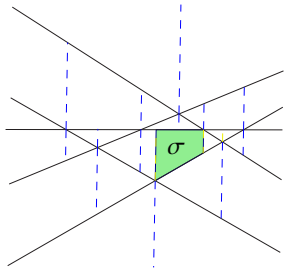
Naturally, we are assuming here that we have geometric primitives that can resolve any decision problem of interest that involve a few geometric entities. For example, for a given segment s and a point p, we would be interested in deciding if p lies vertically below s. From a theoretical point of view, all these primitives require a constant amount of computation, and are "easy". In the real world, numerical issues and degeneracies make implementing these primitives surprisingly challenging. We are going to ignore this major headache here, but the reader should be aware of it.

We will be interested in computing the arrangement $\mathcal{A}(\mathsf{S})$ and a representation of it that makes it easy to manipulate. In particular, we would like to be able to quickly resolve questions of the type (i) are two points in

the same face?, (ii) can one traverse from one point to the other without crossing any segment?, etc. The naive representation of each face as polygons (potentially with holes) is not conducive to carrying out such tasks, since a polygon might be arbitrarily complicated. Instead, we will prefer to break the arrangement into smaller canonical tiles.

To this end, a *vertical trapezoid* is a quadrangle with two vertical sides. The breaking of the faces into such trapezoids is the *vertical decomposition* of the arrangement $\mathcal{A}(\mathsf{S})$.

Formally, for a subset $\mathsf{R} \subseteq \mathsf{S}$, let $\mathcal{A}^{|}(\mathsf{R})$ denote the *vertical decomposition* of the plane formed by the arrangement $\mathcal{A}(\mathsf{R})$ of the segments of R. This is the partition of the plane into interior disjoint vertical trapezoids formed by erecting vertical walls through each vertex of $\mathcal{A}^{|}(\mathsf{R})$. Formally, a ***vertex*** of $\mathcal{A}^{|}(\mathsf{R})$ is either an endpoint of a segment of R or an intersection point of two of its segments. From each such vertex we shoot up (similarly, down) a vertical ray till it hits a segment of R or it continues all the way to infinity. See the figure on the right.



Note that a vertical trapezoid is defined by at most four segments: two segments defining its ceiling and floor and two segments defining the two intersection points that induce the two vertical walls on its boundary. Of course, a vertical trapezoid might be degenerate and thus be defined by fewer segments (i.e., an unbounded vertical trapezoid or a triangle with a vertical segment as one of its sides).

Vertical decomposition breaks the faces of the arrangement that might be arbitrarily complicated into entities (i.e., vertical trapezoids) of constant complexity. This makes handling arrangements (decomposed into vertical trapezoid) much easier computationally.

In the following, we assume that the n segments of S have k pairwise intersection points overall, and we want to compute the arrangement $\mathcal{A} = \mathcal{A}(\mathsf{S})$; namely, compute the edges, vertices, and faces of $\mathcal{A}(\mathsf{S})$. One possible way is the following: Compute a random permutation of the segments of S: $\mathsf{S} = \langle \mathsf{s}_1, \ldots, \mathsf{s}_n \rangle$. Let $\mathsf{S}_i = \langle \mathsf{s}_1, \ldots, \mathsf{s}_i \rangle$ be the prefix of length i of S. Compute $\mathcal{A}^{|}(\mathsf{S}_i)$ from $\mathcal{A}^{|}(\mathsf{S}_{i-1})$, for $i = 1, \ldots, n$. Clearly, $\mathcal{A}^{|}(\mathsf{S}) = \mathcal{A}^{|}(\mathsf{S}_n)$, and we can extract $\mathcal{A}(\mathsf{S})$ from it. Namely, in the ith iteration, we insert the segment s_i into the arrangement $\mathcal{A}^{|}(\mathsf{S}_{i-1})$.

This technique of building the arrangement by inserting the segments one by one is called ***randomized incremental construction***.

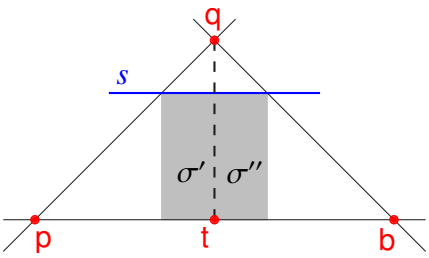
Who need these pesky arrangements anyway? The reader might wonder who needs arrangements? As a concrete examples, consider a situation where you are give several maps of a city containing different layers of information (i.e., streets map, sewer map, electric lines map, train lines map, etc). We would like to compute the overlay map formed by putting all these maps on top of each other. For example, we might be interested in figuring out if there are any buildings lying on a planned train line, etc.

More generally, think about a set of general constraints in $\mathbb{R}^d$. Each constraint is bounded by a surface, or a patch of a surface. The decomposition of $\mathbb{R}^d$ formed by the arrangement of these surfaces gives us a description of the parametric space in a way that is algorithmically useful. For example, finding if there is a point inside all the constraints, when all the constraints are induced by linear inequalities, is linear programming. Namely, arrangements are a useful way to think about any parametric space partitioned by various constraints.

21.1.1. Randomized incremental construction (RIC)

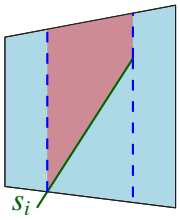
Imagine that we had computed the arrangement $\mathcal{B}_{i-1} = \mathcal{A}^{|}(\mathsf{S}_{i-1})$. In the ith iteration we compute $\mathcal{B}_i$ by inserting s_i into the arrangement $\mathcal{B}_{i-1}$. This involves splitting some trapezoids (and merging some others).

As a concrete example, consider the figure on the right. Here we insert s in the arrangement. To this end we split the "vertical trapezoids" $\triangle\mathsf{pqt}$ and $\triangle\mathsf{bqt}$, each into three trapezoids. The two trapezoids σ' and σ'' now need to be merged together to form the new trapezoid which appears in the vertical decomposition of the new arrangement. (Note that the figure does not show all the trapezoids in the vertical decomposition.)



To facilitate this, we need to compute the trapezoids of $\mathcal{B}_{i-1}$ that intersect s_i. This is done by maintaining a *conflict graph*. Each trapezoid $\sigma \in \mathcal{A}^{|}(\mathsf{S}_{i-1})$ maintains a *conflict list* $\mathrm{cl}(\sigma)$ of *all* the segments of S that intersect its interior. In particular, the conflict list of σ cannot contain any segment of S_{i-1}, and as such it contains only the segments of $\mathsf{S} \setminus \mathsf{S}_{i-1}$ that intersect its interior. We also maintain a similar structure for each segment, listing all the trapezoids of $\mathcal{A}^{|}(\mathsf{S}_{i-1})$ that it currently intersects (in its interior). We maintain those lists with cross pointers, so that given an entry (σ, s) in the conflict list of σ, we can find the entry (s, σ) in the conflict list of s in constant time.

Thus, given s_i, we know what trapezoids need to be split (i.e., all the trapezoids in $\mathrm{cl}(\mathsf{s}_i)$). Splitting a trapezoid σ by a segment s_i is the operation of computing a set of (at most) four trapezoids that cover σ and have s_i on their boundary. We compute those new trapezoids, and next we need to compute the conflict lists of the new trapezoids. This can be easily done by taking the conflict list of a trapezoid $\sigma \in \mathrm{cl}(\mathsf{s}_i)$ and distributing its segments among the $O(1)$ new trapezoids that cover σ. Using careful implementation, this requires a linear time in the size of the conflict list of σ.



Note that only trapezoids that intersect s_i in their interior get split. Also, we need to update the conflict lists for the segments (that were not inserted yet).

We next sketch the low-level details involved in maintaining these conflict lists. For a segment s that intersects the interior of a trapezoid σ, we maintain the pair (s, σ). For every trapezoid σ, in the current vertical decomposition, we maintain a doubly linked list of all such pairs that contain σ. Similarly, for each segment s we maintain the doubly linked list of all such pairs that contain s. Finally, each such pair contains two pointers to the location in the two respective lists where the pair is being stored.

It is now straightforward to verify that using this data-structure we can implement the required operations in linear time in the size of the relevant conflict lists.

In the above description, we ignored the need to merge adjacent trapezoids if they have identical floor and ceiling – this can be done by a somewhat straightforward and tedious implementation of the vertical decomposition data-structure, by providing pointers between adjacent vertical trapezoids and maintaining the conflict list sorted (or by using hashing) so that merge operations can be done quickly. In any case, this can be done in linear time in the input/output size involved, as can be verified.

21.1.1.1. Analysis

Claim 21.1.1. *The (amortized) running time of constructing $\mathcal{B}_i$ from $\mathcal{B}_{i-1}$ is proportional to the size of the conflict lists of the vertical trapezoids in $\mathcal{B}_i \setminus \mathcal{B}_{i-1}$ (and the number of such new trapezoids).*

Proof: Observe that we can charge all the work involved in the ith iteration to either the conflict lists of the newly created trapezoids or the deleted conflict lists. Clearly, the running time of the algorithm in the ith iteration is linear in the total size of these conflict lists. Observe that every conflict gets charged twice – when it is being created and when it is being deleted. As such, the (amortized) running time in the ith iteration is proportional to the total length of the newly created conflict lists. ■

Thus, to bound the running time of the algorithm, it is enough to bound the expected size of the destroyed

conflict lists in ith iteration (and sum this bound on the n iterations carried out by the algorithm). Or alternatively, bound the expected size of the conflict lists created in the ith iteration.

Lemma 21.1.2. *Let S be a set of n segments (in general position[2]) with k intersection points. Let S_i be the first i segments in a random permutation of S. The expected size of $\mathcal{B}_i = \mathcal{A}^{\mid}(\mathsf{S}_i)$, denoted by $\tau(i)$ (i.e., the number of trapezoids in $\mathcal{B}_i$), is $O\left(i + k(i/n)^2\right)$.*

Proof: Consider[3] an intersection point $\mathsf{p} = s \cap s'$, where $s, s' \in \mathsf{S}$. The probability that p is present in $\mathcal{A}^{\mid}(\mathsf{S}_i)$ is equivalent to the probability that both s and s' are in S_i. This probability is

$$\alpha = \frac{\binom{n-2}{i-2}}{\binom{n}{i}} = \frac{(n-2)!}{(i-2)!\,(n-i)!} \cdot \frac{i!\,(n-i)!}{n!} = \frac{i(i-1)}{n(n-1)}.$$

For each intersection point p in $\mathcal{A}(\mathsf{S})$ define an indicator variable X_{p}, which is 1 if the two segments defining p are in the random sample S_i and 0 otherwise. We have that $\mathbf{E}\left[X_{\mathsf{p}}\right] = \alpha$, and as such, by linearity of expectation, the expected number of intersection points in the arrangement $\mathcal{A}(\mathsf{S}_i)$ is

$$\mathbf{E}\left[\sum_{\mathsf{p}\in V} X_{\mathsf{p}}\right] = \sum_{\mathsf{p}\in V} \mathbf{E}\left[X_{\mathsf{p}}\right] = \sum_{\mathsf{p}\in V} \alpha = k\alpha,$$

where V is the set of k intersection points of $\mathcal{A}(\mathsf{S})$. Also, every endpoint of a segment of S_i contributed its two endpoints to the arrangement $\mathcal{A}(\mathsf{S}_i)$. Thus, we have that the expected number of vertices in $\mathcal{A}(\mathsf{S}_i)$ is

$$2i + \frac{i(i-1)}{n(n-1)}k.$$

Now, the number of trapezoids in $\mathcal{A}^{\mid}(\mathsf{S}_i)$ is proportional to the number of vertices of $\mathcal{A}(\mathsf{S}_i)$, which implies the claim. ■

21.1.2. Backward analysis

In the following, we would like to consider the total amount of work involved in the ith iteration of the algorithm. The way to analyze these iterations is (conceptually) to run the algorithm for the first i iterations and then run "backward" the last iteration.

So, imagine that the overall size of the conflict lists of the trapezoids of $\mathcal{B}_i$ is W_i and the total size of the conflict lists created only in the ith iteration is C_i.

We are interested in bounding the expected size of C_i, since this is (essentially) the amount of work done by the algorithm in this iteration. Observe that the structure of $\mathcal{B}_i$ is defined independently of the permutation S_i and depends only on the (unordered) set $\mathsf{S}_i = \{\mathsf{s}_1, \ldots, \mathsf{s}_i\}$. So, fix S_i. What is the probability that s_i is a specific

[2] In this case, no two intersection points of input segments are the same, no two intersection points (or vertices) have the same x-coordinate, no two segments lie on the same line, etc. Making the geometric algorithm work correctly for all degenerate inputs is a huge task that can usually be handled by tedious and careful implementation. Thus, we will always assume general position of the input. In other words, in theory all geometric inputs are inherently good, while in practice they are all evil (as anybody who tried to implement geometric algorithms can testify). The reader is encouraged not to use this to draw any conclusions on the human condition.

[3] The proof is provided in excruciating detail to get the reader used to this kind of argumentation. I would apologize for this pain, but it is a minor trifle, not to be mentioned, when compared to the other offenses in this book.

segment s of S_i? Clearly, this is $1/i$ since this is the probability of s being the last element in a permutation of the i elements of S_i (i.e., we consider a random permutation of S_i).

Now, consider a trapezoid $\sigma \in \mathcal{B}_i$. If σ was created in the ith iteration, then s_i must be one of the (at most four) segments that define it. Indeed, if s_i is not one of the segments that define σ, then σ existed in the vertical decomposition before s_i was inserted. Since $\mathcal{B}_i$ is independent of the internal ordering of S_i, it follows that $\mathbf{Pr}[\sigma \in (\mathcal{B}_i \setminus \mathcal{B}_{i-1})] \leq 4/i$. In particular, the overall size of the conflict lists in the end of the ith iteration is

$$W_i = \sum_{\sigma \in \mathcal{B}_i} |\operatorname{cl}(\sigma)|.$$

As such, the expected overall size of the conflict lists created in the ith iteration is

$$\mathbf{E}\Big[C_i \,\Big|\, \mathcal{B}_i\Big] \leq \sum_{\sigma \in \mathcal{B}_i} \frac{4}{i} |\operatorname{cl}(\sigma)| \leq \frac{4}{i} W_i.$$

By Lemma 21.1.2, the expected size of $\mathcal{B}_i$ is $O\big(i + ki^2/n^2\big)$. Let us guess (for the time being) that on average the size of the conflict list of a trapezoid of $\mathcal{B}_i$ is about $O(n/i)$. In particular, assume that we know that

$$\mathbf{E}\big[W_i\big] = O\left(\left(i + \frac{i^2}{n^2}k\right)\frac{n}{i}\right) = O\left(n + k\frac{i}{n}\right),$$

by Lemma 21.1.2, implying

$$\mathbf{E}\big[C_i\big] = \mathbf{E}\Big[\mathbf{E}\big[C_i \,\big|\, \mathcal{B}_i\big]\Big] \leq \mathbf{E}\left[\frac{4}{i}W_i\right] = \frac{4}{i}\,\mathbf{E}\big[W_i\big] = O\left(\frac{4}{i}\left(n + \frac{ki}{n}\right)\right) = O\left(\frac{n}{i} + \frac{k}{n}\right), \tag{21.1}$$

using Lemma 21.7.2$_{\text{p16}}$. In particular, the expected (amortized) amount of work in the ith iteration is proportional to $\mathbf{E}[C_i]$. Thus, the overall expected running time of the algorithm is

$$\mathbf{E}\left[\sum_{i=1}^{n} C_i\right] = \sum_{i=1}^{n} O\left(\frac{n}{i} + \frac{k}{n}\right) = O\big(n \log n + k\big).$$

Theorem 21.1.3. *Given a set S of n segments in the plane with k intersections, one can compute the vertical decomposition of $\mathcal{A}(\mathsf{S})$ in expected $O(n \log n + k)$ time.*

Intuition and discussion. What remains to be seen is how we came up with the guess that the average size of a conflict list of a trapezoid of $\mathcal{B}_i$ is about $O(n/i)$. Note that using ε-nets implies that the bound $O((n/i)\log i)$ holds with constant probability (see Theorem 21.7.1$_{\text{p15}}$) for all trapezoids in this arrangement. As such, this result is only slightly surprising. To prove this, we present in the next section a "strengthening" of ε-nets to geometric settings.

To get some intuition on how we came up with this guess, consider a set P of n points on the line and a random sample R of i points from P. Let $\widehat{\mathcal{I}}$ be the partition of the real line into (maximal) open intervals by the endpoints of R, such that these intervals do not contain points of R in their interior.

Consider an interval (i.e., a one-dimensional trapezoid) of $\widehat{\mathcal{I}}$. It is intuitively clear that this interval (in expectation) would contain $O(n/i)$ points. Indeed, fix a point x on the real line, and imagine that we pick each point with probability i/n to be in the random sample. The random variable which is the number of points of P we have to scan starting from x and going to the right of x till we "hit" a point that is in the random sample behaves like a geometric variable with probability i/n, and as such its expected value is n/i. The same argument

works if we scan P to the left of x. We conclude that the number of points of P in the interval of $\widehat{\mathcal{I}}$ that contains x but does not contain any point of R is $O(n/i)$ in expectation.

Of course, the vertical decomposition case is more involved, as each vertical trapezoid is defined by four input segments. Furthermore, the number of possible vertical trapezoids is larger. Instead of proving the required result for this special case, we will prove a more general result which can be applied in a lot of other settings.

21.2. General settings

21.2.1. Notation

Let S be a set of objects. For a subset $\mathsf{R} \subseteq \mathsf{S}$, we define a collection of 'regions' called $\mathcal{F}(\mathsf{R})$. For the case of vertical decomposition of segments (i.e., Theorem 21.1.3), the objects are segments, the regions are trapezoids, and $\mathcal{F}(\mathsf{R})$ is the set of vertical trapezoids in $\mathcal{A}^{|}(\mathsf{R})$. Let

$$\mathcal{T} = \mathcal{T}(\mathsf{S}) = \bigcup_{\mathsf{R} \subseteq \mathsf{S}} \mathcal{F}(\mathsf{R})$$

denote the set of ***all possible regions*** defined by subsets of S.

In the vertical trapezoids case, the set $\mathcal{T}$ is the set of all vertical trapezoids that can be defined by any subset of the given input segments.

We associate two subsets $D(\sigma), K(\sigma) \subseteq \mathsf{S}$ with each region $\sigma \in \mathcal{T}$.

The ***defining set*** $D(\sigma)$ of σ is the subset of S defining the region σ (the precise requirements from this set are specified in the axioms below). We assume that for every $\sigma \in \mathcal{T}$, $|D(\sigma)| \leq d$ for a (small) constant d. The constant d is sometime referred to as the *combinatorial dimension*. In the case of Theorem 21.1.3, each trapezoid σ is defined by at most four segments (or lines) of S that define the region covered by the trapezoid σ, and this set of segments is $D(\sigma)$. See Figure 21.1.

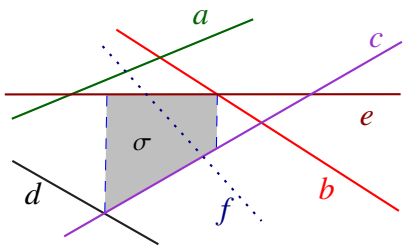


Figure 21.1: $D(\sigma) = \{b, c, d, e\}$ and $K(\sigma) = \{f\}$.

The *stopping set* $K(\sigma)$ of σ is the set of objects of S such that including any object of $K(\sigma)$ in R prevents σ from appearing in $\mathcal{F}(\mathsf{R})$. In many applications $K(\sigma)$ is just the set of objects intersecting the cell σ; this is also the case in Theorem 21.1.3, where $K(\sigma)$ is the set of segments of S intersecting the interior of the trapezoid σ (see Figure 21.1). Thus, the stopping set of a region σ, in many cases, is just the conflict list of this region, when it is being created by an RIC algorithm. The ***weight*** of σ is $\omega(\sigma) = |K(\sigma)|$.

Axioms. Let $\mathsf{S}, \mathcal{F}(\mathsf{R}), D(\sigma)$, and $K(\sigma)$ be such that for any subset $\mathsf{R} \subseteq \mathsf{S}$, the set $\mathcal{F}(\mathsf{R})$ satisfies the following axioms:

(i) For any $\sigma \in \mathcal{F}(\mathsf{R})$, we have $D(\sigma) \subseteq \mathsf{R}$ and $\mathsf{R} \cap K(\sigma) = \emptyset$.
(ii) If $D(\sigma) \subseteq \mathsf{R}$ and $K(\sigma) \cap \mathsf{R} = \emptyset$, then $\sigma \in \mathcal{F}(\mathsf{R})$.

21.2.1.1. Examples of the general framework

(A) **Vertical decomposition.** Discussed above.
(B) **Points on a line.** Let S be a set of n points on the real line. For a set $\mathsf{R} \subseteq \mathsf{S}$, let $\mathcal{F}(\mathsf{R})$ be the set of atomic intervals of the real lines formed by R; that is, the partition of the real line into maximal connected sets (i.e., intervals and rays) that do not contain a point of R in their interior.

Clearly, in this case, an interval $\mathsf{I} \in \mathcal{F}(\mathsf{R})$ the defining set of I (i.e., $D(\mathsf{I})$) is the set containing the (one or two) endpoints of I in R. The stopping set of an I is the set $K(\mathsf{I})$, which is the set of all points of S contained in I.

(C) **Vertices of the convex-hull in 2d.** Consider a set S of n points in the plane. A vertex on the convex hull is defined by the point defining the vertex, and the two edges before and after it on the convex hull. To this end, a *certified vertex* of the convex hull (say this vertex is q) is a triplet $(\mathsf{p}, \mathsf{q}, \mathsf{r})$, such that p, q and r are consecutive vertices of $\mathcal{CH}(\mathsf{S})$ (say, in clockwise order). Observe, that computing the convex-hull of S is equivalent to computing the set of certified vertices of S.
For a set $\mathsf{R} \subseteq \mathsf{S}$, let $\mathcal{F}(\mathsf{R})$ denote the set of certified vertices of R (i.e., this is equivalent to the set of vertices of the convex-hull of R. For a certified vertex $\sigma \in \mathcal{F}(\mathsf{R})$, its defining set is the set of three vertices $\mathsf{p}, \mathsf{q}, \mathsf{r}$ that (surprise, surprise) define it. Its stopping set, is the set of all points in S, that either on the "wrong" side of the line spanning pq, or on the "wrong" side of the line spanning qr. Equivalently, $K(\sigma)$ is the set of all points $\mathsf{t} \in \mathsf{S} \setminus \mathsf{R}$, such that the convex-hull of $\mathsf{p}, \mathsf{q}, \mathsf{r}$, and t does not form a convex quadrilateral.

(D) **Edges of the convex-hull in 3d.**
Let S be a set of points in three dimensions. An edge e of the convex-hull of a set $\mathsf{R} \subseteq ObjSet$ of points in $\mathbb{R}^3$ is defined by two vertices of S, and it can be certified as being on the convex hull $\mathcal{CH}(\mathsf{R})$, by the two faces f, f' adjacent to e. If all the points of R are on the "right" side of both these two faces then e is an edge of the convex hull of R. Computing all the certified edges of S is equivalent to computing the convex-hull of S.
In the following, assume that each face of any convex-hull of a subset of points of S is a triangle. As such, a face of the convex-hull would be defined by three points. Formally, the *butterfly* of an edge e of $\mathcal{CH}(\mathsf{R})$ is $(\mathsf{e}, \mathsf{p}, \mathsf{q})$, where $pnt, \mathsf{q} \in \mathsf{R}$, and such that all the points of R are on the same side as q of the plane spanned by e and p (we have symmetric condition requiring that all the points of S are on the same as p of the plane spanned by e and q).
For a set $\mathsf{R} \subseteq \mathsf{P}$, let $\mathcal{F}(\mathsf{R})$ be its set of butterflies. Clearly, computing all the butterflies of S (i.e., $\mathcal{F}(\mathsf{S})$) is equivalent to computing the convex-hull of S.
For a butterfly $\sigma = (\mathsf{e}, \mathsf{p}, \mathsf{q}) \in \mathcal{F}(\mathsf{R})$ its defining set (i.e., $D(\sigma)$) is a set of four points (i.e., the two points defining its edge e, and the to additional vertices defining the two faces $Face$ and f' adjacent to it). Its stopping set $K(\sigma)$, is the set of all the points of $\mathsf{S} \setminus \mathsf{R}$ that of different sides of the plane spanned by e and p (resp. e and q) than q (resp. p) [here, the stopping set is the union of these two sets].

(E) **Delaunay triangles in 2d.**
For a set of S of n points in the plane. Consider a subset $\mathsf{R} \subseteq \mathsf{S}$. A ***Delaunay circle*** of R is a disc D that has three points $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$ of R on its boundary, and no points of R in its interior. Naturally, these three points define a ***Delaunay triangle*** $\triangle = \triangle \mathsf{p}_1\mathsf{p}_2\mathsf{p}_3$. The defining set is $D(\triangle) = \{\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3\}$, and the stopping set $K(\triangle)$ is the set of all points in S that are contained in the interior of the disk D.

21.2.2. Analysis

In the following, S is a set of n objects complying with axioms (i) and (ii).

The challenge. What makes the analysis not easy is that there are dependencies between the defining set of a region and its stopping set (i.e., conflict list). In particular, we have the following difficulties

(A) The defining set might be of different sizes depending on the region σ being considered.

(B) Even if all the regions have a defining set of the same size d (say, 4 as in the case of vertical trapezoids), it is not true that every d objects define a valid region. For example, for the case of segments, the four segments might be vertically separated from each other (i.e., think about them as being four disjoint intervals on the real line), and they do not define a vertical trapezoid together. Thus, our analysis is

going to be a bit loopy loop – we are going to assume we know how many regions exists (in expectation) for a random sample of certain size, and use this to derive the desired bounds.

21.2.2.1. On the probability of a region to be created

Inherently, to analyze a randomized algorithm using this framework, we will be interested in the probability that a certain region would be created. Thus, let

$$\rho_{r,n}(d,k)$$

denote the probability that a region $\sigma \in \mathcal{T}$ appears in $\mathcal{F}(\mathsf{R})$, where its defining set is of size d, its stopping set is of size k, R is a random sample of size r from a set S, and $n = |\mathsf{S}|$. Specifically, σ is a *feasible* region that might be created by an algorithm computing $\mathcal{F}(\mathsf{R})$.

The sampling model. For describing algorithms it is usually easier to work with samples created by picking a subset of a certain size (without repetition) from the original set of objects. Usually, in the algorithmic applications this would be done by randomly permuting the objects and interpreting a prefix of this permutation as a random sample. Insisting on analyzing this framework in the "right" sampling model creates some non-trivial technical pain.

Lemma 21.2.1. *We have that* $\rho_{r,n}(d,k) \approx \left(1-\frac{r}{n}\right)^k \left(\frac{r}{n}\right)^d$. *Formally,*

$$\frac{1}{2^{2d}}\left(1-4\cdot\frac{r}{n}\right)^k\left(\frac{r}{n}\right)^d \;\le\; \rho_{r,n}(d,k) \;\le\; 2^{2d}\left(1-\frac{1}{2}\cdot\frac{r}{n}\right)^k\left(\frac{r}{n}\right)^d. \tag{21.2}$$

Proof: Let σ be the region under consideration that is defined by d objects and having k stoppers (i.e., $k = K(\sigma)$). We are interested in the probability of σ being created when taking a sample of size r (without repetition) from a set S of n objects. Clearly, this probability is $\rho_{r,n}(d,k) = \binom{n-d-k}{r-d} / \binom{n}{r}$, as we have to pick the d defining objects into the random sample and avoid picking any of the k stoppers. A tedious but careful calculation, delegated to Section 21.4, implies Eq. (21.2).

Instead, here is an elegant argument for why this estimate is correct in a slightly different sampling model. We pick every element of S into the sample R with probability r/n, and this is done independently for each object. In expectation, the random sample is of size r, and clearly the probability that σ is created is the probability that we pick its d defining objects (that is, $(r/n)^d$) multiplied by the probability that we did not pick any of its k stoppers (that is, $(1-r/n)^k$). ■

Remark 21.2.2. The bounds of Eq. (21.2) hold only when r, d, and k are in certain (reasonable) ranges. For the sake of simplicity of exposition we ignore this minor issue. With care, all our arguments work when one pays careful attention to this minor technicality.

21.2.2.2. On exponential decay

For any natural number r and a number $t > 0$, consider R to be a random sample of size r from S without repetition. We will refer to a region $\sigma \in \mathcal{F}(\mathsf{R})$ as being ***t*-heavy** if $\omega(\sigma) \ge t \cdot \frac{n}{r}$. Let $\mathcal{F}_{\ge t}(\mathsf{R})$ denote all the t-heavy regions of $\mathcal{F}(\mathsf{R})$.[④]

[④] These are the regions that are at least t times overweight. Speak about an obesity problem.

Intuitively, and somewhat incorrectly, we expect the average weight of a region of $\mathcal{F}(\mathsf{R})$ to be roughly n/r. We thus expect the size of this set to drop fast as t increases. Indeed, Lemma 21.2.1 tells us that a trapezoid of weight $t\,(n/r)$ has probability

$$\begin{aligned}\rho_{r,n}\Big(d, t\cdot\frac{n}{r}\Big) &\approx \Big(1-\frac{r}{n}\Big)^{t\,(n/r)}\Big(\frac{r}{n}\Big)^d \approx \exp(-t)\cdot\Big(\frac{r}{n}\Big)^d \approx \exp(-t+1)\cdot\Big(1-\frac{r}{n}\Big)^{n/r}\Big(\frac{r}{n}\Big)^d \\ &\approx \exp(-t+1)\cdot\rho_{r,n}(d, n/r)\end{aligned}$$

to be created, since $(1-r/n)^{n/r} \approx 1/e$. Namely, a t-heavy region has exponentially lower probability to be created than a region of weight n/r. We next formalize this argument.

Lemma 21.2.3. *Let $r \le n$ and let t be parameters, such that $1 \le t \le r/d$. Furthermore, let R be a sample of size r, and let R' be a sample of size $r' = \lfloor r/t \rfloor$, both from S. Let $\sigma \in \mathcal{T}$ be a region with weight $\omega(\sigma) \ge t\,(n/r)$. Then, $\mathbf{Pr}\big[\sigma \in \mathcal{F}(\mathsf{R})\big] = O\Big(\exp\Big(-\frac{t}{2}\Big)t^d\,\mathbf{Pr}\big[\sigma \in \mathcal{F}(\mathsf{R}')\big]\Big)$.*

Proof: For the sake of simplicity of exposition, assume that $k = \omega(\sigma) = t\,(n/r)$. By Lemma 21.2.1 (i.e., Eq. (21.2)) we have

$$\begin{aligned}\frac{\mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R})]}{\mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R}')]} = \frac{\rho_{r,n}(d,k)}{\rho_{r',n}(d,k)} &\le \frac{2^{2d}\Big(1-\frac{1}{2}\cdot\frac{r}{n}\Big)^k\Big(\frac{r}{n}\Big)^d}{\frac{1}{2^{2d}}\Big(1-4\frac{r'}{n}\Big)^k\Big(\frac{r'}{n}\Big)^d} \\ &\le 2^{4d}\exp\Big(-\frac{kr}{2n}\Big)\Big(1+8\frac{r'}{n}\Big)^k\Big(\frac{r}{r'}\Big)^d \le 2^{4d}\exp\Big(8\frac{kr'}{n}-\frac{kr}{2n}\Big)\Big(\frac{r}{r'}\Big)^d \\ &= 2^{4d}\exp\Big(8\frac{tn\lfloor r/t\rfloor}{nr}-\frac{tnr}{2nr}\Big)\Big(\frac{r}{\lfloor r/t\rfloor}\Big)^d = O\Big(\exp(-t/2)t^d\Big),\end{aligned}$$

since $1/(1-x) \le 1+2x$ for $x \le 1/2$ and $1+y \le \exp(y)$, for all y. (The constant in the above $O(\cdot)$ depends exponentially on d.) ■

Let

$$\mathbf{E}f(r) = \mathbf{E}[|\mathcal{F}(\mathsf{R})|] \quad \text{and} \quad \mathbf{E}f_{\ge t}(r) = \mathbf{E}[|\mathcal{F}_{\ge t}(\mathsf{R})|]\,,$$

where the expectation is over random subsets $\mathsf{R} \subseteq \mathsf{S}$ of size r. Note that $\mathbf{E}f(r) = \mathbf{E}f_{\ge 0}(r)$ is the expected number of regions created by a random sample of size r. In words, $\mathbf{E}f_{\ge t}(r)$ is the expected number of regions in a structure created by a sample of r random objects, such that these regions have weight which is t times larger than the "expected" weight (i.e., n/r). In the following, we assume that $\mathbf{E}f(r)$ is a monotone increasing function.

Lemma 21.2.4 (The exponential decay lemma). *Given a set S of n objects and parameters $r \le n$ and $1 \le t \le r/d$, where $d = \max_{\sigma\in\mathcal{T}(\mathsf{S})}|D(\sigma)|$, if axioms (i) and (ii) above hold for any subset of S, then*

$$\mathbf{E}f_{\ge t}(r) = O\Big(t^d\exp(-t/2)\,\mathbf{E}f(r)\Big). \tag{21.3}$$

Proof: Let R be a random sample of size r from S and let R' be a random sample of size $r' = \lfloor r/t \rfloor$ from S. Let $H = \bigcup_{X\subseteq\mathsf{S},|X|=r}\mathcal{F}_{\ge t}(X)$ denote the set of all t-heavy regions that might be created by a sample of size r. In the following, the expectation is taken over the content of the random samples R and R'.

For a region σ, let X_σ be the indicator variable that is 1 if and only if $\sigma \in \mathcal{F}(\mathsf{R})$. By linearity of expectation and since $\mathbf{E}[X_\sigma] = \mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R})]$, we have

$$\begin{aligned}
\mathbf{E}\mathfrak{f}_{\geq t}(r) &= \mathbf{E}\Big[|\mathcal{F}_{\geq t}(\mathsf{R})|\Big] = \mathbf{E}\left[\sum_{\sigma\in H} X_\sigma\right] = \sum_{\sigma\in H} \mathbf{E}[X_\sigma] = \sum_{\sigma\in H} \mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R})] \\
&= O\left(t^d \exp(-t/2) \sum_{\sigma\in H} \mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R}')]\right) = O\left(t^d \exp(-t/2) \sum_{\sigma\in \mathcal{T}} \mathbf{Pr}[\sigma \in \mathcal{F}(\mathsf{R}')]\right) \\
&= O\Big(t^d \exp(-t/2)\, \mathbf{E}f(r')\Big) = O\Big(t^d \exp(-t/2)\, \mathbf{E}f(r)\Big),
\end{aligned}$$

by Lemma 21.2.3 and since $\mathbf{E}f(r)$ is a monotone increasing function. ■

21.2.2.3. Bounding the moments

Consider a different randomized algorithm that in a first round samples r objects, $\mathsf{R} \subseteq \mathsf{S}$ (say, segments), computes the arrangement induced by these r objects (i.e., $\mathcal{A}^{|}(\mathsf{R})$), and then inside each region σ it computes the arrangement of the $\omega(\sigma)$ objects intersecting the interior of this region, using an algorithm that takes $O((\omega(\sigma))^c)$ time, where $c > 0$ is some fixed constant. The overall expected running time of this algorithm is

$$\mathbf{E}\left[\sum_{\sigma\in\mathcal{F}(\mathsf{R})} \Big(\omega(\sigma)\Big)^c\right].$$

We are now able to bound this quantity.

Theorem 21.2.5 (Bounded moments theorem). *Let $\mathsf{R} \subseteq \mathsf{S}$ be a random subset of size r. Let $\mathbf{E}f(r) = \mathbf{E}[|\mathcal{F}(\mathsf{R})|]$ and let $c \geq 1$ be an arbitrary constant. Then,*

$$\mathbf{E}\left[\sum_{\sigma\in\mathcal{F}(\mathsf{R})} \Big(\omega(\sigma)\Big)^c\right] = O\Big(\mathbf{E}f(r)\left(\frac{n}{r}\right)^c\Big).$$

Proof: Let $\mathsf{R} \subseteq \mathsf{S}$ be a random sample of size r. Observe that all the regions with weight in the range $\left[(t-1)\frac{n}{r}, t\cdot\frac{n}{r}\right)$ are in the set $\mathcal{F}_{\geq t-1}(\mathsf{R}) \setminus \mathcal{F}_{\geq t}(\mathsf{R})$. As such, we have by Lemma 21.2.4 that

$$\begin{aligned}
\mathbf{E}\left[\sum_{\sigma\in\mathcal{F}(\mathsf{R})} \omega(\sigma)^c\right] &\leq \mathbf{E}\left[\sum_{t\geq 1}\left(t\frac{n}{r}\right)^c \left(|\mathcal{F}_{\geq t-1}(\mathsf{R})| - |\mathcal{F}_{\geq t}(\mathsf{R})|\right)\right] \leq \mathbf{E}\left[\sum_{t\geq 1}\left(t\frac{n}{r}\right)^c |\mathcal{F}_{\geq t-1}(\mathsf{R})|\right] \\
&\leq \left(\frac{n}{r}\right)^c \sum_{t\geq 0} (t+1)^c \cdot \mathbf{E}[|\mathcal{F}_{\geq t}(\mathsf{R})|] \\
&= \left(\frac{n}{r}\right)^c \sum_{t\geq 0} (t+1)^c\, \mathbf{E}\mathfrak{f}_{\geq t}(r) = \left(\frac{n}{r}\right)^c \sum_{t\geq 0} O\Big((t+1)^{c+d} \exp(-t/2)\, \mathbf{E}f(r)\Big) \\
&= O\left(\mathbf{E}f(r)\left(\frac{n}{r}\right)^c \sum_{t\geq 0} (t+1)^{c+d} \exp(-t/2)\right) = O\Big(\mathbf{E}f(r)\left(\frac{n}{r}\right)^c\Big),
\end{aligned}$$

since c and d are both constants. ■

21.3. Applications

21.3.1. Analyzing the RIC algorithm for vertical decomposition

We remind the reader that the input of the algorithm of Section 21.1.2 is a set S of n segments with k intersections, and it uses randomized incremental construction to compute the vertical decomposition of the arrangement $\mathcal{A}(\mathsf{S})$.

Lemma 21.1.2 shows that the number of vertical trapezoids in the randomized incremental construction is in expectation $\mathbf{E}f(i) = O\left(i + k\,(i/n)^2\right)$. Thus, by Theorem 21.2.5 (used with $c = 1$), we have that the total expected size of the conflict lists of the vertical decomposition computed in the ith step is

$$\mathbf{E}[W_i] = \mathbf{E}\left[\sum_{\sigma\in\mathcal{B}_i} \omega(\sigma)\right] = O\left(\mathbf{E}f(i)\frac{n}{i}\right) = O\left(n + k\frac{i}{n}\right).$$

This is the missing piece in the analysis of Section 21.1.2. Indeed, the amortized work in the ith step of the algorithm is $O(W_i/i)$ (see Eq. $(21.1)_{\text{p5}}$), and as such, the expected running time of this algorithm is

$$\mathbf{E}\left[O\left(\sum_{i=1}^{n} \frac{W_i}{i}\right)\right] = O\left(\sum_{i=1}^{n} \frac{1}{i}\left(n + k\frac{i}{n}\right)\right) = O(n\log n + k).$$

This implies Theorem 21.1.3.

21.3.2. Cuttings

Let S be a set of n lines in the plane, and let r be an arbitrary parameter. A $(1/r)$-***cutting*** of S is a partition of the plane into constant complexity regions such that each region intersects at most n/r lines of S. It is natural to try to minimize the number of regions in the cutting, as cuttings are a natural tool for performing "divide and conquer".

Consider the range space having S as its ground set and vertical trapezoids as its ranges (i.e., given a vertical trapezoid σ, its corresponding range is the set of all lines of S that intersect the interior of σ). This range space has a VC dimension which is a constant as can be easily verified. Let $X \subseteq \mathsf{S}$ be an ε-net for this range space, for $\varepsilon = 1/r$. By Theorem $21.7.1_{\text{p15}}$ (ε-net theorem), there exists such an ε-net X of this range space, of size $O((1/\varepsilon)\log(1/\varepsilon)) = O(r\log r)$. In fact, Theorem $21.7.1_{\text{p15}}$ states that an appropriate random sample is an ε-net with non-zero probability, which implies, by the probabilistic method, that such a net (of this size) exists.

Lemma 21.3.1. *There exists a $(1/r)$-cutting of a set of lines S in the plane of size $O\left((r\log r)^2\right)$.*

Proof: Consider the vertical decomposition $\mathcal{A}^{\mid}(X)$, where X is as above. We claim that this collection of trapezoids is the desired cutting.

The bound on the size is immediate, as the complexity of $\mathcal{A}^{\mid}(X)$ is $O\left(|X|^2\right)$ and $|X| = O(r\log r)$.

As for correctness, consider a vertical trapezoid σ in the arrangement $\mathcal{A}^{\mid}(X)$. It does not intersect any of the lines of X in its interior, since it is a trapezoid in the vertical decomposition $\mathcal{A}^{\mid}(X)$. Now, if σ intersected more than n/r lines of S in its interior, where $n = |\mathsf{S}|$, then it must be that the interior of σ intersects one of the lines of X, since X is an ε-net for S, a contradiction.

It follows that σ intersects at most $\varepsilon n = n/r$ lines of S in its interior. ■

Claim 21.3.2. *Any $(1/r)$-cutting in the plane of n lines contains at least $\Omega\left(r^2\right)$ regions.*

Proof: An arrangement of n lines (in general position) has $M = \binom{n}{2}$ intersections. However, the number of intersections of the lines intersecting a single region in the cutting is at most $m = \binom{n/r}{2}$. This implies that any cutting must be of size at least $M/m = \Omega\left(n^2/(n/r)^2\right) = \Omega\left(r^2\right)$. ■

We can get cuttings of size matching the above lower bound using the moments technique.

Theorem 21.3.3. *Let* S *be a set of* n *lines in the plane, and let* r *be a parameter. One can compute a* $(1/r)$*-cutting of* S *of size* $O(r^2)$.

Proof: Let $\mathsf{R} \subseteq \mathsf{S}$ be a random sample of size r, and consider its vertical decomposition $\mathcal{A}^{\mid}(\mathsf{R})$. If a vertical trapezoid $\sigma \in \mathcal{A}^{\mid}(\mathsf{R})$ intersects at most n/r lines of S, then we can add it to the output cutting. The other possibility is that a σ intersects $t(n/r)$ lines of S, for some $t > 1$, and let $\mathrm{cl}(\sigma) \subset \mathsf{S}$ be the conflict list of σ (i.e., the list of lines of S that intersect the interior of σ). Clearly, a $(1/t)$-cutting for the set $\mathrm{cl}(\sigma)$ forms a vertical decomposition (clipped inside σ) such that each trapezoid in this cutting intersects at most n/r lines of S. Thus, we compute such a cutting inside each such "heavy" trapezoid using the algorithm (implicit in the proof) of Lemma 21.3.1, and these subtrapezoids to the resulting cutting. Clearly, the size of the resulting cutting inside σ is $O\left(t^2 \log^2 t\right) = O\left(t^4\right)$. The resulting two-level partition is clearly the required cutting. By Theorem 21.2.5, the expected size of the cutting is

$$O\left(\mathbf{E}f(r) + \mathbf{E}\left[\sum_{\sigma \in \mathcal{F}(\mathsf{R})} \left(2\frac{\omega(\sigma)}{n/r}\right)^4\right]\right) = O\left(\mathbf{E}f(r) + \left(\frac{r}{n}\right)^4 \mathbf{E}\left[\sum_{\sigma \in \mathcal{F}(\mathsf{R})} (\omega(\sigma))^4\right]\right)$$
$$= O\left(\mathbf{E}f(r) + \left(\frac{r}{n}\right)^4 \cdot \mathbf{E}f(r)\left(\frac{n}{r}\right)^4\right) = O(\mathbf{E}f(r)) = O\left(r^2\right),$$

since $\mathbf{E}f(r)$ is proportional to the complexity of $\mathcal{A}(\mathsf{R})$ which is $O(r^2)$. ■

21.4. Bounds on the probability of a region to be created

Here we prove Lemma 21.2.1$_{\mathrm{p8}}$ in the "right" sampling model. The casual reader is encouraged to skip this section, as it contains mostly tedious (and not very insightful) calculations.

Let S be a given set of n objects. Let $\rho_{r,n}(d,k)$ be the probability that a region $\sigma \in \mathcal{T}$ whose defining set is of size d and whose stopping set is of size k appears in $\mathcal{F}(\mathsf{R})$, where R is a random sample from S of size r (without repetition).

Lemma 21.4.1. *We have* $\rho_{r,n}(d,k) = \dfrac{\binom{n-d-k}{r-d}}{\binom{n}{r}} = \dfrac{\binom{n-d-k}{r-d}}{\binom{n}{r-d}} \cdot \dfrac{\binom{r}{d}}{\binom{n-(r-d)}{d}} = \dfrac{\binom{n-d-k}{r-d}}{\binom{n-d}{r-d}} \cdot \dfrac{\binom{r}{d}}{\binom{n}{d}}$.

Proof: So, consider a region σ with d defining objects in $D(\sigma)$ and k detractors in $K(\sigma)$. We have to pick the d defining objects of $D(\sigma)$ to be in the random sample R of size r but avoid picking any of the k objects of $K(\sigma)$ to be in R.

The second part follows since $\binom{n}{r} = \binom{n}{r-d}\binom{n-(r-d)}{d} / \binom{r}{d}$. Indeed, for the right-hand side first pick a sample of size $r-d$ and then a sample of size d from the remaining objects. Merging the two random samples, we get a random sample of size r. However, since we do not care if an object is in the first sample or second sample, we observe that every such random sample is being counted $\binom{r}{d}$ times.

The third part is easier, as it follows from $\binom{n}{r-d}\binom{n-(r-d)}{d} = \binom{n}{d}\binom{n-d}{r-d}$. The two sides count the different ways to pick two subsets from a set of size n, the first one of size d and the second one of size $r-d$. ■

Lemma 21.4.2. *For $M \geq m \geq t \geq 0$, we have* $\left(\frac{m-t}{M-t}\right)^t \leq \frac{\binom{m}{t}}{\binom{M}{t}} \leq \left(\frac{m}{M}\right)^t$.

Proof: We have that $\alpha = \frac{\binom{m}{t}}{\binom{M}{t}} = \frac{m!}{(m-t)!t!}\frac{(M-t)!t!}{M!} = \frac{m}{M}\cdot\frac{m-1}{M-1}\cdots\frac{m-t+1}{M-t+1}$. Now, since $M \geq m$, we have that $\frac{m-i}{M-i} \leq \frac{m}{M}$, for all $i \geq 0$. As such, the maximum (resp. minimum) fraction on the right-hand size is m/M (resp. $\frac{m-t+1}{M-t+1}$). As such, we have $\left(\frac{m-t}{M-t}\right)^t \leq \left(\frac{m-t+1}{M-t+1}\right)^t \leq \alpha \leq (m/M)^t$. ■

Lemma 21.4.3. *Let $0 \leq X, Y \leq N$. We have that* $\left(1 - \frac{X}{N}\right)^Y \leq \left(1 - \frac{Y}{2N}\right)^X$.

Proof: Since $1 - \alpha \leq \exp(-\alpha) \leq (1 - \alpha/2)$, for $0 \leq \alpha \leq 1$, it follows that

$$\left(1 - \frac{X}{N}\right)^Y \leq \exp\left(-\frac{XY}{N}\right) = \left(\exp\left(-\frac{Y}{n}\right)\right)^X \leq \left(1 - \frac{Y}{2n}\right)^X.$$ ■

Lemma 21.4.4. *For $2d \leq r \leq n/8$ and $k \leq n/2$, we have that*

$$\frac{1}{2^{2d}}\left(1 - 4\cdot\frac{r}{n}\right)^k\left(\frac{r}{n}\right)^d \;\leq\; \rho_{r,n}(d,k) \;\leq\; 2^{2d}\left(1 - \frac{1}{2}\cdot\frac{r}{n}\right)^k\left(\frac{r}{n}\right)^d.$$

Proof: By Lemma 21.4.1, Lemma 21.4.2, and Lemma 21.4.3 we have

$$\begin{aligned}\rho_{r,n}(d,k) = \frac{\binom{n-d-k}{r-d}}{\binom{n-d}{r-d}}\cdot\frac{\binom{r}{d}}{\binom{n}{d}} &\leq \left(\frac{n-d-k}{n-d}\right)^{r-d}\left(\frac{r}{n}\right)^d \;\leq\; \left(1 - \frac{k}{n}\right)^{r-d}\left(\frac{r}{n}\right)^d \leq 2^d\left(1 - \frac{k}{n}\right)^r\left(\frac{r}{n}\right)^d \\ &\leq 2^d\left(1 - \frac{r}{2n}\right)^k\left(\frac{r}{n}\right)^d,\end{aligned}$$

since $k \leq n/2$. As for the other direction, by similar argumentation, we have

$$\begin{aligned}\rho_{r,n}(d,k) = \frac{\binom{n-d-k}{r-d}}{\binom{n}{r-d}}\cdot\frac{\binom{r}{d}}{\binom{n-(r-d)}{d}} &\geq \left(\frac{n-d-k-(r-d)}{n-(r-d)}\right)^{r-d}\left(\frac{r-d}{n-(r-d)-d}\right)^d \\ &= \left(1 - \frac{d+k}{n-(r-d)}\right)^{r-d}\left(\frac{r-d}{n-r}\right)^d \;\geq\; \left(1 - \frac{d+k}{n/2}\right)^r\left(\frac{r/2}{n}\right)^d \\ &\geq \frac{1}{2^d}\left(1 - \frac{4r}{n}\right)^{d+k}\left(\frac{r}{n}\right)^d \geq \frac{1}{2^{2d}}\left(1 - \frac{4r}{n}\right)^k\left(\frac{r}{n}\right)^d,\end{aligned}$$

by Lemma 21.4.3 (setting $N = n/4$, $X = r$, and $Y = d + k$) and since $r \geq 2d$ and $4r/n \leq 1/2$. ■

21.5. Bibliographical notes

The technique described in this chapter is generally attributed to the work by Clarkson and Shor [CS89], which is historically inaccurate as the technique was developed by Clarkson [Cla88]. Instead of mildly confusing the matter by referring to it as the Clarkson technique, we decided to make sure to really confuse the reader and

refer to it as the *moments technique*. The Clarkson technique [Cla88] is in fact more general and implies a connection between the number of "heavy" regions and "light" regions. The general framework can be traced back to the earlier paper [Cla87]. This implies several beautiful results, some of which we cover later in the book.

For the full details of the algorithm of Section 21.1, the interested reader is refereed to the books [dBCKO08, BY98]. Interestingly, in some cases the merging stage can be skipped; see [Har00a].

Agarwal *et al.* [AMS98] presented a slightly stronger variant than the original version of Clarkson [Cla88] that allows a region to disappear even if none of the members of its stopping set are in the random sample. This stronger setting is used in computing the vertical decomposition of a single face in an arrangement (instead of the whole arrangement). Here an insertion of a faraway segment of the random sample might cut off a portion of the face of interest. In particular, in the settings of Agarwal *et al.* Axiom (ii) is replaced by the following:

(ii) If $\sigma \in \mathcal{F}(\mathsf{R})$ and R' is a subset of R with $D(\sigma) \subseteq \mathsf{R}'$, then $\sigma \in \mathcal{F}(\mathsf{R}')$.

Interestingly, Clarkson [Cla88] did not prove Theorem 21.2.5 using the exponential decay lemma but gave a direct proof. In fact, his proof implicitly contains the exponential decay lemma. We chose the current exposition since it is more modular and provides a better intuition of what is really going on and is hopefully slightly simpler. In particular, Lemma 21.2.1 is inspired by the work of Sharir [Sha03].

The exponential decay lemma (Lemma 21.2.4) was proved by Chazelle and Friedman [CF90]. The work of Agarwal *et al.* [AMS98] is a further extension of this result. Another analysis was provided by Clarkson *et al.* [CMS93].

Another way to reach similar results is using the technique of Mulmuley [Mul94], which relies on a direct analysis on 'stoppers' and 'triggers'. This technique is somewhat less convenient to use but is applicable to some settings where the moments technique does not apply directly. Also, his concept of the omega function might explain why randomized incremental algorithms perform better in practice than their worst case analysis [**?**].

Backwards analysis in geometric settings was first used by Chew [Che86] and was formalized by Seidel [Sei93]. It is similar to the "leave one out" argument used in statistics for cross validation. The basic idea was probably known to the Greeks (or Russians or French) at some point in time.

(Naturally, our summary of the development is cursory at best and not necessarily accurate, and all possible disclaimers apply. A good summary is provided in the introduction of [Sei93].)

Sampling model. As a rule of thumb all the different sampling approaches are similar and yield similar results. For example, we used such an alternative sampling approach in the "proof" of Lemma 21.2.1. It is a good idea to use whichever sampling scheme is the easiest to analyze in figuring out what's going on. Of course, a formal proof requires analyzing the algorithm in the sampling model its uses.

Lazy randomized incremental construction. If one wants to compute a single face that contains a marking point in an arrangement of curves, then the problem in using randomized incremental construction is that as you add curves, the region of interest shrinks, and regions that were maintained should be ignored. One option is to perform flooding in the vertical decomposition to figure out what trapezoids are still reachable from the marking point and maintaining only these trapezoids in the conflict graph. Doing it in each iteration is way too expensive, but luckily one can use a lazy strategy that performs this cleanup only a logarithmic number of times (i.e., you perform a cleanup in an iteration if the iteration number is, say, a power of 2). This strategy complicates the analysis a bit; see [dBDS95] for more details on this *lazy randomized incremental construction* technique. An alternative technique was suggested by the author for the (more restricted) case of planar arrangements; see [Har00b]. The idea is to compute only what the algorithm really needs to compute the output, by computing the vertical decomposition in an exploratory online fashion. The details are unfortunately overwhelming although the algorithm seems to perform quite well in practice.

Cuttings. The concept of cuttings was introduced by Clarkson. The first optimal size cuttings were constructed by Chazelle and Friedman [CF90], who proved the exponential decay lemma to this end. Our elegant proof follows the presentation by de Berg and Schwarzkopf [dBS95]. The problem with this approach is that the constant involved in the cutting size is awful[5]. Matoušek [Mat98] showed that there $(1/r)$-cuttings with $8r^2 + 6r + 4$ trapezoids, by using level approximation. A different approach was taken by the author [Har00a], who showed how to get cuttings which seem to be quite small (i.e., constant-wise) in practice. The basic idea is to do randomized incremental construction but at each iteration greedily add all the trapezoids with conflict list small enough to the cutting being output. One can prove that this algorithm also generates $O(r^2)$ cuttings, but the details are not trivial as the framework described in this chapter is not applicable for analyzing this algorithm.

Cuttings also can be computed in higher dimensions for hyperplanes. In the plane, cuttings can also be computed for well-behaved curves; see [SA95].

Another fascinating concept is *shallow cuttings*. These are cuttings covering only portions of the arrangement that are in the "bottom" of the arrangement. Matoušek came up with the concept [Mat92]. See [AES99, CCH09] for extensions and applications of shallow cuttings.

Even more on randomized algorithms in geometry. We have only scratched the surface of this fascinating topic, which is one of the cornerstones of "modern" computational geometry. The interested reader should have a look at the books by Mulmuley [Mul94], Sharir and Agarwal [SA95], Matoušek [Mat02], and Boissonnat and Yvinec [BY98].

21.6. Exercises

Exercise 21.6.1 (Convex hulls incrementally). Let P be a set of n points in the plane.

(A) Describe a randomized incremental algorithm for computing the convex hull $\mathcal{CH}(\mathsf{P})$. Bound the expected running time of your algorithm.

(B) Assume that for any subset of P, its convex hull has complexity t (i.e., the convex hull of the subset has t edges). What is the expected running time of your algorithm in this case? If your algorithm is not faster for this case (for example, think about the case where $t = O(\log n)$), describe a variant of your algorithm which is faster for this case.

Exercise 21.6.2 (Compressed quadtree made incremental). Given a set P of n points in $\mathbb{R}^d$, describe a randomized incremental algorithm for building a compressed quadtree for P that works in expected $O(dn \log n)$ time. Prove the bound on the running time of your algorithm.

21.7. From previous lectures

Theorem 21.7.1 (ε-net theorem, [HW87]). *Let $(X, \mathcal{R})$ be a range space of* VC *dimension δ, let* x *be a finite subset of X, and suppose that $0 < \varepsilon \leq 1$ and $\varphi < 1$. Let N be a set obtained by m random independent draws from* x*, where*

$$m \geq \max\left(\frac{4}{\varepsilon} \lg \frac{4}{\varphi}, \frac{8\delta}{\varepsilon} \lg \frac{16}{\varepsilon} \right). \tag{21.4}$$

Then N is an ε-net for x *with probability at least $1 - \varphi$.*

[5] This is why all computations related to cuttings should be done on a waiter's bill pad. As Douglas Adams put it: "On a waiter's bill pad, reality and unreality collide on such a fundamental level that each becomes the other and anything is possible, within certain parameters."

Lemma 21.7.2. *For any two random variables X and Y, we have* $\mathbf{E}\Big[\mathbf{E}\big[X \,\big|\, Y\big]\Big] = \mathbf{E}\big[X\big]$.

Bibliography

[AES99] P. K. Agarwal, A. Efrat, and M. Sharir. Vertical decomposition of shallow levels in 3-dimensional arrangements and its applications. *SIAM J. Comput.*, 29:912–953, 1999.

[AMS98] P. K. Agarwal, J. Matoušek, and O. Schwarzkopf. Computing many faces in arrangements of lines and segments. *SIAM J. Comput.*, 27(2):491–505, 1998.

[BY98] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, 1998.

[CCH09] C. Chekuri, K. L. Clarkson., and S. Har-Peled. On the set multi-cover problem in geometric settings. In *Proc. 25th Annu. Sympos. Comput. Geom.* (SoCG), pages 341–350, 2009.

[CF90] B. Chazelle and J. Friedman. A deterministic view of random sampling and its use in geometry. *Combinatorica*, 10(3):229–249, 1990.

[Che86] L. P. Chew. Building Voronoi diagrams for convex polygons in linear expected time. Technical Report PCS-TR90-147, Dept. Math. Comput. Sci., Dartmouth College, Hanover, NH, 1986.

[Cla87] K. L. Clarkson. New applications of random sampling in computational geometry. *Discrete Comput. Geom.*, 2:195–222, 1987.

[Cla88] K. L. Clarkson. Applications of random sampling in computational geometry, II. In *Proc. 4th Annu. Sympos. Comput. Geom.* (SoCG), pages 1–11, New York, NY, USA, 1988. ACM.

[CMS93] K. L. Clarkson, K. Mehlhorn, and R. Seidel. Four results on randomized incremental constructions. *Comput. Geom. Theory Appl.*, 3(4):185–212, 1993.

[CS89] K. L. Clarkson and P. W. Shor. Applications of random sampling in computational geometry, II. *Discrete Comput. Geom.*, 4:387–421, 1989.

[dBCKO08] M. de Berg, O. Cheong, M. van Kreveld, and M. H. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Santa Clara, CA, USA, 3rd edition, 2008.

[dBDS95] M. de Berg, K. Dobrindt, and O. Schwarzkopf. On lazy randomized incremental construction. *Discrete Comput. Geom.*, 14:261–286, 1995.

[dBS95] M. de Berg and O. Schwarzkopf. Cuttings and applications. *Internat. J. Comput. Geom. Appl.*, 5:343–355, 1995.

[Har00a] S. Har-Peled. Constructing planar cuttings in theory and practice. *SIAM J. Comput.*, 29(6):2016–2039, 2000.

[Har00b] S. Har-Peled. Taking a walk in a planar arrangement. *SIAM J. Comput.*, 30(4):1341–1367, 2000.

[HW87] D. Haussler and E. Welzl. ε-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[Mat92] J. Matoušek. Reporting points in halfspaces. *Comput. Geom. Theory Appl.*, 2(3):169–186, 1992.

[Mat98] J. Matoušek. On constants for cuttings in the plane. *Discrete Comput. Geom.*, 20:427–448, 1998.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Grad. Text in Math.* Springer, 2002.

[Mul94] K. Mulmuley. *Computational Geometry: An Introduction Through Randomized Algorithms*. Prentice Hall, Englewood Cliffs, NJ, 1994.

[SA95] M. Sharir and P. K. Agarwal. *Davenport-Schinzel Sequences and Their Geometric Applications*. Cambridge University Press, New York, 1995.

[Sei93] R. Seidel. Backwards analysis of randomized geometric algorithms. In J. Pach, editor, *New Trends in Discrete and Computational Geometry*, volume 10 of *Algorithms and Combinatorics*, pages 37–68. Springer-Verlag, 1993.

[Sha03] M. Sharir. The Clarkson-Shor technique revisited and extended. *Comb., Prob. & Comput.*, 12(2):191–201, 2003.

Chapter 22

Primality testing

By Sariel Har-Peled, December 30, 2015[1]

> *"The world is what it is; men who are nothing, who allow themselves to become nothing, have no place in it."*
>
> — Bend in the river, V.S. Naipaul

Introduction – how to read this write-up

In this note, we present a simple randomized algorithms for primality testing. The challenge is that it requires a non-trivial amount of number theory, which is not the purpose of this course. Nevertheless, this note is more or less self contained, and all necessary background is provided (assuming some basic mathematical familiarity with groups, fields and modulo arithmetic). It is however not really necessary to understand all the number theory material needed, and the reader can take it as given. In particular, I recommend to read the number theory background part without reading all of the proofs (at least on first reading). Naturally, a complete and total understanding of this material one needs to read everything carefully.

The description of the primality testing algorithm in this write-up is not minimal – there are shorter descriptions out there. However, it is modular – assuming the number theory machinery used is correct, the algorithm description is relatively straightforward.

22.1. Number theory background

22.1.1. Modulo arithmetic

22.1.1.1. Prime and coprime

For integer numbers x and y, let $x \mid y$ denotes that x divides y. The *greatest common divisor* (*gcd*) of two numbers x and y, denoted by $\gcd(x, y)$, is the largest integer that divides both x and y. The *least common multiple* (*lcm*) of x and y, denoted by $\operatorname{lcm}(x, y) = xy/\gcd(x, y)$, is the smallest integer α, such that $x \mid \alpha$ and $y \mid \alpha$. An integer number $p > 0$ is *prime* if it is divisible only by 1 and itself (we will consider 1 not to be prime).

Some standard definitions:

$$\begin{aligned} x, y \text{ are } \textit{coprime} &\iff \gcd(x,y) = 1, \\ \textit{quotient} \text{ of } x/y &\iff x \operatorname{div} y = \lfloor x/y \rfloor, \\ \textit{remainder} \text{ of } x/y &\iff x \bmod y = x - y\lfloor x/y \rfloor. \end{aligned}$$

The remainder $x \bmod y$ is sometimes referred to as *residue*.

22.1.1.2. Computing gcd

Computing the gcd of two numbers is a classical algorithm, see code on the right – proving that it indeed returns the right result follows by an easy induction. It is easy to verify that if the input is made out of $\log n$ bits, then this algorithm takes $O(\text{poly}(\log n))$ time (i.e., it is polynomial in the input size). Indeed, doing basic operations on numbers (i.e., multiplication, division, addition, subtraction, etc) with total of ℓ bits takes $O(\ell^2)$ time (naively – faster algorithms are known).

```
EuclidGCD(a, b):
    if (b = 0)
        return a
    else
        return EuclidGCD(b, a mod b)
```

Exercise 22.1.1. Show that $\gcd(F_n, F_{n-1}) = 1$, where F_i is the ith Fibonacci number. Argue that for two consecutive Fibonacci numbers **EuclidGCD**(F_n, F_{n-1}) takes $O(n)$ time, if every operation takes $O(1)$ time.

Lemma 22.1.2. *For all $\alpha, \beta > 0$ integers, there are integer numbers x and y, such that $\gcd(\alpha,\beta) = \alpha x + \beta y$, which can be computed in polynomial time; that is, $O(\text{poly}(\log\alpha + \log\beta))$.*

Proof: If $\alpha = \beta$ then the claim trivially holds. Otherwise, assume that $\alpha > \beta$ (otherwise, swap them), and observe that $\gcd(\alpha,\beta) = \gcd(\alpha \bmod \beta, \beta)$. In particular, by induction, there are integers x', y', such that $\gcd(\alpha \bmod \beta, \beta) = x'(\alpha \bmod \beta) + y'\beta$. However, $\tau = \alpha \bmod \beta = \alpha - \beta\lfloor\alpha/\beta\rfloor$. As such, we have

$$\gcd(\alpha,\beta) = \gcd(\alpha \bmod \beta, \beta) = x'\left(\alpha - \beta\lfloor\alpha/\beta\rfloor\right) + y'\beta = x'\alpha + (y' - \beta\lfloor\alpha/\beta\rfloor)\beta,$$

as claimed. The running time follows immediately by modifying **EuclidGCD** to compute these numbers. ■

We use $\alpha \equiv \beta \pmod n$ or $\alpha \equiv_n \beta$ to denote that α and β are *congruent modulo* n; that is $\alpha \bmod n = \beta \bmod n$. Or put differently, we have $n \mid (\alpha - \beta)$. The set $\mathbb{Z}_n = \{0, \ldots, n-1\}$ form a *group* under addition modulo n (see Definition 22.1.9$_{\text{p4}}$ for a formal definition of a group). The more interesting creature is $\mathbb{Z}_n^* = \left\{x \,\middle|\, x \in \{1,\ldots,n\}, x > 0, \text{ and } \gcd(x,n) = 1\right\}$, which is a *group* modulo n under multiplication.

Remark 22.1.3. Observe that $\mathbb{Z}_1^* = \{1\}$, while for $n > 1$, $\mathbb{Z}_n^*$ does not contain n.

Lemma 22.1.4. *For any element $\alpha \in \mathbb{Z}_n^*$, there exists a unique inverse element $\beta = \alpha^{-1} \in \mathbb{Z}_n^*$ such that $\alpha * \beta \equiv_n 1$. Furthermore, the inverse can be computed in polynomial time*[②].

Proof: Since $\alpha \in \mathbb{Z}_n^*$, we have that $\gcd(\alpha, n) = 1$. As such, by Lemma 22.1.2, there exists x and y integers, such that $x\alpha + yn = 1$. That is $x\alpha \equiv 1 \pmod n$, and clearly $\beta := x \bmod n$ is the desired inverse, and it can be computed in polynomial time by Lemma 22.1.2.

As for uniqueness, assume that there are two inverses β, β' to $\alpha < n$, such that $\beta < \beta' < n$. But then $\beta\alpha \equiv_n \beta'\alpha \equiv_n 1$, which implies that $n \mid (\beta' - \beta)\alpha$, which implies that $n \mid \beta' - \beta$, which is impossible as $0 < \beta' - \beta < n$. ■

[②] Again, as is everywhere in this chapter, the polynomial time is in the number of bits needed to specify the input.

It is now straightforward, but somewhat tedious, to verify the following (the interested reader that had not encountered this stuff before can spend some time proving this).

Lemma 22.1.5. *The set $\mathbb{Z}_n$ under the + operation modulo n is a group, as is $\mathbb{Z}_n^*$ under multiplication modulo n. More importantly, for a prime number p, $\mathbb{Z}_p$ forms a field with the $+, *$ operations modulo p (see Definition 22.1.17$_{p6}$).*

22.1.1.3. The Chinese remainder theorem

Theorem 22.1.6 (Chinese remainder theorem). *Let $n_1, \ldots, n_k$ be coprime numbers, and let $n = n_1 n_2 \cdots n_k$. For any residues $r_1 \in \mathbb{Z}_{n_1}, \ldots, r_k \in \mathbb{Z}_{n_k}$, there is a unique $r \in \mathbb{Z}_n$, which can be computed in polynomial time, such that $r \equiv r_i \pmod{n_i}$, for $i = 1, \ldots, k$.*

Proof: By the coprime property of the n_is it follows that $\gcd(n_i, n/n_i) = 1$. As such, $n/n_i \in \mathbb{Z}_{n_i}^*$, and it has a unique inverse m_i modulo n_i; that is $(n/n_i)m_i \equiv 1 \pmod{n_i}$. So set $r = \sum_i r_i m_i n/n_i$. Observe that for $i \neq j$, we have that $n_j \mid (n/n_i)$, and as such $r_i m_i n/n_i \pmod{n_j} \equiv 0 \pmod{n_j}$. As such, we have

$$r \bmod n_j = \left(\sum_i \left(r_i m_i \frac{n}{n_i} \bmod n_j \right) \right) \bmod n_j = \left(r_j m_j \frac{n}{n_j} \bmod n_j \right) \bmod n_j = r_j * 1 \bmod n_j = r_j.$$

As for uniqueness, if there is another such number r', such that $r < r' < n$, then $r' - r \pmod{n_i} = 0$ implying that $n_i \mid r' - r$, for all i. Since all the n_is are coprime, this implies that $n \mid r' - r$, which is of course impossible. ■

Lemma 22.1.7 (Fast exponentiation). *Given numbers b, c, n, one can compute $b^c \bmod n$ in polynomial time.*

Proof: The key property we need is that

$$xy \bmod n = \Big((x \bmod n)\,(y \bmod n) \Big) \bmod n.$$

Now, if c is even, then we can compute

$$b^c \bmod n = \left(b^{c/2} \right)^2 \bmod n = \left(b^{c/2} \bmod n \right)^2 \bmod n.$$

Similarly, if c is odd, we have

$$b^c \bmod n = (b \bmod n) \left(b^{(c-1)/2} \right)^2 \bmod n = (b \bmod n) \left(b^{(c-1)/2} \bmod n \right)^2 \bmod n.$$

Namely, computing $b^c \bmod n$ can be reduced to recursively computing $b^{\lfloor c/2 \rfloor} \bmod n$, and a constant number of operations (on numbers that are smaller than n). Clearly, the depth of the recursion is $O(\log c)$. ■

22.1.1.4. Euler totient function

The *Euler totient function* $\phi(n) = \left| \mathbb{Z}_n^* \right|$ is the number of positive integer numbers that at most n and are coprime with n. If n is prime then $\phi(n) = n - 1$.

Lemma 22.1.8. *Let $n = p_1^{k_1} \cdots p_t^{k_t}$, where the p_is are prime numbers and the k_is are positive integers (this is the* prime factorization *of n). Then $\phi(n) = \prod_{i=1}^{t} p_i^{k_i - 1}(p_i - 1)$. and this quantity can be computed in polynomial time if the factorization is given.*

Proof: Observe that $\phi(1) = 1$ (see Remark 22.1.3), and for a prime number p, we have that $\phi(p) = p - 1$. Now, for $k > 1$, and p prime we have that $\phi(p^k) = p^{k-1}(p-1)$, as a number $x \le p^k$ is coprime with p^k, if and only if $x \bmod p \neq 0$, and $(p-1)/p$ fraction of the numbers in this range have this property.

Now, if n and m are relative primes, then $\gcd(x, nm) = 1 \iff \gcd(x, n) = 1$ and $\gcd(x, m) = 1$. In particular, there are $\phi(n)\phi(m)$ pairs $(\alpha, \beta) \in \mathbb{Z}_n^* \times \mathbb{Z}_m^*$, such that $\gcd(\alpha, n) = 1$ and $\gcd(\beta, m) = 1$. By the Chinese remainder theorem (Theorem 22.1.6), each such pair represents a unique number in the range $1, \ldots, nm$, as desired.

Now, the claim follows by easy induction on the prime factorization of the given number. ■

22.1.2. Structure of the modulo group $\mathbb{Z}_n$

22.1.2.1. Some basic group theory

Definition 22.1.9. A *group* is a set, $\mathcal{G}$, together with an operation $\times$ that combines any two elements a and b to form another element, denoted $a \times b$ or ab. To qualify as a group, the set and operation, $(\mathsf{G}, \times)$, must satisfy the following:

(A) (Closure) For all $a, b \in \mathcal{G}$, the result of the operation, $a \times b \in \mathcal{G}$.
(B) (Associativity) For all $a, b, c \in \mathcal{G}$, we have $(a \times b) \times c = a \times (b \times c)$.
(C) (Identity element) There exists an element $\mathsf{i} \in \mathcal{G}$, called the *identity element*, such that for every element $a \in \mathcal{G}$, the equation $\mathsf{i} \times a = a \times \mathsf{i} = a$ holds.
(D) (Inverse element) For each $a \in \mathcal{G}$, there exists an element $b \in \mathcal{G}$ such that $a \times b = b \times a = \mathsf{i}$.

A group is *abelian* (aka, *commutative group*) if for all $a, b \in \mathcal{G}$, we have that $a \times b = b \times a$.

In the following we restrict our attention to abelian groups since it makes the discussion somewhat simpler. In particular, some of the claims below holds even without the restriction to abelian groups.

The identity element is unique. Indeed, if both $f, g \in \mathcal{G}$ are identity elements, then $f = f \times g = g$. Similarly, for every element $x \in \mathcal{G}$ there exists a unique inverse $y = x^{-1}$. Indeed, if there was another inverse z, then $y = y \times \mathsf{i} = y \times (x \times z) = (y \times x) \times z = \mathsf{i} \times z = z$.

22.1.2.2. Subgroups

For a group $\mathcal{G}$, a subset $\mathcal{H} \subseteq \mathcal{G}$ that is also a group (under the same operation) is a *subgroup*.

For $x, y \in \mathcal{G}$, let us define $x \sim y$ if $x/y \in \mathcal{H}$. Here $x/y = xy^{-1}$ and y^{-1} is the inverse of y in $\mathcal{G}$. Observe that $(y/x)(x/y) = \left(yx^{-1}\right)\left(xy^{-1}\right) = \mathsf{i}$. That is y/x is the inverse of x/y, and it is in $\mathcal{H}$. But that implies that $x \sim y \implies y \sim x$. Now, if $x \sim y$ and $y \sim z$, then $x/y, y/z \in \mathcal{H}$. But then $x/y \times y/z \in \mathcal{H}$, and furthermore $x/y \times y/z = xy^{-1}yz^{-1} = xz^{-1} = x/z$. that is $x \sim z$. Together, this implies that $\sim$ is an equivalence relationship.

Furthermore, observe that if $x/y = x/z$ then $y^{-1} = x^{-1}(x/y) = x^{-1}(x/z) = z^{-1}$, that is $y = z$. In particular, the equivalence class of $x \in \mathcal{G}$, is $[x] = \left\{ z \in \mathcal{G} \;\middle|\; x \sim z \right\}$. Observe that if $x \in \mathcal{H}$ then $\mathsf{i}/x = \mathsf{i}x^{-1} = x^{-1} \in \mathcal{H}$, and thus $\mathsf{i} \sim x$. That is $\mathcal{H} = [x]$. The following is now easy.

Lemma 22.1.10. *Let $\mathcal{G}$ be an abelian group, and let $\mathcal{H} \subseteq \mathcal{G}$ be a subgroup. Consider the set $\mathcal{G}/\mathcal{H} = \left\{ [x] \;\middle|\; x \in \mathcal{G} \right\}$. We claim that $\left|[x]\right| = \left|[y]\right|$ for any $x, y \in \mathcal{G}$. Furthermore $\mathcal{G}/\mathcal{H}$ is a group (that is, the* quotient group*), with $[x] \times [y] = [x \times y]$.*

Proof: Pick an element $\alpha \in [x]$, and $\beta \in [y]$, and consider the mapping $f(x) = x\alpha^{-1}\beta$. We claim that f is one to one and onto from $[x]$ to $[y]$. For any $\gamma \in [x]$, we have that $\gamma\alpha^{-1} = \gamma/\alpha \in \mathcal{H}$ As such, $f(\gamma) = \gamma\alpha^{-1}\beta \in [\beta] = [y]$. Now, for any $\gamma, \gamma' \in [x]$ such that $\gamma \neq \gamma'$, we have that if $f(\gamma) = \gamma\alpha^{-1}\beta = \gamma'\alpha^{-1}\beta = f(\gamma')$, then by multiplying by $\beta^{-1}\alpha$, we have that $\gamma = \gamma'$. That is, f is one to one, implying that $\left|[x]\right| = \left|[y]\right|$.

The second claim follows by careful but tediously checking that the conditions in the definition of a group holds. ■

Lemma 22.1.11. *For a finite abelian group $\mathcal{G}$ and a subgroup $\mathcal{H} \subseteq \mathcal{G}$, we have that $|\mathcal{H}|$ divides $|\mathcal{G}|$.*

Proof: By Lemma 22.1.10, we have that $|\mathcal{G}| = |\mathcal{H}| \cdot |\mathcal{G}/\mathcal{H}|$, as $\mathcal{H} = [\mathsf{i}]$. ■

22.1.2.3. Cyclic groups

Lemma 22.1.12. *For a finite group $\mathcal{G}$, and any element $g \in \mathcal{G}$, the set $\langle g \rangle = \left\{ g^i \;\middle|\; i \geq 0 \right\}$ is a group.*

Proof: Since $\mathcal{G}$ is finite, there are integers $i > j \geq 1$, such that $i \neq j$ and $g^i = g^j$, but then $g^j \times g^{i-j} = g^i = g^j$. That is $g^{i-j} = \mathsf{i}$ and, by definition, we have $g^{i-j} \in \langle g \rangle$. It is now straightforward to verify that the other properties of a group hold for $\langle g \rangle$. ■

In particular, for an element $g \in \mathcal{G}$, we define its *order* as $\operatorname{ord}(g) = \left|\langle g \rangle\right|$, which clearly is the minimum positive integer m, such that $g^m = \mathsf{i}$. Indeed, for $j > m$, observe that $g^j = g^{j \bmod m} \in X = \left\{ \mathsf{i}, g, g^2, \ldots, g^{m-1} \right\}$, which implies that $\langle g \rangle = X$.

A group $\mathcal{G}$ is *cyclic*, if there is an element $g \in \mathcal{G}$, such that $\langle g \rangle = \mathcal{G}$. In such a case g is a *generator* of $\mathcal{G}$.

Lemma 22.1.13. *For any finite abelian group $\mathcal{G}$, and any $g \in \mathcal{G}$, we have that* $\operatorname{ord}(g)$ *divides* $|\mathcal{G}|$*, and* $g^{|\mathcal{G}|} = \mathsf{i}$.

Proof: By Lemma 22.1.12, the set $\langle g \rangle$ is a subgroup of $\mathcal{G}$. By Lemma 22.1.11, we have that $\operatorname{ord}(g) = \left|\langle g \rangle\right| \mid |\mathcal{G}|$. As such, $g^{|\mathcal{G}|} = \left(g^{\operatorname{ord}(g)}\right)^{|\mathcal{G}|/\operatorname{ord}(g)} = \left(\mathsf{i}\right)^{|\mathcal{G}|/\operatorname{ord}(g)} = \mathsf{i}$. ■

22.1.2.4. Modulo group

Lemma 22.1.14. *For any integer n, consider the additive group $\mathbb{Z}_n$. Then, for any $x \in \mathbb{Z}_n$, we have that $x \cdot \operatorname{ord}(x) = \operatorname{lcm}(x, n)$. In particular,* $\operatorname{ord}(x) = \dfrac{\operatorname{lcm}(n, x)}{x} = \dfrac{n}{\gcd(n, x)}$*. If n is prime, and $x \neq 0$ then* $\operatorname{ord}(x) = |\mathbb{Z}_n| = n$*, and $\mathbb{Z}_n$ is a cyclic group.*

Proof: We are working modulo n here under additions, and the identity element is 0. As such, $x \cdot \operatorname{ord}(x) \equiv_n 0$, which implies that $n \mid x \operatorname{ord}(x)$. By definition, $\operatorname{ord}(x)$ is the minimal number that has this property, implying that $\operatorname{ord}(x) = \dfrac{\operatorname{lcm}(n, x)}{x}$. Now, $\operatorname{lcm}(n, x) = nx/\gcd(n, x)$. The second claim is now easy. ■

Theorem 22.1.15. ***(Euler's theorem)*** *For all n and $x \in \mathbb{Z}_n^*$, we have $x^{\phi(n)} \equiv 1 \pmod{n}$.*

(Fermat's theorem) *If p is a prime then $\forall x \in \mathbb{Z}_p^* \quad x^{p-1} \equiv 1 \pmod{p}$.*

Proof: The group $\mathbb{Z}_n^*$ is abelian and has $\phi(n)$ elements, with 1 being the identity element (duh!). As such, by Lemma 22.1.13, we have that $x^{\phi(n)} = x^{|\mathbb{Z}_n^*|} \equiv 1 \pmod{n}$, as claimed.

The second claim follows by setting $n = p$, and recalling that $\phi(p) = p - 1$, if p is a prime. ■

One might be tempted to think that Lemma 22.1.14 implies that if p is a prime then $\mathbb{Z}_p^*$ is a cyclic group, but this does not follow, as the cardinality of $\mathbb{Z}_p^*$ is $\phi(p) = p - 1$, which is not a prime number (for $p > 2$). To prove that $\mathbb{Z}_p^*$ is cyclic, let us go back shortly to the totient function.

Lemma 22.1.16. *For any $n > 0$, we have $\sum_{d \mid n} \phi(d) = n$.*

Proof: For any $g > 0$, let $V_g = \left\{ x \;\middle|\; x \in \{1, \ldots, n\} \text{ and } \gcd(x, n) = g \right\}$. Now, $x \in V_g \iff \gcd(x, n) = g \iff \gcd(x/g, n/g) = 1 \iff x/g \in \mathbb{Z}_{n/g}^*$. Since $V_1, V_2, \ldots, V_n$ form a partition of $\{1, \ldots, n\}$, it follows that

$$n = \sum_g \left|V_g\right| = \sum_{g \mid n} \left|\mathbb{Z}_{n/g}^*\right| = \sum_{g \mid n} \phi(n/g) = \sum_{d \mid n} \phi(d).$$ ■

22.1.2.5. Fields

Definition 22.1.17. A *field* is an algebraic structure $\langle \mathbb{F}, +, *, 0, 1 \rangle$ consisting of two abelian groups:

(A) $\mathbb{F}$ under $+$, with 0 being the identity element.

(B) $\mathbb{F} \setminus \{0\}$ under $*$, with 1 as the identity element (here $0 \neq 1$).

Also, the following property (*distributivity of multiplication over addition*) holds:

$$\forall a, b, c \in \mathbb{F} \qquad a * (b + c) = (a * b) + (a * c).$$

We need the following: A polynomial p of degree k over a field $\mathbb{F}$ has at most k roots. indeed, if p has the root α then it can be written as $p(x) = (x - \alpha)q(x)$, where $q(x)$ is a polynomial of one degree lower. To see this, we divide $p(x)$ by the polynomial $(x - \alpha)$, and observe that $p(x) = (x - \alpha)q(x) + \beta$, but clearly $\beta = 0$ since $p(\alpha) = 0$. As such, if p had t roots $\alpha_1, \ldots, \alpha_t$, then $p(x) = q(x) \prod_{i=1}^{t} (x - \alpha_i)$, which implies that p would have degree at least t.

22.1.2.6. $\mathbb{Z}_p^*$ is cyclic for prime numbers

For a prime number p, the group $\mathbb{Z}_p^*$ has size $\phi(p) = p - 1$, which is not a prime number for $p > 2$. As such, Lemma 22.1.13 does not imply that there must be an element in $\mathbb{Z}_p^*$ that has order $p - 1$ (and thus $\mathbb{Z}_p^*$ is cyclic). Instead, our argument is going to be more involved and less direct.

Lemma 22.1.18. *For $k < n$, let $R_k = \left\{ x \in \mathbb{Z}_p^* \;\middle|\; \operatorname{ord}(x) = k \right\}$ be the set of all numbers in $\mathbb{Z}_p^*$ that are of order k. We have that $|R_k| \leq \phi(k)$.*

Proof: Clearly, all the elements of R_k are roots of the polynomial $x^k - 1 = 0 \pmod{n}$. By the above, this polynomial has at most k roots. Now, if R_k is not empty, then it contains an element $x \in R_k$ of order k, which implies that for all $i < j \leq k$, we have that $x^i \not\equiv x^j \pmod{n}$, as the order of x is the size of $\langle x \rangle$, and the minimum k such that $x^k \equiv 1 \pmod{n}$. In particular, we have that $R_k \subseteq \langle x \rangle$, as for $y = x^j$, we have that $y^k \equiv_n x^{jk} \equiv_n 1^j \equiv_n 1$.

Observe that for $y = x^i$, if $g = \gcd(k, i) > 1$, then $y^{k/g} \equiv_n x^{i(k/g)} \equiv_n x^{\operatorname{lcm}(i,k)} \equiv_n 1$; that is, $\operatorname{ord}(y) \leq k/g < k$, and $y \notin R_k$. As such, R_k contains only elements of x^i such that $\gcd(i, k) = 1$. That is $R_k \subseteq \mathbb{Z}_k^*$. The claim now readily follows as $\left|\mathbb{Z}_k^*\right| = \phi(k)$. ■

Lemma 22.1.19. *For any prime p, the group $\mathbb{Z}_p^*$ is cyclic.*

Proof: For $p = 2$ the claim trivially holds, so assume $p > 2$. If the set R_{p-1}, from Lemma 22.1.18, is not empty, then there is $g \in R_{p-1}$, it has order $p - 1$, and it is a generator of $\mathbb{Z}_p^*$, as $\left|\mathbb{Z}_p^*\right| = p - 1$, implying that $\mathbb{Z}_p^* = \langle g \rangle$ and this group is cyclic.

Now, by Lemma 22.1.13, we have that for any $y \in \mathbb{Z}_p^*$, we have that $\operatorname{ord}(y) \mid p - 1 = \left|\mathbb{Z}_p^*\right|$. This implies that R_k is empty if k does not divides $p - 1$. On the other hand, $R_1, \ldots, R_{p-1}$ form a partition of $\mathbb{Z}_p^*$. As such, we have that

$$p - 1 = \left|\mathbb{Z}_p^*\right| = \sum_{k \mid p-1} |R_k| \leq \sum_{k \mid p-1} \phi(k) = p - 1,$$

by Lemma 22.1.18 and Lemma 22.1.16$_{\text{p5}}$, implying that the inequality in the above display is equality, and for all $k \mid p - 1$, we have that $|R_k| = \phi(k)$. In particular, $\left|R_{p-1}\right| = \phi(p - 1) > 0$, and by the above the claim follows. ■

22.1.2.7. $\mathbb{Z}_n^*$ is cyclic for powers of a prime

Lemma 22.1.20. *Consider any odd prime p, and any integer $c \geq 1$, then the group $\mathbb{Z}_n^*$ is cyclic, where $n = p^c$.*

Proof: Let g be a generator of $\mathbb{Z}_p^*$. Observe that $g^{p-1} \equiv 1 \pmod{p}$. The number $g < p$, and as such p does not divide g, and also p does not divide g^{p-2}, and also p does not divide $p-1$. As such, p^2 does not divide $\Delta = (p-1)g^{p-2}p$; that is, $\Delta \not\equiv 0 \pmod{p^2}$. As such, we have that

$$(g+p)^{p-1} \equiv g^{p-1} + \binom{p-1}{1} g^{p-2}p \equiv g^{p-1} + \Delta \not\equiv g^{p-1} \pmod{p^2}$$

$$\implies \qquad (g+p)^{p-1} \not\equiv 1 \pmod{p^2} \qquad \text{or} \qquad g^{p-1} \not\equiv 1 \pmod{p^2}.$$

Renaming $g+p$ to be g, if necessary, we have that $g^{p-1} \not\equiv 1 \pmod{p^2}$, but by Theorem 22.1.15$_{\text{p5}}$, $g^{p-1} \equiv 1 \pmod{p}$. As such, $g^{p-1} = 1 + \beta p$, where p does not divide β. Now, we have

$$g^{p(p-1)} = (1+\beta p)^p = 1 + \binom{p}{1}\beta p + \beta p^3 \text{<whatever>} = 1 + \gamma_1 p^2,$$

where γ_1 is an integer (the p^3 is not a typo – the binomial coefficient contributes at least one factor of p – here we are using that $p > 2$). In particular, as p does not divides β, it follows that p does not divides γ_1 either. Let us apply this argumentation again to

$$g^{p^2(p-1)} = \left(1 + \gamma_1 p^2\right)^p = 1 + \gamma_1 p^3 + p^4 \text{<whatever>} = 1 + \gamma_2 p^3,$$

where again p does not divides γ_2. Repeating this argument, for $i = 1, \ldots, c-2$, we have

$$\alpha_i = g^{p^i(p-1)} = \left(g^{p^{i-1}(p-1)}\right)^p = \left(1 + \gamma_{i-1}p^i\right)^p = 1 + \gamma_{i-1}p^{i+1} + p^{i+2} \text{<whatever>} = 1 + \gamma_i p^{i+1},$$

where p does not divides γ_i. In particular, this implies that $\alpha_{c-2} = 1 + \gamma_{c-2}p^{c-1}$ and p does not divides γ_{c-2}. This in turn implies that $\alpha_{c-2} \not\equiv 1 \pmod{p^c}$.

Now, the order of g in $\mathbb{Z}_n$, denoted by k, must divide $\left|\mathbb{Z}_n^*\right|$ by Lemma 22.1.13$_{\text{p5}}$. Now $\left|\mathbb{Z}_n^*\right| = \phi(n) = p^{c-1}(p-1)$, see Lemma 22.1.8$_{\text{p3}}$. So, $k \mid p^{c-1}(p-1)$. Also, $\alpha_{c-2} \not\equiv 1 \pmod{p^c}$. implies that k does not divides $p^{c-2}(p-1)$. It follows that $p^{c-1} \mid k$. So, let us write $k = p^{c-1}k'$, where $k' \leq (p-1)$. This, by definition, implies that $g^k \equiv 1 \pmod{p^c}$. Now, $g^p \equiv g \pmod{p}$, because g is a generator of $\mathbb{Z}_p^*$. As such, we have that

$$g^k \equiv_p g^{p^\delta k'} \equiv_p (g^p)^{p^{\delta-1}k'} \equiv_p (g)^{p^{\delta-1}k'} \equiv_p \ldots \equiv_p (g)^{k'} \equiv_p \left(g^k \bmod p^c\right) \bmod p \equiv_p 1.$$

Namely, $g^{k'} \equiv 1 \pmod{p}$, which implies, as g as a generator of $\mathbb{Z}_p^*$, that either $k' = 1$ or $k' = p-1$. The case $k' = 1$ is impossible, as this implies that $g = 1$, and it can not be the generator of $\mathbb{Z}_p^*$. We conclude that $k = p^{c-1}(p-1)$; that is, $\mathbb{Z}_n^*$ is cyclic. ■

22.1.3. Quadratic residues

22.1.3.1. Quadratic residue

Definition 22.1.21. An integer α is a *quadratic residue* modulo a positive integer n, if $\gcd(\alpha, n) = 1$ and for some integer β, we have $\alpha \equiv \beta^2 \pmod{n}$.

Theorem 22.1.22 (Euler's criterion). *Let p be an odd prime, and $\alpha \in \mathbb{Z}_p^*$. We have that*

(A) $\alpha^{(p-1)/2} \equiv_p \pm 1$.
(B) If α *is a quadratic residue, then* $\alpha^{(p-1)/2} \equiv_p 1$.
(C) If α *is not a quadratic residue, then* $\alpha^{(p-1)/2} \equiv_p -1$.

Proof: (A) Let $\gamma = \alpha^{(p-1)/2}$, and observe that $\gamma^2 \equiv_p \alpha^{p-1} \equiv 1$, by Fermat's theorem (Theorem 22.1.15$_{\text{p5}}$), which implies that γ is either $+1$ or -1, as the polynomial $x^2 - 1$ has at most two roots over a field.

(B) Let $\alpha \equiv_p \beta^2$, and again by Fermat's theorem, we have $\alpha^{(p-1)/2} \equiv_p \beta^{p-1} \equiv_p 1$.

(C) Let X be the set of elements in $\mathbb{Z}_p^*$ that are not quadratic residues, and consider $\alpha \in X$. Since $\mathbb{Z}_p^*$ is a group, for any $x \in \mathbb{Z}_p^*$ there is a unique $y \in \mathbb{Z}_p^*$ such that $xy \equiv_p \alpha$. As such, we partition $\mathbb{Z}_p^*$ into pairs $C = \left\{ \{x, y\} \;\middle|\; x, y \in \mathbb{Z}_p^* \text{ and } xy \equiv_p \alpha \right\}$. We have that

$$\tau \equiv_p \prod_{\beta \in \mathbb{Z}_p^*} \beta \equiv_p \prod_{\{x,y\} \in C} xy \equiv_p \prod_{\{x,y\} \in C} \alpha \equiv_p \alpha^{(p-1)/2}.$$

Let consider a similar set of pair, but this time for 1: $D = \left\{ \{x, y\} \;\middle|\; x, y \in \mathbb{Z}_p^*, x \neq y \text{ and } xy \equiv_p 1 \right\}$. Clearly, D does not contain -1 and 1, but all other elements in $\mathbb{Z}_p^*$ are in D. As such,

$$\tau \equiv_p \prod_{\beta \in \mathbb{Z}_p^*} \beta \equiv_p (-1)1 \prod_{\{x,y\} \in D} xy \equiv_p \prod_{\{x,y\} \in D} 1 \equiv_p -1. \qquad \blacksquare$$

22.1.3.2. Legendre symbol

For an odd prime p, and an integer a with $\gcd(a, n) = 1$, the *Legendre symbol* $(a \mid p)$ is one if a is a quadratic residue modulo p, and -1 otherwise (if $p \mid a$, we define $(a \mid p) = 0$). Euler's criterion (Theorem 22.1.22) implies the following equivalent definition.

Definition 22.1.23. The *Legendre symbol*, for a prime number p, and $a \in \mathbb{Z}_p^*$, is

$$(a \mid p) = a^{(p-1)/2} \pmod{p}.$$

The following is easy to verify.

Lemma 22.1.24. *Let* p *be an odd prime, and let* a, b *be integer numbers. We have:*
(i) $(-1 \mid p) = (-1)^{(p-1)/2}$.
(ii) $(a \mid p)(b \mid p) = (ab \mid p)$.
(iii) If $a \equiv_p b$ *then* $(a \mid p) = (b \mid p)$.

Lemma 22.1.25 (Gauss' lemma). *Let* p *be an odd prime and let* a *be an integer that is not divisible by* p. *Let* $X = \left\{ \alpha_j = ja \pmod{p} \;\middle|\; j = 1, \ldots, (p-1)/2 \right\}$, *and* $L = \left\{ x \in X \;\middle|\; x > p/2 \right\} \subseteq X$. *Then* $(a \mid p) = (-1)^n$, *where* $n = |L|$.

Proof: Observe that for any distinct i, j, such that $1 \leq i \leq j \leq (p-1)/2$, we have that $ja \equiv ia \pmod{p}$ implies that $(j - i)a \equiv 0 \pmod{p}$, which is impossible as $j - i < p$ and $\gcd(a, p) = 1$. As such, all the elements of X are distinct, and $|X| = (p-1)/2$. We have a somewhat stronger property: If $ja \equiv p - ia \pmod{p}$ implies $(j + i)a \equiv 0 \pmod{p}$, which is impossible. That is, $S = X \setminus L$, and $\overline{L} = \left\{ p - \ell \;\middle|\; \ell \in L \right\}$ are disjoint, and $S \cup \overline{L} = \left\{ 1, \ldots, (p-1)/2 \right\}$. As such,

$$\left(\frac{p-1}{2}\right)! \equiv \prod_{x \in S} x \cdot \prod_{y \in L} (p - y) \equiv (-1)^n \prod_{x \in S} x \cdot \prod_{y \in L} y \equiv (-1)^n \prod_{j=1}^{(p-1)/2} ja \equiv (-1)^n a^{(p-1)/2} \left(\frac{p-1}{2}\right)! \pmod{p}.$$

Dividing both sides by $(-1)^n((p-1)/2)!$, we have that $(a \mid p) \equiv a^{(p-1)/2} \equiv (-1)^n \pmod{p}$, as claimed. $\blacksquare$

Lemma 22.1.26. *If p is an odd prime, and $a > 2$ and $\gcd(a,p) = 1$ then $(a \mid p) = (-1)^{\Delta}$, where $\Delta = \sum_{j=1}^{(p-1)/2} \lfloor ja/p \rfloor$. Furthermore, we have $(2 \mid p) = (-1)^{(p^2-1)/8}$.*

Proof: Using the notation of Lemma 22.1.25, we have

$$\sum_{j=1}^{(p-1)/2} ja = \sum_{j=1}^{(p-1)/2} \Big(\lfloor ja/p \rfloor p + (ja \bmod p) \Big) = \Delta p + \sum_{x \in S} x + \sum_{y \in L} y = (\Delta + n)p + \sum_{x \in S} x - \sum_{y \in \overline{L}} y$$
$$= (\Delta + n)p + \sum_{j=1}^{(p-1)/2} j - 2 \sum_{y \in \overline{L}} y.$$

Rearranging, and observing that $\sum_{j=1}^{(p-1)/2} j = \frac{p-1}{2} \cdot \frac{1}{2}\Big(\frac{p-1}{2} + 1\Big) = \frac{p^2-1}{8}$. We have that

$$(a-1)\frac{p^2-1}{8} = (\Delta + n)p - 2\sum_{y \in \overline{L}} y. \qquad \Longrightarrow \qquad (a-1)\frac{p^2-1}{8} \equiv (\Delta + n)p \pmod 2. \tag{22.1}$$

Observe that $p \equiv 1 \pmod 2$, and for any x we have that $x \equiv -x \pmod 2$. As such, and if a is odd, then the above implies that $n \equiv \Delta \pmod 2$. Now the claim readily follows from Lemma 22.1.25.

As for $(2 \mid p)$, setting $a = 2$, observe that $\lfloor ja/p \rfloor = 0$, for $j = 0, \ldots (p-1)/2$, and as such $\Delta = 0$. Now, Eq. (22.1) implies that $\frac{p^2-1}{8} \equiv n \pmod 2$, and the claim follows from Lemma 22.1.25. ■

Theorem 22.1.27 (*Law of quadratic reciprocity*). *If p and q are distinct odd primes, then*

$$(p \mid q) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}} (q \mid p).$$

Proof: Let $S = \Big\{(x,y) \;\Big|\; 1 \le x \le (p-1)/2 \text{ and } 1 \le y \le (q-1)/2\Big\}$. As $\operatorname{lcm}(p,q) = pq$, it follows that there are no $(x,y) \in S$, such that $qx = py$, as all such numbers are strict smaller than pq. Now, let

$$S_1 = \Big\{(x,y) \in S \;\Big|\; qx > py\Big\} \qquad \text{and} \qquad S_2 = \Big\{(x,y) \in S \;\Big|\; qx < py\Big\}.$$

Now, $(x,y) \in S_1 \iff 1 \le x \le (p-1)$, and $1 \le y \le \lfloor qx/p \rfloor$. As such, we have $|S_1| = \sum_{x=1}^{(p-1)/2} \lfloor qx/p \rfloor$, and similarly $|S_2| = \sum_{y=1}^{(q-1)/2} \lfloor py/q \rfloor$. We have

$$\tau = \frac{p-1}{2} \cdot \frac{q-1}{2} = |S| = |S_1| + |S_2| = \underbrace{\sum_{x=1}^{(p-1)/2} \lfloor qx/p \rfloor}_{\tau_1} + \underbrace{\sum_{y=1}^{(q-1)/2} \lfloor py/q \rfloor}_{\tau_2}.$$

The claim now readily follows by Lemma 22.1.26, as $(-1)^{\tau} = (-1)^{\tau_1}(-1)^{\tau_2} = (p \mid q)(q \mid p)$. ■

22.1.3.3. Jacobi symbol

Definition 22.1.28. For any integer a, and an odd number n with prime factorization $n = p_1^{k_1} \cdots p_t^{k_t}$, its *Jacobi symbol* is

$$[\![a \mid n]\!] = \prod_{i=1}^{t} (a \mid p_i)^{k_i}.$$

Claim 22.1.29. *For odd integers $n_1, \ldots, n_k$, we have that $\sum_{i=1}^{k}(n_i-1)/2 \equiv \left(\prod_{i=1}^{k} n_i - 1\right)/2 \pmod 2$.*

Proof: We prove for two odd integers x and y, and apply this repeatedly to get the claim. Indeed, we have $\frac{x-1}{2} + \frac{y-1}{2} \equiv \frac{xy-1}{2} \pmod 2 \iff 0 \equiv \frac{xy-x+1-y+1-1}{2} \pmod 2 \iff 0 \equiv \frac{xy-x-y+1}{2} \pmod 2 \iff 0 \equiv \frac{(x-1)(y-1)}{2} \pmod 2$, which is obviously true. ■

Lemma 22.1.30 (*Law of quadratic reciprocity*). *For n and m positive odd integers, we have that $[\![n \mid m]\!] = (-1)^{\frac{n-1}{2}\frac{m-1}{2}} [\![m \mid n]\!]$.*

Proof: Let $n = \prod_{i=1}^{\nu} p_i$ and Let $m = \prod_{j=1}^{\mu} q_j$ be the prime factorization of the two numbers (allowing repeated factors). If they share a common factor p, then both $[\![n \mid m]\!]$ and $[\![m \mid n]\!]$ contain a zero term when expanded, as $(n \mid p) = (m \mid p) = 0$. Otherwise, we have

$$\begin{aligned}
[\![n \mid m]\!] &= \prod_{i=1}^{\nu}\prod_{j=1}^{\mu} \left[\!\left[p_i \mid q_j \right]\!\right] = \prod_{i=1}^{\nu}\prod_{j=1}^{\mu} \left(p_i \mid q_j\right) = \prod_{i=1}^{\nu}\prod_{j=1}^{\mu} (-1)^{(q_j-1)/2\cdot(p_i-1)/2} \left(q_j \mid p_i\right) \\
&= \underbrace{\prod_{i=1}^{\nu}\prod_{j=1}^{\mu} (-1)^{(q_j-1)/2\cdot(p_i-1)/2}}_{s} \cdot \left(\prod_{i=1}^{\nu}\prod_{j=1}^{\mu} \left(q_j \mid p_i\right) \right) = s\,[\![m \mid n]\!].
\end{aligned}$$

by Theorem 22.1.27. As for the value of s, observe that

$$s = \prod_{i=1}^{\nu}\left(\prod_{j=1}^{\mu} (-1)^{(q_j-1)/2} \right)^{(p_i-1)/2} = \prod_{i=1}^{\nu}\left((-1)^{(m-1)/2}\right)^{(p_i-1)/2} = \left(\prod_{i=1}^{\nu} (-1)^{(p_i-1)/2} \right)^{(m-1)/2} = (-1)^{(n-1)/2\cdot(m-1)/2},$$

by repeated usage of Claim 22.1.29. ■

Lemma 22.1.31. *For odd integers n and m, we have that $\frac{n^2-1}{8} + \frac{m^2-1}{8} \equiv \frac{n^2m^2-1}{8} \pmod 2$.*

Proof: For an odd integer n, we have that either (i) $2 \mid n-1$ and $4 \mid n+1$, or (ii) $4 \mid n-1$ and $2 \mid n+1$. As such, $8 \mid n^2-1 = (n-1)(n+1)$. In particular, $64 \mid \left(n^2-1\right)\left(m^2-1\right)$. We thus have that

$$\begin{aligned}
\frac{\left(n^2-1\right)\left(m^2-1\right)}{8} \equiv 0 \pmod 2 &\iff \frac{n^2m^2-n^2-m^2+1}{8} \equiv 0 \pmod 2 \\
&\iff \frac{n^2m^2-1}{8} \equiv \frac{n^2+m^2-2}{8} \pmod 2 \\
&\iff \frac{n^2-1}{8} + \frac{m^2-1}{8} \equiv \frac{n^2m^2-1}{8} \pmod 2.
\end{aligned}$$

■

Lemma 22.1.32. *Let m, n be odd integers, and a, b be any integers. We have the following:*

(A) $[\![ab \mid n]\!] = [\![a \mid n]\!]\,[\![b \mid n]\!]$.
(B) $[\![a \mid nm]\!] = [\![a \mid n]\!]\,[\![a \mid m]\!]$.
(C) If $a \equiv b \pmod n$ *then* $[\![a \mid n]\!] = [\![b \mid n]\!]$.
(D) If $\gcd(a, n) > 1$ *then* $[\![a \mid n]\!] = 0$.
(E) $[\![1 \mid n]\!] = 1$.

(F) $[\![2 \mid n]\!] = (-1)^{(n^2-1)/8}$.
(G) $[\![n \mid m]\!] = (-1)^{\frac{n-1}{2}\frac{m-1}{2}} [\![m \mid n]\!]$.

Proof: (A) Follows immediately, as $(ab \mid p_i) = (a \mid p_i)(b \mid p_i)$, see Lemma 22.1.24$_{\text{p8}}$.
(B) Immediate from definition.
(C) Follows readily from Lemma 22.1.24$_{\text{p8}}$ (iii).
(D) Indeed, if $p \mid \gcd(a, n)$ and $p > 1$, then $(a \mid p)^k = (0 \mid p)^k = 0$ appears as a term in $[\![a \mid n]\!]$.
(E) Obvious by definition.
(F) By Lemma 22.1.26$_{\text{p9}}$, for a prime p, we have $(2 \mid p) = (-1)^{(p^2-1)/8}$. As such, writing $n = \prod_{i=1}^{t} p_i$ as a product of primes (allowing repeated primes), we have

$$[\![2 \mid n]\!] = \prod_{i=1}^{t} (2 \mid p_i) = \prod_{i=1}^{t} (-1)^{(p_i^2-1)/8} = (-1)^{\Delta},$$

where $\Delta = \sum_{i=1}^{t} (p_i^2 - 1)/8$. As such, we need to compute the $\Delta \pmod 2$, which by Lemma 22.1.31, is

$$\Delta \equiv \sum_{i=1}^{t} \frac{p_i^2 - 1}{8} \equiv \frac{\prod_{i=1}^{t} p_i^2 - 1}{8} \equiv \frac{n^2 - 1}{8} \pmod 2,$$

and as such $[\![2 \mid n]\!] = (-1)^{\Delta} = (-1)^{(n^2-1)/8}$.
(G) This is Lemma 22.1.30. ■

22.1.3.4. Jacobi(a, n): Computing the Jacobi symbol

Given a and n (n is an odd number), we are interested in computing (in polynomial time) the Jacobi symbol $[\![a \mid n]\!]$. The algorithm **Jacobi**(a, n) works as follows:

(A) If $a = 0$ then **return** 0 `// Since [[0 | n]] = 0.`
(B) If $a > n$ then **return** **Jacobi**$(a \pmod n, n)$ `// Lemma 22.1.32 (C)`
(C) If $\gcd(a, n) > 1$ then **return** 0 `// Lemma 22.1.32 (D)`
(D) If $a = 2$ then
  (I) Compute $\Delta = n^2 - 1 \pmod{16}$,
  (II) Return $(-1)^{\Delta/8 \pmod 2}$ `// As (n^2-1)/8 ≡ Δ/8 (mod 2), and by Lemma 22.1.32 (F)`
(E) If $2 \mid a$ then **return** **Jacobi**$(2, n)$ * **Jacobi**$(a/2, n)$ `// Lemma 22.1.32 (A)`
`// Must be that a and b are both odd, a < n, and they are coprime`
(F) $a' := a \pmod 4$, $n' := n \pmod 4$, $\beta = (a' - 1)(n' - 1)/4$.
return $(-1)^{\beta}$ **Jacobi**(n, a) `// By Lemma 22.1.32 (G)`

Ignoring the recursive calls, all the operations takes polynomial time. Clearly, computing **Jacobi**$(2, n)$ takes polynomial time. Otherwise, observe that **Jacobi** reduces its input size by say, one bit, at least every two recursive calls, and except the $a = 2$ case, it always perform only a single call. Thus, it follows that its running time is polynomial. We thus get the following.

Lemma 22.1.33. *Given integers a and n, where n is odd, then $[\![a \mid n]\!]$ can be computed in polynomial time.*

22.1.3.5. Subgroups induced by the Jacobi symbol

For an n, consider the set

$$J_n = \left\{ a \in \mathbb{Z}_n^* \;\middle|\; [\![a \mid n]\!] \equiv a^{(n-1)/2} \bmod n \right\}. \tag{22.2}$$

Claim 22.1.34. *The set J_n is a subgroup of $\mathbb{Z}_n^*$.*

Proof: For $a, b \in J_n$, we have that $[\![ab \mid n]\!] \equiv [\![a \mid n]\!]\,[\![b \mid n]\!] \equiv a^{(n-1)/2} b^{(n-1)/2} \equiv (ab)^{(n-1)/2} \bmod n$, implying that $ab \in J_n$. Now, $[\![1 \mid n]\!] = 1$, so $1 \in J_n$. Now, for $a \in J_n$, let a^{-1} the inverse of a (which is a number in $\mathbb{Z}_n^*$). Observe that $a(a^{-1}) = kn + 1$, for some k, and as such, we have

$$1 = [\![1 \mid n]\!] = [\![kn + 1 \mid n]\!] = \left[\!\left[aa^{-1} \mid n\right]\!\right] = [\![kn + 1 \mid n]\!] = [\![a \mid n]\!] \left[\!\left[a^{-1} \mid n\right]\!\right].$$

And modulo n, we have

$$1 \equiv [\![a \mid n]\!] \left[\!\left[a^{-1} \mid n\right]\!\right] \equiv a^{(n-1)/2} \left[\!\left[a^{-1} \mid n\right]\!\right] \bmod n.$$

Which implies that $\left(a^{-1}\right)^{(n-1)/2} \equiv \left[\!\left[a^{-1} \mid n\right]\!\right] \bmod n$. That is $a^{-1} \in J_n$.

Namely, J_n contains the identity, it is closed under inverse and multiplication, and it is now easy to verify that fulfill the other requirements to be a group. ■

Lemma 22.1.35. *Let n be an odd integer that is composite, then $|J_n| \leq \left|\mathbb{Z}_n^*\right|/2$.*

Proof: Let has the prime factorization $n = \prod_{i=1}^{t} p_i^{k_i}$. Let $q = p_1^{k_1}$, and $m = n/q$. By Lemma 22.1.20$_{\text{p7}}$, the group $\mathbb{Z}_q^*$ is cyclic, and let g be its generator. Consider the element $a \in \mathbb{Z}_n^*$ such that

$$a \equiv g \bmod q \qquad \text{and} \qquad a \equiv 1 \bmod m.$$

Such a number a exists and its unique, by the Chinese remainder theorem (Theorem 22.1.6$_{\text{p3}}$). In particular, let $m = \prod_{i=2}^{t} p_i^{k_i}$, and observe that, for all i, we have $a \equiv 1 \pmod{p_i}$, as $p_i \mid m$. As such, writing the Jacobi symbol explicitly, we have

$$[\![a \mid n]\!] = [\![a \mid q]\!] \prod_{i=2}^{t} (a \mid p_i)^{k_i} = [\![a \mid q]\!] \prod_{i=2}^{t} (1 \mid p_i)^{k_i} = [\![a \mid q]\!] \prod_{i=2}^{t} 1 = [\![a \mid q]\!] = [\![g \mid q]\!].$$

since $a \equiv g \pmod{q}$, and Lemma 22.1.32$_{\text{p10}}$ (C). At this point there are two possibilities:

(A) If $k_1 = 1$, then $q = p_1$, and $[\![g \mid q]\!] = (g \mid q) = g^{(q-1)/2} \pmod{q}$. But g is a generator of $\mathbb{Z}_q^*$, and its order is $q-1$. As such $g^{(q-1)/2} \equiv -1 \pmod{q}$, see Definition 22.1.23$_{\text{p8}}$. We conclude that $[\![a \mid n]\!] = -1$. If we assume that $J_n = \mathbb{Z}_n^*$, then $[\![a \mid n]\!] \equiv a^{(n-1)/2} \equiv -1 \pmod{n}$. Now, as $m \mid n$, we have

$$a^{(n-1)/2} \equiv_m \left(a^{(n-1)/2} \bmod n\right) \bmod m \equiv_m -1.$$

But this contradicts the choice of a as $a \equiv 1 \pmod{m}$.

(B) If $k_1 > 1$ then $q = p_1^{k_1}$. Arguing as above, we have that $[\![a \mid n]\!] = (-1)^{k_1}$. Thus, if we assume that $J_n = \mathbb{Z}_n^*$, then $a^{(n-1)/2} \equiv -1 \pmod{n}$ or $a^{(n-1)/2} \equiv 1 \pmod{n}$. This implies that $a^{n-1} \equiv 1 \pmod{n}$. Thus, $a^{n-1} \equiv 1 \pmod{q}$.
Now $a \equiv g \bmod q$, and thus $g^{n-1} \equiv 1 \pmod{q}$. This implies that the order of g in $\mathbb{Z}_q^*$ must divide $n-1$. That is $\operatorname{ord}(g) = \phi(q) \mid n-1$. Now, since $k_1 \geq 2$, we have that $p_1 \mid \phi(q) = \left(p_1^{k_1}\right)(p_1 - 1)$, see Lemma 22.1.8$_{\text{p3}}$. We conclude that $p_1 \mid n-1$ and $p_1 \mid n$, which is of course impossible, as $p_1 > 1$.

We conclude that J_n must be a proper subgroup of $\mathbb{Z}_n^*$, but, by Lemma 22.1.11$_{\text{p5}}$, it must be that $|J_n| \mid \left|\mathbb{Z}_n^*\right|$. But this implies that $|J_n| \leq \left|\mathbb{Z}_n^*\right|/2$. ■

22.2. Primality testing

The primality test is now easy[③]. Indeed, given a number n, first check if it is even (duh!). Otherwise, randomly pick a number $r \in \{2, \ldots, n-1\}$. If $\gcd(r, n) > 1$ then the number is composite. Otherwise, check if $r \in J_n$ (see Eq. (22.2)$_{\text{p11}}$), by computing $x = [\![r \mid n]\!]$ in polynomial time, see Section 22.1.3.4$_{\text{p11}}$, and $x' = a^{(n-1)/2} \bmod n$. (see Lemma 22.1.7$_{\text{p3}}$). If $x = x'$ then the algorithm returns is prime, otherwise it returns it is composite.

Theorem 22.2.1. *Given a number n, and a parameter $\delta > 0$, there is a randomized algorithm that, decides if the given number is prime or composite. The running time of the algorithm is $O((\log n)^c \log(1/\delta))$, where c is some constant. If the algorithm returns that n is composite then it is. If the algorithm returns that n is prime, then is wrong with probability at most δ.*

Proof: Run the above algorithm $m = O(\log(1/\delta))$ times. If any of the runs returns that it is composite then the algorithm return that n is composite, otherwise the algorithms returns that it is a prime.

If the algorithm fails, then n is a composite, and let $r_1, \ldots, r_m$ be the random numbers the algorithm picked. The algorithm fails only if $r_1, \ldots, r_m \in J_n$, but since $|J_n| \leq \left|\mathbb{Z}_n^2\right|/2$, by Lemma 22.1.35$_{\text{p12}}$, it follows that this happens with probability at most $\left(|J_n| / \left|\mathbb{Z}_n^2\right|\right)^m \leq 1/2^m \leq \delta$, as claimed. ■

22.2.1. Distribution of primes

In the following, let $\pi(n)$ denote the number of primes between 1 and n. Here, we prove that $\pi(n) = \Theta(n/\log n)$.

Lemma 22.2.2. *Let Δ be the product of all the prime numbers p, where $m < p \leq 2m$. We have that $\Delta \leq \binom{2m}{m}$.*

Proof: Let X be the product of the all composite numbers between m and $2m$, we have

$$\binom{2m}{m} = \frac{2m \cdot (2m-1) \cdots (m+2) \cdot (m+1)}{m \cdot (m-1) \cdots 2 \cdot 1} = \frac{X \cdot \Delta}{m \cdot (m-1) \cdots 2 \cdot 1}.$$

Since none of the numbers between 2 and m divides any of the factors of Δ, it must be that the number $\frac{X}{m \cdot (m-1) \cdots 2 \cdot 1}$ is an integer number, as $\binom{2m}{m}$ is an integer. Therefore, $\binom{2m}{m} = c \cdot \Delta$, for some integer $c > 0$, implying the claim. ■

Lemma 22.2.3. *The number of prime numbers between m and $2m$ is $O(m/\ln m)$.*

Proof: Let us denote all primes between m and $2m$ as $p_1 < p_2 < \cdots < p_k$. Since $p_1 \geq m$, it follows from Lemma 22.2.2 that $m^k \leq \prod_{i=1}^{k} p_i \leq \binom{2m}{m} \leq 2^{2m}$. Now, taking log of both sides, we have $k \lg m \leq 2m$. Namely, $k \leq 2m/\lg m$. ■

Lemma 22.2.4. $\pi(n) = O(n/\ln n)$.

Proof: Let the number of primes less than n be $\Pi(n)$, then by Lemma 22.2.3, there exist some positive constant C, such that for all $\forall n \geq N$, we have $\Pi(2n) - \Pi(n) \leq C \cdot n/\ln n$. Namely, $\Pi(2n) \leq C \cdot n/\ln n + \Pi(n)$. Thus, $\Pi(2n) \leq \sum_{i=0}^{\lceil \lg n \rceil} \left(\Pi\left(2n/2^i\right) - \Pi\left(2n/2^{i+1}\right)\right) \leq \sum_{i=0}^{\lceil \lg n \rceil} C \cdot \frac{n/2^i}{\ln(n/2^i)} = O\left(\frac{n}{\ln n}\right)$, by observing that the summation behaves like a decreasing geometric series. ■

[③] One could even say "trivial" with heavy Russian accent.

Lemma 22.2.5. *For integers m, k and a prime p, if $p^k \mid \binom{2m}{m}$, then $p^k \le 2m$.*

Proof: Let $T(p,m)$ be the number of times p appear in the prime factorization of $m!$. Formally, $T(p,m)$ is the highest number k such that p^k divides $m!$. We claim that $T(p,m) = \sum_{i=1}^{\infty} \left\lfloor m/p^i \right\rfloor$. Indeed, consider an integer $\beta \le m$, such that $\beta = p^t\gamma$, where γ is an integer that is not divisible by p. Observe that β contributes exactly to the first t terms of the summation of $T(p,m)$ – namely, its contribution to $m!$ as far as powers of p is counted correctly.

Let α be the maximum number such that p^α divides $\binom{2m}{m} = \frac{2m!}{m!m!}$. Clearly,

$$\alpha = T(p,2m) - 2T(p,m) = \sum_{i=1}^{\infty} \left(\left\lfloor \frac{2m}{p^i} \right\rfloor - 2\left\lfloor \frac{m}{p^i} \right\rfloor \right).$$

It is easy to verify that for any integers x, y, we have that $0 \le \left\lfloor \frac{2x}{y} \right\rfloor - 2\left\lfloor \frac{x}{y} \right\rfloor \le 1$. In particular, let k be the largest number such that $\left(\left\lfloor \frac{2m}{p^k} \right\rfloor - 2\left\lfloor \frac{m}{p^k} \right\rfloor \right) = 1$, and observe that $T(p,2m) \le k$ as only the proceedings $k-1$ terms might be non-zero in the summation of $T(p,2m)$. But this implies that $\left\lfloor 2m/p^k \right\rfloor \ge 1$, which implies in turn that $p^k \le 2m$, as desired. ■

Lemma 22.2.6. $\pi(n) = \Omega(n/\ln n)$.

Proof: Assume $\binom{2m}{m}$ have k prime factors, and thus can be written as $\binom{2m}{m} = \prod_{i=1}^{k} p_i^{n_i}$, By Lemma 22.2.5, we have $p_i^{n_i} \le 2m$. Of course, the above product might not include some prime numbers between 1 and $2m$, and as such k is a lower bound on the number of primes in this range; that is, $k \le \pi(2m)$. This implies $\frac{2^{2m}}{2m} \le \binom{2m}{m} \le \prod_{i=1}^{k} 2m = (2m)^k$. By taking lg of both sides, we have $\frac{2m - \lg(2m)}{\lg(2m)} \le k \le \pi(2m)$. ■

We summarize the result.

Theorem 22.2.7. *Let $\pi(n)$ be the number of distinct prime numbers between 1 and n. We have that $\pi(n) = \Theta(n/\ln n)$.*

22.3. Bibliographical notes

Miller [Mil76] presented the primality testing algorithm which runs in deterministic polynomial time but relies on Riemann's Hypothesis (which is still open). Later on, Rabin [Rab80] showed how to convert this algorithm to a randomized algorithm, without relying on the Riemann's hypothesis.

This write-up is based on various sources – starting with the description in [MR95], and then filling in some details from various sources on the web.

What is currently missing from the write-up is a description of the RSA encryption system. This would hopefully be added in the future. There are of course typos in these notes – let me know if you find any.

Bibliography

[Mil76] G. L. Miller. Riemann's hypothesis and tests for primality. *J. Comput. Sys. Sci.*, 13(3):300–317, 1976.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, UK, 1995.

[Rab80] M. O. Rabin. Probabilistic algorithm for testing primality. *J. Number Theory*, 12(1):128–138, 1980.

Chapter 23

Finite Metric Spaces and Partitions

By Sariel Har-Peled, December 30, 2015[①]

23.1. Finite Metric Spaces

Definition 23.1.1. A *metric space* is a pair $(\mathcal{X}, \mathsf{d})$ where $\mathcal{X}$ is a set and $\mathsf{d} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a *metric*, satisfying the following axioms: (i) $\mathsf{d}(x, y) = 0$ iff $x = y$, (ii) $\mathsf{d}(x, y) = \mathsf{d}(y, x)$, and (iii) $\mathsf{d}(x, y) + \mathsf{d}(y, z) \geq \mathsf{d}(x, z)$ (triangle inequality).

For example, $\mathbb{R}^2$ with the regular Euclidean distance is a metric space.

It is usually of interest to consider the finite case, where $\mathcal{X}$ is an n-point set. Then, the function d can be specified by $\binom{n}{2}$ real numbers. Alternatively, one can think about $(\mathcal{X}, \mathsf{d})$ is a weighted complete graph, where we specify positive weights on the edges, and the resulting weights on the edges comply with the triangle inequality.

In fact, finite metric spaces rise naturally from (sparser) graphs. Indeed, let $G = (\mathcal{X}, E)$ be an undirected weighted graph defined over $\mathcal{X}$, and let $\mathsf{d}_G(x, y)$ be the length of the shortest path between x and y in G. It is easy to verify that $(\mathcal{X}, \mathsf{d}_G)$ is a finite metric space. As such if the graph G is sparse, it provides a compact representation to the finite space $(\mathcal{X}, \mathsf{d}_G)$.

Definition 23.1.2. Let $(\mathcal{X}, d)$ be an n-point metric space. We denote the *open ball* of radius r about $x \in \mathcal{X}$, by $\mathbf{b}(x, r) = \left\{ y \in \mathcal{X} \;\middle|\; \mathsf{d}(x, y) < r \right\}$.

Underling our discussion of metric spaces are algorithmic applications. The hardness of various computational problems depends heavily on the structure of the finite metric space. Thus, given a finite metric space, and a computational task, it is natural to try to map the given metric space into a new metric where the task at hand becomes easy.

Example 23.1.3. For example, computing the diameter is not trivial in two dimensions, but is easy in one dimension. Thus, if we could map points in two dimensions into points in one dimension, such that the diameter is preserved, then computing the diameter becomes easy. In fact, this approach yields an efficient approximation algorithm, see Exercise 23.7.3 below.

Of course, this mapping from one metric space to another, is going to introduce error. We would be interested in minimizing the error introduced by such a mapping.

Definition 23.1.4. Let $(\mathcal{X}, \mathsf{d}_\mathcal{X})$ and (Y, d_Y) be metric spaces. A mapping $f : \mathcal{X} \to Y$ is called an *embedding*, and is *C-Lipschitz* if $\mathsf{d}_Y(f(x), f(y)) \leq C \cdot \mathsf{d}_\mathcal{X}(x, y)$ for all $x, y \in \mathcal{X}$. The mapping f is called *K-bi-Lipschitz* if there exists a $C > 0$ such that

$$CK^{-1} \cdot \mathsf{d}_\mathcal{X}(x, y) \leq \mathsf{d}_Y(f(x), f(y)) \leq C \cdot \mathsf{d}_\mathcal{X}(x, y),$$

for all $x, y \in \mathcal{X}$.

The least K for which f is K-bi-Lipschitz is called the *distortion* of f, and is denoted $\operatorname{dist}(f)$. The least distortion with which $\mathcal{X}$ may be embedded in Y is denoted $c_Y(\mathcal{X})$.

There are several powerful results in this vain, that show the existence of embeddings with low distortion that would be presented:

1. Probabilistic trees - every finite metric can be randomly embedded into a tree such that the "expected" distortion for a specific pair of points is $O(\log n)$.

2. Bourgain embedding - shows that any n-point metric space can be embedded into (finite dimensional) metric space with $O(\log n)$ distortion.

3. Johnson-Lindenstrauss lemma - shows that any n-point set in Euclidean space with the regular Euclidean distance can be embedded into $\mathbb{R}^k$ with distortion $(1 + \varepsilon)$, where $k = O(\varepsilon^{-2} \log n)$.

23.2. Examples

23.2.0.0.1. What is distortion? When considering a mapping $f : \mathcal{X} \to \mathbb{R}^d$ of a metric space $(\mathcal{X}, \mathsf{d})$ to $\mathbb{R}^d$, it would useful to observe that since $\mathbb{R}^d$ can be scaled, we can consider f to be an an expansion (i.e., no distances shrink). Furthermore, we can in fact assume that there is at least one pair of points $x, y \in \mathcal{X}$, such that $\mathsf{d}(x, y) = \|x - y\|$. As such, we have $\operatorname{dist}(f) = \max_{x,y} \frac{\|x-y\|}{\mathsf{d}(x,y)}$.

23.2.0.0.2. Why distortion is necessary? Consider the a graph $G = (V, E)$ with one vertex s connected to three other vertices a, b, c, where the weights on the edges are all one (i.e., G is the star graph with three leafs). We claim that G can not be embedded into Euclidean space with distortion $\leq \sqrt{2}$. Indeed, consider the associated metric space (V, d_G) and an (expansive) embedding $f : V \to \mathbb{R}^d$.

Consider the triangle formed by $\triangle = a'b'c'$, where $a' = f(a), b' = f(b)$ and $c' = f(c)$. Next, consider the following quantity $\max(\|a' - s'\|, \|b' - s'\|, \|c' - s'\|)$ which lower bounds the distortion of f. This quantity is minimized when $r = \|a' - s'\| = \|b' - s'\| = \|c' - s'\|$. Namely, s' is the center of the smallest enclosing circle of $\triangle$. However, r is minimize when all the edges of $\triangle$ are of equal length, and are in fact of length $\mathsf{d}_G(a, b) = 2$. It follows that $\operatorname{dist}(f) \geq r \geq 2/\sqrt{3}$.

It is known that $\Omega(\log n)$ distortion is necessary in the worst case. This is shown using expanders [Mat02].

23.2.1. Hierarchical Tree Metrics

The following metric is quite useful in practice, and nicely demonstrate why algorithmically finite metric spaces are useful.

Definition 23.2.1. *Hierarchically well-separated tree* (HST) is a metric space defined on the leaves of a rooted tree T. To each vertex $u \in T$ there is associated a label $\Delta_u \geq 0$ such that $\Delta_u = 0$ if and only if u is a leaf of T. The labels are such that if a vertex u is a child of a vertex v then $\Delta_u \leq \Delta_v$. The distance between two leaves $x, y \in T$ is defined as $\Delta_{\mathrm{lca}(x,y)}$, where $\mathrm{lca}(x, y)$ is the least common ancestor of x and y in T.

A HST T is a *k-HST* if for a vertex $v \in T$, we have that $\Delta_v \leq \Delta_{\overline{\mathrm{p}}(v)}/k$, where $\overline{\mathrm{p}}(v)$ is the parent of v in T.

Note that a HST is a very limited metric. For example, consider the cycle $G = C_n$ of n vertices, with weight one on the edges, and consider an expansive embedding f of G into a HST HST. It is easy to verify, that there must be two consecutive nodes of the cycle, which are mapped to two different subtrees of the root r of HST. Since HST is expansive, it follows that $\Delta_r \geq n/2$. As such, $\mathrm{dist}(f) \geq n/2$. Namely, HSTs fail to faithfully represent even very simple metrics.

23.2.2. Clustering

One natural problem we might want to solve on a graph (i.e., finite metric space) $(\mathcal{X}, \mathsf{d})$ is to partition it into clusters. One such natural clustering is the *k-median clustering*, where we would like to choose a set $C \subseteq \mathcal{X}$ of k centers, such that $\nu_C(\mathcal{X}, \mathsf{d}) = \sum_{\mathsf{q} \in \mathcal{X}} \mathsf{d}(\mathsf{q}, C)$ is minimized, where $\mathsf{d}(\mathsf{q}, C) = \min_{c \in C} \mathsf{d}(\mathsf{q}, c)$ is the distance of q to its closest center in C.

It is known that finding the optimal k-median clustering in a (general weighted) graph is NP-complete. As such, the best we can hope for is an approximation algorithm. However, if the structure of the finite metric space $(\mathcal{X}, \mathsf{d})$ is simple, then the problem can be solved efficiently. For example, if the points of $\mathcal{X}$ are on the real line (and the distance between a and b is just $|a - b|$), then k-median can be solved using dynamic programming.

Another interesting case is when the metric space $(\mathcal{X}, \mathsf{d})$ is a HST. Is not too hard to prove the following lemma. See Exercise 23.7.1.

Lemma 23.2.2. *Let $(\mathcal{X}, \mathsf{d})$ be a HST defined over n points, and let $k > 0$ be an integer. One can compute the optimal k-median clustering of $\mathcal{X}$ in $O(k^2 n)$ time.*

Thus, if we can embed a general graph G into a HST HST, with low distortion, then we could approximate the k-median clustering on G by clustering the resulting HST, and "importing" the resulting partition to the original space. The quality of approximation, would be bounded by the distortion of the embedding of G into HST.

23.3. Random Partitions

Let $(\mathcal{X}, d)$ be a finite metric space. Given a partition $P = \{C_1, \ldots, C_m\}$ of $\mathcal{X}$, we refer to the sets C_i as *clusters*. We write $\mathcal{P}_{\mathcal{X}}$ for the set of all partitions of $\mathcal{X}$. For $x \in \mathcal{X}$ and a partition $P \in \mathcal{P}_{\mathcal{X}}$ we denote by $P(x)$ the unique cluster of P containing x. Finally, the set of all probability distributions on $\mathcal{P}_{\mathcal{X}}$ is denoted $\mathcal{D}_{\mathcal{X}}$.

23.3.1. Constructing the partition

Let $\Delta = 2^u$ be a prescribed parameter, which is the required diameter of the resulting clusters. Choose, uniformly at random, a permutation π of $\mathcal{X}$ and a random value $\alpha \in [1/4, 1/2]$. Let $R = \alpha\Delta$, and observe that it is uniformly distributed in the interval $[\Delta/4, \Delta/2]$.

The partition is now defined as follows: A point $x \in \mathcal{X}$ is assigned to the cluster C_y of y, where y is the first point in the permutation in distance $\leq R$ from x. Formally,

$$C_y = \left\{ x \in \mathcal{X} \;\middle|\; x \in \mathbf{b}(y, R) \text{ and } \pi(y) \leq \pi(z) \text{ for all } z \in \mathcal{X} \text{ with } x \in \mathbf{b}(z, R) \right\}.$$

Let $P = \{C_y\}_{y \in \mathcal{X}}$ denote the resulting partition.

Here is a somewhat more intuitive explanation: Once we fix the radius of the clusters R, we start scooping out balls of radius R centered at the points of the random permutation π. At the ith stage, we scoop out only the remaining mass at the ball centered at x_i of radius r, where x_i is the ith point in the random permutation.

23.3.2. Properties

Lemma 23.3.1. *Let $(\mathcal{X}, d)$ be a finite metric space, $\Delta = 2^u$ a prescribed parameter, and let P be the partition of $\mathcal{X}$ generated by the above random partition. Then the following holds:*

(i) For any $C \in P$, we have $\operatorname{diam}(C) \leq \Delta$.

(ii) Let x be any point of $\mathcal{X}$, and t a parameter $\leq \Delta/8$. Then,

$$\mathbf{Pr}[\mathbf{b}(x,t) \not\subseteq P(x)] \leq \frac{8t}{\Delta} \ln \frac{b}{a},$$

where $a = |\mathbf{b}(x, \Delta/8)|$, $b = |\mathbf{b}(x, \Delta)|$.

Proof: Since $C_y \subseteq \mathbf{b}(y, R)$, we have that $\operatorname{diam}(C_y) \leq \Delta$, and thus the first claim holds.

Let U be the set of points of $\mathbf{b}(x, \Delta)$, such that $w \in U$ iff $\mathbf{b}(w, R) \cap \mathbf{b}(x, t) \neq \emptyset$. Arrange the points of U in increasing distance from x, and let $w_1, \ldots, w_{b'}$ denote the resulting order, where $b' = |U|$. Let $I_k = [d(x, w_k) - t, d(x, w_k) + t]$ and write $\mathcal{E}_k$ for the event that w_k is the first point in π such that $\mathbf{b}(x, t) \cap C_{w_k} \neq \emptyset$, and yet $\mathbf{b}(x, t) \not\subseteq C_{w_k}$. Note that if $w_k \in \mathbf{b}(x, \Delta/8)$, then $\mathbf{Pr}[\mathcal{E}_k] = 0$ since $\mathbf{b}(x, t) \subseteq \mathbf{b}(x, \Delta/8) \subseteq \mathbf{b}(w_k, \Delta/4) \subseteq \mathbf{b}(w_k, R)$.

In particular, $w_1, \ldots, w_a \in \mathbf{b}(x, \Delta/8)$ and as such $\mathbf{Pr}[\mathcal{E}_1] = \cdots = \mathbf{Pr}[\mathcal{E}_a] = 0$. Also, note that if $\mathsf{d}(x, w_k) < R - t$ then $\mathbf{b}(w_k, R)$ contains $\mathbf{b}(x, t)$ and as such $\mathcal{E}_k$ can not happen. Similarly, if $\mathsf{d}(x, w_k) > R + t$ then $\mathbf{b}(w_k, R) \cap \mathbf{b}(x, t) = \emptyset$ and $\mathcal{E}_k$ can not happen. As such, if $\mathcal{E}_k$ happen then $R - t \leq \mathsf{d}(x, w_k) \leq R + t$. Namely, if $\mathcal{E}_k$ happen then $R \in I_k$. Namely, $\mathbf{Pr}[\mathcal{E}_k] = \mathbf{Pr}[\mathcal{E}_k \cap (R \in I_k)] = \mathbf{Pr}[R \in I_k] \cdot \mathbf{Pr}[\mathcal{E}_k \,|\, R \in I_k]$. Now, R is uniformly distributed in the interval $[\Delta/4, \Delta/2]$, and I_k is an interval of length $2t$. Thus, $\mathbf{Pr}[R \in I_k] \leq 2t/(\Delta/4) = 8t/\Delta$.

Next, to bound $\mathbf{Pr}[\mathcal{E}_k \,|\, R \in I_k]$, we observe that $w_1, \ldots, w_{k-1}$ are closer to x than w_k and their distance to $\mathbf{b}(x, t)$ is smaller than R. Thus, if any of them appear before w_k in π then $\mathcal{E}_k$ does not happen. Thus, $\mathbf{Pr}[\mathcal{E}_k \,|\, R \in I_k]$ is bounded by the probability that w_k is the first to appear in π out of $w_1, \ldots, w_k$. But this probability is $1/k$, and thus $\mathbf{Pr}[\mathcal{E}_k \,|\, R \in I_k] \leq 1/k$.

We are now ready for the kill. Indeed,

$$\begin{aligned}
\mathbf{Pr}[\mathbf{b}(x,t) \not\subseteq P(x)] &= \sum_{k=1}^{b'} \mathbf{Pr}[\mathcal{E}_k] = \sum_{k=a+1}^{b'} \mathbf{Pr}[\mathcal{E}_k] = \sum_{k=a+1}^{b'} \mathbf{Pr}[R \in I_k] \cdot \mathbf{Pr}[\mathcal{E}_k \,|\, R \in I_k] \\
&\leq \sum_{k=a+1}^{b'} \frac{8t}{\Delta} \cdot \frac{1}{k} \leq \frac{8t}{\Delta} \ln \frac{b'}{a} \leq \frac{8t}{\Delta} \ln \frac{b}{a},
\end{aligned}$$

since $\sum_{k=a+1}^{b} \frac{1}{k} \leq \int_a^b \frac{dx}{x} = \ln \frac{b}{a}$ and $b' \leq b$. ∎

23.4. Probabilistic embedding into trees

In this section, given n-point finite metric $(\mathcal{X}, \mathsf{d})$. we would like to embed it into a HST. As mentioned above, one can verify that for any embedding into HST, the distortion in the worst case is $\Omega(n)$. Thus, we define

a randomized algorithm that embed $(\mathcal{X}, d)$ into a tree. Let T be the resulting tree, and consider two points $x, y \in \mathcal{X}$. Consider the *random variable* $\mathsf{d}_T(x, y)$. We constructed the tree T such that distances never shrink; i.e. $\mathsf{d}(x, y) \leq \mathsf{d}_T(x, y)$. The *probabilistic distortion* of this embedding is $\max_{x,y} \mathbf{E}\left[\frac{\mathsf{d}_T(x,y)}{\mathsf{d}(x,y)}\right]$. Somewhat surprisingly, one can find such an embedding with logarithmic probabilistic distortion.

Theorem 23.4.1. *Given n-point metric $(\mathcal{X}, d)$ one can randomly embed it into a 2-HST with probabilistic distortion $\leq 24 \ln n$.*

Proof: The construction is recursive. Let $\operatorname{diam}(P)$, and compute a random partition of $\mathcal{X}$ with cluster diameter $\operatorname{diam}(P)/2$, using the construction of Section 23.3.1. We recursively construct a 2-HST for each cluster, and hang the resulting clusters on the root node v, which is marked by $\Delta_v = \operatorname{diam}(P)$. Clearly, the resulting tree is a 2-HST.

For a node $v \in T$, let $\mathcal{X}(v)$ be the set of points of $\mathcal{X}$ contained in the subtree of v.

For the analysis, assume $\operatorname{diam}(P) = 1$, and consider two points $x, y \in \mathcal{X}$. We consider a node $v \in T$ to be in level i if $\operatorname{level}(v) = \lceil \lg \Delta_v \rceil = i$. The two points x and y correspond to two leaves in T, and let $\widehat{u}$ be the least common ancestor of x and y in t. We have $\mathsf{d}_T(x, y) \leq 2^{\operatorname{level}(v)}$. Furthermore, note that along a path the levels are strictly monotonically increasing.

In fact, we are going to be conservative, and let w be the first ancestor of x, such that $\mathbf{b} = \mathbf{b}(x, \mathsf{d}(x, y))$ is not completely contained in $\mathcal{X}(u_1), \ldots, \mathcal{X}(u_m)$, where $u_1, \ldots, u_m$ are the children of w. Clearly, $\operatorname{level}(w) > \operatorname{level}(\widehat{u})$. Thus, $\mathsf{d}_T(x, y) \leq 2^{\operatorname{level}(w)}$.

Consider the path σ from the root of T to x, and let $\mathcal{E}_i$ be the event that $\mathbf{b}$ is not fully contained in $\mathcal{X}(v_i)$, where v_i is the node of σ of level i (if such a node exists). Furthermore, let Y_i be the indicator variable which is 1 if $\mathcal{E}_i$ is the first to happened out of the sequence of events $\mathcal{E}_0, \mathcal{E}_{-1}, \ldots$. Clearly, $\mathsf{d}_T(x, y) \leq \sum Y_i 2^i$.

Let $t = \mathsf{d}(x, y)$ and $j = \lfloor \lg \mathsf{d}(x, y) \rfloor$, and $n_i = \left|\mathbf{b}(x, 2^i)\right|$ for $i = 0, \ldots, -\infty$. We have

$$\mathbf{E}[\mathsf{d}_T(x, y)] \leq \sum_{i=j}^{0} \mathbf{E}[Y_i]\, 2^i \leq \sum_{i=j}^{0} 2^i \mathbf{Pr}\left[\mathcal{E}_i \cap \overline{\mathcal{E}_{i-1}} \cap \overline{\mathcal{E}_{i-1}} \cdots \overline{\mathcal{E}_0}\right] \leq \sum_{i=j}^{0} 2^i \cdot \frac{8t}{2^i} \ln \frac{n_i}{n_{i-3}},$$

by Lemma 23.3.1. Thus,

$$\mathbf{E}[\mathsf{d}_T(x, y)] \leq 8t \ln\left(\prod_{i=j}^{0} \frac{n_i}{n_{i-3}}\right) \leq 8t \ln(n_0 \cdot n_1 \cdot n_2) \leq 24t \ln n.$$

It thus follows, that the expected distortion for x and y is $\leq 24 \ln n$. ∎

23.4.1. Application: approximation algorithm for k-median clustering

Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space, and let k be an integer number. We would like to compute the optimal k-median clustering. Number, find a subset $C_{\text{opt}} \subseteq \mathcal{X}$, such that $\nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d})$ is minimized, see Section 23.2.2. To this end, we randomly embed $(\mathcal{X}, \mathsf{d})$ into a HST HST using Theorem 23.4.1. Next, using Lemma 23.2.2, we compute the optimal k-median clustering of HST. Let C be the set of centers computed. We return C together with the partition of $\mathcal{X}$ it induces as the required clustering.

Theorem 23.4.2. *Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space. One can compute in polynomial time a k-median clustering of $\mathcal{X}$ which has expected price $O(\alpha \log n)$, where α is the price of the optimal k-median clustering of $(\mathcal{X}, \mathsf{d})$.*

Proof: The algorithm is described above, and the fact that its running time is polynomial can be easily be verified. To prove the bound on the quality of the clustering, for any point $p \in \mathcal{X}$, let $center(p)$ denote the closest point in C_{opt} to p according to d, where C_{opt} is the set of k-medians in the optimal clustering. Let C be the set of k-medians returned by the algorithm, and let HST be the HST used by the algorithm. We have

$$\beta = \nu_C(\mathcal{X}, \mathsf{d}) \le \nu_C(\mathcal{X}, \mathsf{d}_{\mathsf{HST}}) \le \nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d}_{\mathsf{HST}}) \le \sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, C_{\text{opt}}) \le \sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, center(p)).$$

Thus, in expectation we have

$$\begin{aligned} \mathbf{E}[\beta] &= \mathbf{E}\left[\sum_{p \in \mathcal{X}} \mathsf{d}_{\mathsf{HST}}(p, center(p))\right] = \sum_{p \in \mathcal{X}} \mathbf{E}[\mathsf{d}_{\mathsf{HST}}(p, center(p))] = \sum_{p \in \mathcal{X}} O(\mathsf{d}(p, center(p)) \log n) \\ &= O\left((\log n) \sum_{p \in \mathcal{X}} \mathsf{d}(p, center(p))\right) = O\Big(\nu_{C_{\text{opt}}}(\mathcal{X}, \mathsf{d}) \log n\Big), \end{aligned}$$

by linearity of expectation and Theorem 23.4.1. ∎

23.5. Embedding any metric space into Euclidean space

Lemma 23.5.1. *Let $(\mathcal{X}, \mathsf{d})$ be a metric, and let $Y \subset \mathcal{X}$. Consider the mapping $f : \mathcal{X} \to \mathbb{R}$, where $f(x) = \mathsf{d}(x, Y) = \min_{y \in Y} \mathsf{d}(x, y)$. Then for any $x, y \in \mathcal{X}$, we have $|f(x) - f(y)| \le \mathsf{d}(x, y)$. Namely f is nonexpansive.*

Proof: Indeed, let x' and y' be the closet points of Y, to x and y, respectively. Observe that $f(x) = \mathsf{d}(x, x') \le \mathsf{d}(x, y') \le \mathsf{d}(x, y) + \mathsf{d}(y, y') = \mathsf{d}(x, y) + f(y)$ by the triangle inequality. Thus, $f(x) - f(y) \le \mathsf{d}(x, y)$. By symmetry, we have $f(y) - f(x) \le \mathsf{d}(x, y)$. Thus, $|f(x) - f(y)| \le \mathsf{d}(x, y)$. ∎

23.5.1. The bounded spread case

Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric. The *spread* of $\mathcal{X}$, denoted by $\Phi(\mathcal{X}) = \frac{\operatorname{diam}(\mathcal{X})}{\min_{x,y \in \mathcal{X}, x \neq y} \mathsf{d}(x,y)}$, is the ratio between the diameter of $\mathcal{X}$ and the distance between the closest pair of points.

Theorem 23.5.2. *Given a n-point metric $\mathcal{Y} = (\mathcal{X}, d)$, with spread Φ, one can embed it into Euclidean space $\mathbb{R}^k$ with distortion $O(\sqrt{\ln \Phi} \ln n)$, where $k = O(\ln \Phi \ln n)$.*

Proof: Assume that $\operatorname{diam}(\mathcal{Y}) = \Phi$ (i.e., the smallest distance in $\mathcal{Y}$ is 1), and let $r_i = 2^{i-2}$, for $i = 1, \ldots, \alpha$, where $\alpha = \lceil \lg \Phi \rceil$. Let $P_{i,j}$ be a random partition of P with diameter r_i, using Theorem 23.4.1, for $i = 1, \ldots, \alpha$ and $j = 1, \ldots, \beta$, where $\beta = \lceil c \log n \rceil$ and c is a large enough constant to be determined shortly.

For each cluster of $P_{i,j}$ randomly toss a coin, and let $V_{i,j}$ be the all the points of $\mathcal{X}$ that belong to clusters in $P_{i,j}$ that got 'T' in their coin toss. For a point $u \in x$, let $f_{i,j}(x) = \mathsf{d}(x, \mathcal{X} \setminus V_{i,j}) = \min_{v \in \mathcal{X} \setminus V_{i,j}} \mathsf{d}(x, v)$, for $i = 0, \ldots, m$ and $j = 1, \ldots, \beta$. Let $F : \mathcal{X} \to \mathbb{R}^{(m+1)\cdot\beta}$ be the embedding, such that $F(x) = (f_{0,1}(x), f_{0,2}(x), \ldots, f_{0,\beta}(x), f_{1,1}(x), f_{0,2}(x), \ldots, f_{1,\beta}(x), \ldots, f_{m,1}(x), f_{m,2}(x), \ldots, f_{m,\beta}(x))$.

Next, consider two points $x, y \in \mathcal{X}$, with distance $\phi = \mathsf{d}(x, y)$. Let k be an integer such that $r_u \le \phi/2 \le r_{u+1}$. Clearly, in any partition of $P_{u,1}, \ldots, P_{u,\beta}$ the points x and y belong to different clusters. Furthermore, with probability half $x \in V_{u,j}$ and $y \notin V_{u,j}$ or $x \notin V_{u,j}$ and $y \in V_{u,j}$, for $1 \le j \le \beta$.

Let $\mathcal{E}_j$ denote the event that $\mathbf{b}(x, \rho) \subseteq V_{u,j}$ and $y \notin V_{u,j}$, for $j = 1, \ldots, \beta$, where $\rho = \phi/(64 \ln n)$. By Lemma 23.3.1, we have

$$\mathbf{Pr}\Big[\mathbf{b}(x, \rho) \not\subseteq P_{u,j}(x)\Big] \le \frac{8\rho}{r_u} \ln n \le \frac{\phi}{8 r_u} \le 1/2.$$

Thus,

$$\begin{aligned}\mathbf{Pr}\Big[\mathcal{E}_j\Big] &= \mathbf{Pr}\Big[\Big(\mathbf{b}(x,\rho)\subseteq P_{u,j}(x)\Big)\cap\Big(x\in V_{u,j}\Big)\cap\Big(y\notin V_{u,j}\Big)\Big]\\ &= \mathbf{Pr}\Big[\mathbf{b}(x,\rho)\subseteq P_{u,j}(x)\Big]\cdot\mathbf{Pr}\Big[x\in V_{u,j}\Big]\cdot\mathbf{Pr}\Big[y\notin V_{u,j}\Big]\geq 1/8,\end{aligned}$$

since those three events are independent. Notice, that if $\mathcal{E}_j$ happens, than $f_{u,j}(x)\geq\rho$ and $f_{u,j}(y)=0$.

Let X_j be an indicator variable which is 1 if $\mathcal{E}_i$ happens, for $j=1,\ldots,\beta$. Let $Z=\sum_j X_j$, and we have $\mu=\mathbf{E}[Z]=\mathbf{E}\Big[\sum_j X_j\Big]\geq\beta/8$. Thus, the probability that only $\beta/16$ of $\mathcal{E}_1,\ldots,\mathcal{E}_\beta$ happens, is $\mathbf{Pr}[Z<(1-1/2)\,\mathbf{E}[Z]]$. By the Chernoff inequality, we have $\mathbf{Pr}[Z<(1-1/2)\,\mathbf{E}[Z]]\leq\exp\Big(-\mu 1/(2\cdot 2^2)\Big)=\exp(-\beta/64)\leq 1/n^{10}$, if we set $c=640$.

Thus, with high probability

$$\|F(x)-F(y)\|\geq\sqrt{\sum_{j=1}^{\beta}\Big(f_{u,j}(x)-f_{u,j}(y)\Big)^2}\geq\sqrt{\rho^2\frac{\beta}{16}}=\sqrt{\beta}\frac{\rho}{4}=\phi\cdot\frac{\sqrt{\beta}}{256\ln n}.$$

On the other hand, $\Big|f_{i,j}(x)-f_{i,j}(y)\Big|\leq\mathsf{d}(x,y)=\phi\leq 64\rho\ln n$. Thus,

$$\|F(x)-F(y)\|\leq\sqrt{\alpha\beta(64\rho\ln n)^2}\leq 64\sqrt{\alpha\beta}\rho\ln n=\sqrt{\alpha\beta}\cdot\phi.$$

Thus, setting $G(x)=F(x)\frac{256\ln n}{\sqrt{\beta}}$, we get a mapping that maps two points of distance ϕ from each other to two points with distance in the range $\Big[\phi,\phi\cdot\sqrt{\alpha\beta}\cdot\frac{256\ln n}{\sqrt{\beta}}\Big]$. Namely, $G(\cdot)$ is an embedding with distortion $O(\sqrt{\alpha}\ln n)=O(\sqrt{\ln\Phi}\ln n)$.

The probability that G fails on one of the pairs, is smaller than $(1/n^{10})\cdot\binom{n}{2}<1/n^8$. In particular, we can check the distortion of G for all $\binom{n}{2}$ pairs, and if any of them fail (i.e., the distortion is too big), we restart the process. ■

23.5.2. The unbounded spread case

Our next task, is to extend Theorem 23.5.2 to the case of unbounded spread. Indeed, let $(\mathcal{X},d)$ be a n-point metric, such that $\operatorname{diam}(\mathcal{X})\leq 1/2$. Again, we look on the different resolutions $r_1,r_2,\ldots$, where $r_i=1/2^{i-1}$. For each one of those resolutions r_i, we can embed this resolution into β coordinates, as done for the bounded case. Then we concatenate the coordinates together.

There are two problems with this approach: (i) the number of resulting coordinates is infinite, and (ii) a pair x,y, might be distorted a "lot" because it contributes to all resolutions, not only to its "relevant" resolutions.

Both problems can be overcome with careful tinkering. Indeed, for a resolution r_i, we are going to modify the metric, so that it ignores short distances (i.e., distances $\leq r_i/n^2$). Formally, for each resolution r_i, let $G_i=(\mathcal{X},\widehat{E_i})$ be the graph where two points x and y are connected if $\mathsf{d}(x,y)\leq r_i/n^2$. Consider a connected component $C\in G_i$. For any two points $x,y\in C$, we have $\mathsf{d}(x,y)\leq n(r_i/n^2)\leq r_i/n$. Let $\mathcal{X}_i$ be the set of connected components of G_i, and define the distances between two connected components $C,C'\in\mathcal{X}_i$, to be $\mathsf{d}_i(C,C')=\mathsf{d}(C,C')=\min_{c\in C,c'\in C'}\mathsf{d}(c,c')$.

It is easy to verify that $(\mathcal{X}_i,\mathsf{d}_i)$ is a metric space (see Exercise 23.7.2). Furthermore, we can naturally embed $(\mathcal{X},\mathsf{d})$ into $(\mathcal{X}_i,\mathsf{d}_i)$ by mapping a point $x\in\mathcal{X}$ to its connected components in $\mathcal{X}_i$. Essentially $(\mathcal{X}_i,\mathsf{d}_i)$ is a snapped version of the metric $(\mathcal{X},d)$, with the advantage that $\Phi((\mathcal{X},\mathsf{d}_i))=O(n^2)$. We now embed $\mathcal{X}_i$ into $\beta=O(\log n)$ coordinates. Next, for any point of $\mathcal{X}$ we embed it into those β coordinates, by using the embedding of its connected component in $\mathcal{X}_i$. Let E_i be the embedding for resolution r_i. Namely, $E_i(x)=$

$(f_{i,1}(x), f_{i,2}(x), \ldots, f_{i,\beta}(x))$, where $f_{i,j}(x) = \min(\mathsf{d}_i(x, \mathcal{X} \setminus V_{i,j}), 2r_i)$. The resulting embedding is $F(x) = \oplus E_i(x) = (E_1(x), E_2(x), \ldots,)$.

Since we slightly modified the definition of $f_{i,j}(\cdot)$, we have to show that $f_{i,j}(\cdot)$ is nonexpansive. Indeed, consider two points $x, y \in \mathcal{X}_i$, and observe that

$$\left|f_{i,j}(x) - f_{i,j}(y)\right| \leq \left|\mathsf{d}_i(x, V_{i,j}) - \mathsf{d}_i(y, V_{i,j})\right| \leq \mathsf{d}_i(x, y) \leq \mathsf{d}(x, y),$$

as a simple case analysis[②] shows.

For a pair $x, y \in \mathcal{X}$, and let $\phi = \mathsf{d}(x, y)$. To see that $F(\cdot)$ is the required embedding (up to scaling), observe that, by the same argumentation of Theorem 23.5.2, we have that with high probability

$$\|F(x) - F(y)\| \geq \phi \cdot \frac{\sqrt{\beta}}{256 \ln n}.$$

To get an upper bound on this distance, observe that for i such that $r_i > \phi n^2$, we have $E_i(x) = E_i(y)$. Thus,

$$\begin{aligned}
\|F(x) - F(y)\|^2 &= \sum_i \|E_i(x) - E_i(y)\|^2 = \sum_{i, r_i < \phi n^2} \|E_i(x) - E_i(y)\|^2 \\
&= \sum_{i, \phi/n^2 < r_i < \phi n^2} \|E_i(x) - E_i(y)\|^2 + \sum_{i, r_i < \phi/n^2} \|E_i(x) - E_i(y)\|^2 \\
&= \beta\phi^2 \lg\left(n^4\right) + \sum_{i, r_i < \phi/n^2} (2r_i)^2 \beta \leq 4\beta\phi^2 \lg n + \frac{4\phi^2\beta}{n^4} \leq 5\beta\phi^2 \lg n.
\end{aligned}$$

Thus, $\|F(x) - F(y)\| \leq \phi\sqrt{5\beta \lg n}$. We conclude, that with high probability, $F(\cdot)$ is an embedding of $\mathcal{X}$ into Euclidean space with distortion $\left(\phi\sqrt{5\beta \lg n}\right) / \left(\phi \cdot \frac{\sqrt{\beta}}{256 \ln n}\right) = O(\log^{3/2} n)$.

We still have to handle the infinite number of coordinates problem. However, the above proof shows that we care about a resolution r_i (i.e., it contributes to the estimates in the above proof) only if there is a pair x and y such that $r_i/n^2 \leq \mathsf{d}(x, y) \leq r_i n^2$. Thus, for every pair of distances there are $O(\log n)$ relevant resolutions. Thus, there are at most $\eta = O(n^2\beta \log n) = O(n^2 \log^2 n)$ relevant coordinates, and we can ignore all the other coordinates. Next, consider the affine subspace h that spans $F(P)$. Clearly, it is $n - 1$ dimensional, and consider the projection $G : \mathbb{R}^\eta \to \mathbb{R}^{n-1}$ that projects a point to its closest point in h. Clearly, $G(F(\cdot))$ is an embedding with the same distortion for P, and the target space is of dimension $n - 1$.

Note, that all this process succeeds with high probability. If it fails, we try again. We conclude:

Theorem 23.5.3 (Low quality Bourgain theorem.). *Given a n-point metric M, one can embed it into Euclidean space of dimension $n - 1$, such that the distortion of the embedding is at most $O(\log^{3/2} n)$.*

Using the Johnson-Lindenstrauss lemma, the dimension can be further reduced to $O(\log n)$. In fact, being more careful in the proof, it is possible to reduce the dimension to $O(\log n)$ directly.

23.6. Bibliographical notes

The partitions we use are due to Calinescu *et al.* [CKR01]. The idea of embedding into spanning trees is due to Alon *et al.* [AKPW95], which showed that one can get a probabilistic distortion of $2^{O\left(\sqrt{\log n \log\log n}\right)}$. Yair Bartal

[②] Indeed, if $f_{i,j}(x) < \mathsf{d}_i(x, V_{i,j})$ and $f_{i,j}(y) < \mathsf{d}_i(x, V_{i,j})$ then $f_{i,j}(x) = 2r_i$ and $f_{i,j}(y) = 2r_i$, which implies the above inequality. If $f_{i,j}(x) = \mathsf{d}_i(x, V_{i,j})$ and $f_{i,j}(y) = \mathsf{d}_i(x, V_{i,j})$ then the inequality trivially holds. The other option is handled in a similar fashion.

realized that by allowing trees with additional vertices, one can get a considerably better result. In particular, he showed [Bar96] that probabilistic embedding into trees can be done with polylogarithmic average distortion. He later improved the distortion to $O(\log n \log\log n)$ in [Bar98]. Improving this result was an open question, culminating in the work of Fakcharoenphol *et al.* [FRT03] which achieve the optimal $O(\log n)$ distortion.

Interestingly, if one does not care about the optimal distortion, one can get similar result (for embedding into probabilistic trees), by first embedding the metric into Euclidean space, then reduce the dimension by the Johnson-Lindenstrauss lemma, and finally, construct an HST by constructing a quadtree over the points. The "trick" is to randomly translate the quadtree. It is easy to verify that this yields $O(\log^4 n)$ distortion. See the survey by Indyk [Ind01] for more details. This random shifting of quadtrees is a powerful technique that was used in getting several result, and it is a crucial ingredient in Arora [Aro98] approximation algorithm for Euclidean TSP.

Our proof of Lemma 23.3.1 (which is originally from [FRT03]) is taken from [KLMN05]. The proof of Theorem 23.5.3 is by Gupta [Gup00].

A good exposition of metric spaces is available in Matoušek [Mat02].

23.7. Exercises

Exercise 23.7.1 (Clustering for HST.). Let $(\mathcal{X}, \mathsf{d})$ be a HST defined over n points, and let $k > 0$ be an integer. Provide an algorithm that computes the optimal k-median clustering of $\mathcal{X}$ in $O(k^2 n)$ time.

[Transform the HST into a tree where every node has only two children. Next, run a dynamic programming algorithm on this tree.]

Exercise 23.7.2 (Partition induced metric.).

(a) Give a counter example to the following claim: Let $(\mathcal{X}, \mathsf{d})$ be a metric space, and let P be a partition of $\mathcal{X}$. Then, the pair (P, d') is a metric, where $\mathsf{d}'(C, C') = \mathsf{d}(C, C') = \min_{x \in C, y \in C'} \mathsf{d}(x, y)$ and $C, C' \in P$.

(b) Let $(\mathcal{X}, \mathsf{d})$ be a n-point metric space, and consider the set $U = \left\{ i \;\middle|\; 2^i \le \mathsf{d}(x, y) \le 2^{i+1}, \text{ for } x, y \in \mathcal{X} \right\}$. Prove that $|U| = O(n)$. Namely, there are only n different resolutions that "matter" for a finite metric space.

Exercise 23.7.3 (Computing the diameter via embeddings.).

(a) (h:1) Let ℓ be a line in the plane, and consider the embedding $f : \mathbb{R}^2 \to \ell$, which is the projection of the plane into ℓ. Prove that f is 1-Lipschitz, but it is not K-bi-Lipschitz for any constant K.

(b) (h:3) Prove that one can find a family of projections $\mathcal{F}$ of size $O(1/\sqrt{\varepsilon})$, such that for any two points $x, y \in \mathbb{R}^2$, for one of the projections $f \in \mathcal{F}$ we have $\mathsf{d}(f(x), f(y)) \ge (1 - \varepsilon)\mathsf{d}(x, y)$.

(c) (h:1) Given a set P of n in the plane, given a $O(n/\sqrt{\varepsilon})$ time algorithm that outputs two points $x, y \in P$, such that $\mathsf{d}(x, y) \ge (1 - \varepsilon)\mathrm{diam}(P)$, where $\mathrm{diam}(P) = \max_{z,w \in P} \mathsf{d}(z, w)$ is the diameter of P.

(d) (h:2) Given P, show how to extract, in $O(n)$ time, a set $Q \subseteq P$ of size $O(\varepsilon^{-2})$, such that $\mathrm{diam}(Q) \ge (1 - \varepsilon/2)\mathrm{diam}(P)$. (Hint: Construct a grid of appropriate resolution.)

In particular, give an $(1 - \varepsilon)$-approximation algorithm to the diameter of P that works in $O(n + \varepsilon^{-2.5})$ time. (There are slightly faster approximation algorithms known for approximating the diameter.)

Acknowledgments

The presentation in this write-up follows closely the insightful suggestions of Manor Mendel.

Bibliography

[AKPW95] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k-server problem. *SIAM J. Comput.*, 24(1):78–100, February 1995.

[Aro98] S. Arora. Polynomial time approximation schemes for Euclidean TSP and other geometric problems. *J. Assoc. Comput. Mach.*, 45(5):753–782, Sept. 1998.

[Bar96] Y. Bartal. Probabilistic approximations of metric space and its algorithmic application. In *Proc. 37th Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), pages 183–193, October 1996.

[Bar98] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *Proc. 30th Annu. ACM Sympos. Theory Comput.* (STOC), pages 161–168, 1998.

[CKR01] G. Calinescu, H. Karloff, and Y. Rabani. Approximation algorithms for the 0-extension problem. In *Proc. 12th ACM-SIAM Sympos. Discrete Algs.* (SODA), pages 8–16, 2001.

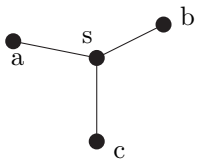
[FRT03] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proc. 35th Annu. ACM Sympos. Theory Comput.* (STOC), pages 448–455, 2003.

[Gup00] A. Gupta. *Embeddings of Finite Metrics*. PhD thesis, University of California, Berkeley, 2000.

[Ind01] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), pages 10–31, 2001. Tutorial.

[KLMN05] R. Krauthgamer, J. R. Lee, M. Mendel, and A. Naor. Measured descent: A new embedding method for finite metric spaces. *Geom. funct. anal. (GAFA)*, 15(4):839–858, 2005.

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Grad. Text in Math.* Springer, 2002.



Chapter 24

Approximate Max Cut

By Sariel Har-Peled, December 30, 2015[①]

24.1. Problem Statement

Given an undirected graph $G = (V, E)$ and nonnegative weights w_{ij} on the edge $ij \in E$, the *maximum cut problem* (MAX CUT) is that of finding the set of vertices S that maximizes the weight of the edges in the cut $(S, \overline{S})$; that is, the weight of the edges with one endpoint in S and the other in $\overline{S}$. For simplicity, we usually set $w_{ij} = O$ for $ij \notin E$ and denote the weight of a cut $(S, \overline{S})$ by $w(S, \overline{S}) = \sum_{i \in S, j \in \overline{j}} w_{ij}$.

This problem is NP-Complete, and hard to approximate within a certain constant.

Given a graph with vertex set $V = 1, \ldots, n$ and nonnegative weights W_{ij}, the weight of the maximum cut $w(S, \overline{S})$ N given by the following integer quadratic program:

$$\begin{aligned} \text{(Q)} \quad \text{Maximize} \quad & \frac{1}{2} \sum_{i<j}, w_{ij}(1 - y_i y_j) \\ \text{subject to:} \quad & y_i \in \{-1, 1\} \qquad \forall i \in V. \end{aligned}$$

Indeed, set $S = \left\{ i \;\middle|\; y_i = 1 \right\}$. Clearly, $w(S, \overline{S}) = \frac{1}{2} \sum_{i<j}, w_{ij}(1 - y_i y_j)$.

Solving quadratic integer programming is of course NP-Hard. Thus, we we will relax it, by thinking about the the numbers y_i as unit vectors in higher dimensional space. If so, the multiplication of the two vectors, is now replaced by dot product. We have:

$$\begin{aligned} (P) \quad \text{Maximize} \quad & \tfrac{1}{2} \textstyle\sum_{i<j} w_{ij}(1 - \left\langle v_i, v_j \right\rangle) \\ \text{subject to:} \quad & v_i \in \mathbb{S}^{(n)} \qquad \forall i \in V, \end{aligned}$$

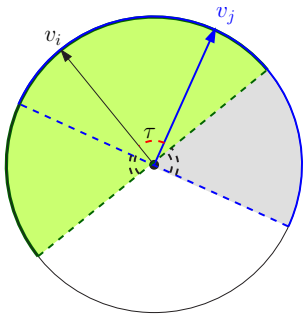
where $\mathbb{S}^{(n)}$ is the n dimensional unit sphere in $\mathbb{R}^{n+1}$. This is an instance of semi-definite programming, which is a special case of convex programming, which can be solved in polynomial time (solved here means approximated within arbitrary constant in polynomial time). Observe that (P) is a relaxation of (Q), and as such the optimal solution of (P) has value larger than the optimal value of (Q).

The intuition is that vectors that correspond to vertices that should be on one side of the cut, and vertices on the other sides, would have vectors which are faraway from each other in (P). Thus, we compute the optimal solution for (P), and we uniformly generate a random vector $\mathbf{r}$ on the unit sphere $\mathbb{S}^{(n)}$. This induces a hyperplane h which passes through the origin and is orthogonal to $\mathbf{r}$. We next assign all the vectors that are on one side of h to S, and the rest to $\overline{S}$.

24.1.1. Analysis

The intuition of the above rounding procedure, is that with good probability, vectors that have big angle between them would be separated by this cut.

Lemma 24.1.1. *We have* $\mathbf{Pr}\Big[\operatorname{sign}(\langle v_i, \mathbf{r}\rangle) \neq \operatorname{sign}\Big(\Big\langle v_j, \mathbf{r}\Big\rangle\Big)\Big] = \frac{1}{\pi}\arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big)$.



Proof: Let us think about the vectors v_i, v_j and $\mathbf{r}$ as being in the plane. To see why this is a reasonable assumption, consider the plane g spanned by v_i and v_j, and observe that for the random events we consider, only the direction of $\mathbf{r}$ matter, which can be decided by projecting $\mathbf{r}$ on g, and normalizing it to have length 1. Now, the sphere is symmetric, and as such, sampling $\mathbf{r}$ randomly from $\mathbb{S}^{(n)}$, projecting it down to g, and then normalizing it, is equivalent to just choosing uniformly a vector from the unit circle.

Now, $\operatorname{sign}(\langle v_i, \mathbf{r}\rangle) \neq \operatorname{sign}\Big(\Big\langle v_j, \mathbf{r}\Big\rangle\Big)$ happens only if $\mathbf{r}$ falls in the double wedge formed by the lines perpendicular to v_i and v_j. The angle of this double wedge is exactly the angle between v_i and v_j. Now, since v_i and v_j are unit vectors, we have $\Big\langle v_i, v_j\Big\rangle = \cos(\tau)$, where $\tau = \angle v_i v_j$. Thus, $\mathbf{Pr}\Big[\operatorname{sign}(\langle v_i, \mathbf{r}\rangle) \neq \operatorname{sign}\Big(\Big\langle v_j, \mathbf{r}\Big\rangle\Big)\Big] = 2\tau/(2\pi) = \frac{1}{\pi}\cdot\arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big)$, as claimed. ■

Theorem 24.1.2. *Let W be the random variable which is the weight of the cut generated by the algorithm. We have*

$$\mathbf{E}[W] = \frac{1}{\pi}\sum_{i<j} w_{ij}\arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big).$$

Proof: Let X_{ij} be an indicator variable which is 1 if ij is in the cut. We have $\mathbf{E}\Big[X_{ij}\Big] = \mathbf{Pr}\Big[\operatorname{sign}(\langle v_i, \mathbf{r}\rangle) \neq \operatorname{sign}\Big(\Big\langle v_j, \mathbf{r}\Big\rangle\Big)\Big] = \frac{1}{\pi}\arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big)$, by Lemma 24.1.1.

Clearly, $W = \sum_{i<j} w_{ij}X_{ij}$, and by linearity of expectation, we have

$$\mathbf{E}[W] = \sum_{i<j} w_{ij}\,\mathbf{E}\Big[X_{ij}\Big] = \sum_{i<j}\arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big).$$ ■

Lemma 24.1.3. *For $-1 \le y \le 1$, we have* $\dfrac{\arccos(y)}{\pi} \ge \alpha\cdot\dfrac{1}{2}(1-y)$, *where* $\alpha = \min_{0\le\psi\le\pi}\dfrac{2}{\pi}\dfrac{\psi}{1-\cos(\psi)}$.

Proof: Set $y = \cos(\psi)$. The inequality now becomes $\frac{\psi}{\pi} \ge \alpha\frac{1}{2}(1-\cos\psi)$. Reorganizing, the inequality becomes $\frac{2}{\pi}\frac{\psi}{1-\cos\psi} \ge \alpha$, which trivially holds by the definition of α. ■

Lemma 24.1.4. $\alpha > 0.87856$.

Proof: Using simple calculus, one can see that α achieves its value for $\psi = 2.331122...$, the nonzero root of $cos\psi + \psi \sin\psi = 1$. ■

Theorem 24.1.5. *The above algorithm computes in expectation a cut of size* $\alpha\text{Opt} \geq 0.87856\text{Opt}$*, where* Opt *is the weight of the maximal cut.*

Proof: Consider the optimal solution to (P), and lets its value be $\gamma \geq \text{Opt}$. We have

$$\mathbf{E}[W] = \frac{1}{\pi}\sum_{i<j} w_{ij} \arccos\Big(\Big\langle v_i, v_j\Big\rangle\Big) \geq \sum_{i<j} w_{ij}\alpha\frac{1}{2}\Big(1 - \Big\langle v_i, v_j\Big\rangle\Big) = \alpha\gamma \geq \alpha\text{Opt},$$

by Lemma 24.1.3. ■

24.2. Semi-definite programming

Let us define a variable $x_{ij} = \Big\langle v_i, v_j\Big\rangle$, and consider the n by n matrix M formed by those variables, where $x_{ii} = 1$ for $i = 1, \ldots, n$. Let V be the matrix having $v_1, \ldots, v_n$ as its columns. Clearly, $M = V^T V$. In particular, this implies that for any non-zero vector $v \in \mathbb{R}^n$, we have $v^T M v = v^T A^T A v = (Av)^T (Av) \geq 0$. A matrix that has this property, is called *semidefinite*. The interesting thing is that any semi-definite matrix P can be represented as a product of a matrix with its transpose; namely, $P = B^T B$. It is easy to observe that if this semi-definite matrix has a diagonal one, then B has rows which are unit vectors. Thus, if we solve (P) and get back a semi-definite matrix, then we can recover the vectors realizing the solution, and use them for the rounding.

In particular, (P) can now be restated as

$$\begin{array}{rll} (SD) \quad \text{Maximize} & & \frac{1}{2}\sum_{i<j} w_{ij}(1 - x_{ij}) \\ & x_{ii} = 1 & \text{for } i = 1, \ldots, n \\ & \text{subject to:} \quad \Big(x_{ij}\Big)_{i=1,\ldots,n, j=1,\ldots,n} & \text{is semi-definite.} \end{array}$$

We are trying to find the optimal value of a linear function over a set which is the intersection of linear constraints and the set of semi-definite matrices.

Lemma 24.2.1. *Let* $\mathcal{U}$ *be the set of* $n \times n$ *semidefinite matrices. The set* $\mathcal{U}$ *is convex.*

Proof: Consider $A, B \in \mathcal{U}$, and observe that for any $t \in [0, 1]$, and vector $v \in \mathbb{R}^n$, we have: $v^T(tA + (1-t)B)v = tv^T Av + (1-t)v^T Bv \geq 0 + 0 \geq 0$, since A and B are semidefinite. ■

Positive semidefinite matrices corresponds to ellipsoids. Indeed, consider the set $x^T Ax = 1$: the set of vectors that solve this equation is an ellipsoid. Also, the eigenvalues of a positive semidefinite matrix are all non-negative real numbers. Thus, given a matrix, we can in polynomial time decide if it is positive semidefinite or not.

Thus, we are trying to optimize a linear function over a convex domain. There is by now machinery to approximately solve those problems to within any additive error in polynomial time. This is done by using interior point method, or the ellipsoid method. See [BV04, GLS93] for more details.

24.3. Bibliographical Notes

The approximation algorithm presented is from the work of Goemans and Williamson [GW95]. Håstad [Hås01] showed that MAX CUT can not be approximated within a factor of $16/17 \approx 0.941176$. Recently, Khot et al. [KKMO04] showed a hardness result that matches the constant of Goemans and Williamson (i.e., one can not approximate it better than ϕ, unless **P** = **NP**). However, this relies on two conjectures, the first one is the "Unique Games Conjecture", and the other one is "Majority is Stablest". The "Majority is Stablest" conjecture was recently proved by Mossel *et al.* [MOO05]. However, it is not clear if the "Unique Games Conjecture" is true, see the discussion in [KKMO04].

The work of Goemans and Williamson was very influential and spurred wide research on using SDP for approximation algorithms. For an extension of the MAX CUT problem where negative weights are allowed and relevant references, see the work by Alon and Naor [AN04].

Bibliography

[AN04] N. Alon and A. Naor. Approximating the cut-norm via grothendieck's inequality. In *Proc. 36th Annu. ACM Sympos. Theory Comput.* (STOC), pages 72–80, 2004.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge, 2004.

[GLS93] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin Heidelberg, 2nd edition, 1993.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, November 1995.

[Hås01] J. Håstad. Some optimal inapproximability results. *J. Assoc. Comput. Mach.*, 48(4):798–859, 2001.

[KKMO04] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for max cut and other 2-variable csps. In *Proc. 45th Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), pages 146–154, 2004. To appear in SICOMP.

[MOO05] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences invariance and optimality. In *Proc. 46th Annu. IEEE Sympos. Found. Comput. Sci.* (FOCS), pages 21–30, 2005.

Chapter 25

Entropy, Randomness, and Information

By Sariel Har-Peled, December 30, 2015[①]

> "If only once - only once - no matter where, no matter before what audience - I could better the record of the great Rastelli and juggle with thirteen balls, instead of my usual twelve, I would feel that I had truly accomplished something for my country. But I am not getting any younger, and although I am still at the peak of my powers there are moments - why deny it? - when I begin to doubt - and there is a time limit on all of us."
>
> – –Romain Gary, The talent scout..

25.1. Entropy

Definition 25.1.1. The *entropy* in bits of a discrete random variable X is given by

$$\mathbb{H}(X) = -\sum_x \mathbf{Pr}[X = x] \lg \mathbf{Pr}[X = x] .$$

Equivalently, $\mathbb{H}(X) = \mathbf{E}\left[\lg \frac{1}{\mathbf{Pr}[X]}\right]$.

The ***binary entropy*** function $\mathbb{H}(p)$ for a random binary variable that is 1 with probability p, is $\mathbb{H}(p) = -p \lg p - (1-p)\lg(1-p)$. We define $\mathbb{H}(0) = \mathbb{H}(1) = 0$.

The function $\mathbb{H}(p)$ is a concave symmetric around $1/2$ on the interval $[0, 1]$ and achieves its maximum at $1/2$. For a concrete example, consider $\mathbb{H}(3/4) \approx 0.8113$ and $\mathbb{H}(7/8) \approx 0.5436$. Namely, a coin that has $3/4$ probably to be heads have higher amount of "randomness" in it than a coin that has probability $7/8$ for heads.

We have that

$$\mathbb{H}(p) = \frac{1}{\ln 2}(-p \ln p - (1-p)\ln(1-p))$$

$$\text{and} \quad \mathbb{H}'(p) = \frac{1}{\ln 2}\left(-\ln p - \frac{p}{p} - (-1)\ln(1-p) - \frac{1-p}{1-p}(-1)\right) = \lg \frac{1-p}{p}.$$

Deploying our amazing ability to compute derivative of simple functions once more, we get that

$$\mathbb{H}''(p) = \frac{1}{\ln 2}\frac{p}{1-p}\left(\frac{p(-1) - (1-p)}{p^2}\right) = -\frac{1}{p(1-p)\ln 2}.$$

Since $\ln 2 \approx 0.693$, we have that $\mathbb{H}''(p) \leq 0$, for all $p \in (0,1)$, and the $\mathbb{H}(\cdot)$ is concave in this range. Also, $\mathbb{H}'(1/2) = 0$, which implies that $\mathbb{H}(1/2) = 1$ is a maximum of the binary entropy. Namely, a balanced coin has the largest amount of randomness in it.

Example 25.1.2. A random variable X that has probability $1/n$ to be i, for $i = 1, \ldots, n$, has entropy $\mathbb{H}(X) = -\sum_{i=1}^{n} \frac{1}{n} \lg \frac{1}{n} = \lg n$.

Note, that the entropy is oblivious to the exact values that the random variable can have, and it is sensitive only to the probability distribution. Thus, a random variables that accepts $-1, +1$ with equal probability has the same entropy (i.e., 1) as a fair coin.

Lemma 25.1.3. *Let X and Y be two independent random variables, and let Z be the random variable (X, T). Then $\mathbb{H}(Z) = \mathbb{H}(X) + \mathbb{H}(Y)$.*

Proof: In the following, summation are over all possible values that the variables can have. By the independence of X and Y we have

$$\begin{aligned}
\mathbb{H}(Z) &= \sum_{x,y} \mathbf{Pr}[(X,Y) = (x,y)] \lg \frac{1}{\mathbf{Pr}[(X,Y) = (x,y)]} \\
&= \sum_{x,y} \mathbf{Pr}[X = x] \mathbf{Pr}[Y = y] \lg \frac{1}{\mathbf{Pr}[X = x] \mathbf{Pr}[Y = y]} \\
&= \sum_{x} \sum_{y} \mathbf{Pr}[X = x] \mathbf{Pr}[Y = y] \lg \frac{1}{\mathbf{Pr}[X = x]} \\
&\quad + \sum_{y} \sum_{x} \mathbf{Pr}[X = x] \mathbf{Pr}[Y = y] \lg \frac{1}{\mathbf{Pr}[Y = y]} \\
&= \sum_{x} \mathbf{Pr}[X = x] \lg \frac{1}{\mathbf{Pr}[X = x]} + \sum_{y} \mathbf{Pr}[Y = y] \lg \frac{1}{\mathbf{Pr}[Y = y]} = \mathbb{H}(X) + \mathbb{H}(Y).
\end{aligned}$$

■

Lemma 25.1.4. *Suppose that nq is integer in the range $[0, n]$. Then $\frac{2^{n\mathbb{H}(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{n\mathbb{H}(q)}$.*

Proof: This trivially holds if $q = 0$ or $q = 1$, so assume $0 < q < 1$. We know that $\binom{n}{nq} q^{nq}(1-q)^{n-nq} \leq (q + (1-q))^n = 1$. As such, since $q^{-nq}(1-q)^{-(1-q)n} = 2^{n(-q \lg q - (1-q)\lg(1-q))} = 2^{n\mathbb{H}(q)}$, we have

$$\binom{n}{nq} \leq q^{-nq}(1-q)^{-(1-q)n} = 2^{n\mathbb{H}(q)}.$$

As for the other direction, we claim that $\mu(nq) = \binom{n}{nq} q^{nq}(1-q)^{n-nq}$ is the largest term in $\sum_{k=0}^{n} \mu(k) = 1$, where $\mu(k) = \binom{n}{k} q^k (1-q)^{n-k}$. Indeed,

$$\Delta_k = \mu(k) - \mu(k+1) = \binom{n}{k} q^k (1-q)^{n-k} \left(1 - \frac{n-k}{k+1} \frac{q}{1-q}\right),$$

and the sign of this quantity is the sign of $(k+1)(1-q) - (n-k)q = k + 1 - kq - q - nq + kq = 1 + k - q - nq$. Namely, $\Delta_k \geq 0$ when $k \geq nq + q - 1$, and $\Delta_k < 0$ otherwise. Namely, $\mu(k) < \mu(k+1)$, for $k < nq$, and $\mu(k) \geq \mu(k+1)$ for $k \geq nq$. Namely, $\mu(nq)$ is the largest term in $\sum_{k=0}^{n} \mu(k) = 1$, and as such it is larger than the average. We have $\mu(nq) = \binom{n}{nq} q^{nq}(1-q)^{n-nq} \geq \frac{1}{n+1}$, which implies

$$\binom{n}{nq} \geq \frac{1}{n+1} q^{-nq}(1-q)^{-(n-nq)} = \frac{1}{n+1} 2^{n\mathbb{H}(q)}.$$

■

Lemma 25.1.4 can be extended to handle non-integer values of q. This is straightforward, and we omit the easy details.

Corollary 25.1.5. *We have:* *(i)* $q \in [0, 1/2] \Rightarrow \binom{n}{\lfloor nq \rfloor} \leq 2^{n\mathbb{H}(q)}$. *(ii)* $q \in [1/2, 1]$ $\binom{n}{\lceil nq \rceil} \leq 2^{n\mathbb{H}(q)}$. *(iii)* $q \in [1/2, 1] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \leq \binom{n}{\lfloor nq \rfloor}$. *(iv)* $q \in [0, 1/2] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \leq \binom{n}{\lceil nq \rceil}$.

The bounds of Lemma 25.1.4 and Corollary 25.1.5 are loose but sufficient for our purposes. As a sanity check, consider the case when we generate a sequence of n bits using a coin with probability q for head, then by the Chernoff inequality, we will get roughly nq heads in this sequence. As such, the generated sequence Y belongs to $\binom{n}{nq} \approx 2^{n\mathbb{H}(q)}$ possible sequences that have similar probability. As such, $\mathbb{H}(Y) \approx \lg \binom{n}{nq} = n\mathbb{H}(q)$, by Example 25.1.2, a fact that we already know from Lemma 25.1.3.

25.1.1. Extracting randomness

Entropy can be interpreted as the amount of unbiased random coin flips can be extracted from a random variable.

Definition 25.1.6. An extraction function $\texttt{Ext}$ takes as input the value of a random variable X and outputs a sequence of bits y, such that $\mathbf{Pr}\Big[\texttt{Ext}(X) = y \;\Big|\; |y| = k\Big] = \frac{1}{2^k}$, whenever $\mathbf{Pr}[|y| = k] \geq 0$, where $|y|$ denotes the length of y.

As a concrete (easy) example, consider X to be a uniform random integer variable out of $0, \ldots, 7$. All that $\texttt{Ext}(x)$ has to do in this case, is just to compute the binary representation of x. However, note that Definition 25.1.6 is somewhat more subtle, as it requires that all extracted sequence of the same length would have the same probability.

Thus, for X a uniform random integer variable in the range $0, \ldots, 11$, the function $\texttt{Ext}(x)$ can output the binary representation for x if $0 \leq x \leq 7$. However, what do we do if x is between 8 and 11? The idea is to output the binary representation of $x - 8$ as a two bit number. Clearly, Definition 25.1.6 holds for this extraction function, since $\mathbf{Pr}\Big[\texttt{Ext}(X) = 00 \;\Big|\; |\texttt{Ext}(X)| = 2\Big] = \frac{1}{4}$, as required. This scheme can be of course extracted for any range.

Theorem 25.1.7. *Suppose that the value of a random variable X is chosen uniformly at random from the integers $\{0, \ldots, m-1\}$. Then there is an extraction function for X that outputs on average (i.e., in expectation) at least $\lfloor \lg m \rfloor - 1 = \lfloor \mathbb{H}(X) \rfloor - 1$ independent and unbiased bits.*

Proof: We represent m as a sum of unique powers of 2, namely $m = \sum_i a_i 2^i$, where $a_i \in \{0, 1\}$. Thus, we decomposed $\{0, \ldots, m-1\}$ into a disjoint union of blocks that have sizes which are distinct powers of 2. If a number falls inside such a block, we output its relative location in the block, using binary representation of the appropriate length (i.e., k if the block is of size 2^k). The fact that this is an extraction function, fulfilling Definition 25.1.6, is obvious.

Now, observe that the claim holds trivially if m is a power of two. Thus, if m is not a power of 2, then in the decomposition if there is a block of size 2^k, and the X falls inside this block, then the entropy is k. Thus, for the inductive proof, assume that are looking at the largest block in the decomposition, that is $m < 2^{k+1}$, and let $u = \Big\lfloor \lg(m - 2^k) \Big\rfloor < k$. It is easy to verify that, for any integer $\alpha > 2^k$, we have $\frac{\alpha - 2^k}{\alpha} \leq \frac{\alpha + 1 - 2^k}{\alpha + 1}$. Furthermore, $m \leq 2^{u+1} + 2^k$. As such, $\frac{m - 2^k}{m} \leq \frac{2^{u+1}}{2^{u+1} + 2^k}$.

Let Y be the random variable which is the number of random bits extract. We have that

$$\begin{aligned}
\mathbf{E}[Y] &\geq \frac{2^k}{m} k + \frac{m - 2^k}{m}\Big(\Big\lfloor \lg(m - 2^k) \Big\rfloor - 1\Big) = k + \frac{m - 2^k}{m}(u - k - 1) \\
&\geq k + \frac{2^{u+1}}{2^{u+1} + 2^k}(u - k - 1) = k - \frac{2^{u+1}}{2^{u+1} + 2^k}(1 + k - u).
\end{aligned}$$

If $u = k-1$, then $\mathbb{H}(X) \geq k - \frac{1}{2}\cdot 2 = k-1$, as required. If $u = k-2$ then $\mathbb{H}(X) \geq k - \frac{1}{3}\cdot 3 = k-1$. Finally, if $u < k-2$ then

$$\mathbf{E}[Y] \geq k - \frac{2^{u+1}}{2^k}(1+k-u) \geq k - \frac{k-u+1}{2^{k-u-1}} \geq k-1,$$

since $\frac{2+i}{2^i} \leq 1$ for $i \geq 2$. ■

Theorem 25.1.8. *Consider a coin that comes up heads with probability $p > 1/2$. For any constant $\delta > 0$ and for n sufficiently large:*

1. *One can extract, from an input of a sequence of n flips, an output sequence of $(1-\delta)n\mathbb{H}(p)$ (unbiased) independent random bits.*
2. *One can not extract more than $n\mathbb{H}(p)$ bits from such a sequence.*

Proof: There are $\binom{n}{j}$ input sequences with exactly j heads, and each has probability $p^j(1-p)^{n-j}$. We map this sequence to the corresponding number in the set $\left\{0,\ldots,\binom{n}{j}-1\right\}$. Note, that this, conditional distribution on j, is uniform on this set, and we can apply the extraction algorithm of Theorem 25.1.7. Let Z be the random variables which is the number of heads in the input, and let B be the number of random bits extracted. We have

$$\mathbf{E}[B] = \sum_{k=0}^{n} \mathbf{Pr}[Z=k]\,\mathbf{E}\Big[B \;\Big|\; Z=k\Big],$$

and by Theorem 25.1.7, we have $\mathbf{E}\Big[B \;\Big|\; Z=k\Big] \geq \left\lfloor \lg\binom{n}{k}\right\rfloor - 1$. Let $\varepsilon < p - 1/2$ be a constant to be determined shortly. For $n(p-\varepsilon) \leq k \leq n(p+\varepsilon)$, we have

$$\binom{n}{k} \geq \binom{n}{\lfloor n(p+\varepsilon)\rfloor} \geq \frac{2^{n\mathbb{H}(p+\varepsilon)}}{n+1},$$

by Corollary 25.1.5 (iii). We have

$$\begin{aligned}
\mathbf{E}[B] &\geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbf{Pr}[Z=k]\,\mathbf{E}\Big[B \;\Big|\; Z=k\Big] \geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbf{Pr}[Z=k]\left(\left\lfloor \lg\binom{n}{k}\right\rfloor - 1\right)\\
&\geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbf{Pr}[Z=k]\left(\lg\frac{2^{n\mathbb{H}(p+\varepsilon)}}{n+1} - 2\right)\\
&= (n\mathbb{H}(p+\varepsilon) - \lg(n+1))\,\mathbf{Pr}[|Z-np| \leq \varepsilon n]\\
&\geq (n\mathbb{H}(p+\varepsilon) - \lg(n+1))\left(1 - 2\exp\left(-\frac{n\varepsilon^2}{4p}\right)\right),
\end{aligned}$$

since $\mu = \mathbf{E}[Z] = np$ and $\mathbf{Pr}\Big[|Z-np| \geq \frac{\varepsilon}{p}pn\Big] \leq 2\exp\left(-\frac{np}{4}\left(\frac{\varepsilon}{p}\right)^2\right) = 2\exp\left(-\frac{n\varepsilon^2}{4p}\right)$, by the Chernoff inequality. In particular, fix $\varepsilon > 0$, such that $\mathbb{H}(p+\varepsilon) > (1-\delta/4)\mathbb{H}(p)$, and since p is fixed $n\mathbb{H}(p) = \Omega(n)$, in particular, for n sufficiently large, we have $-\lg(n+1) \geq -\frac{\delta}{10}n\mathbb{H}(p)$. Also, for n sufficiently large, we have $2\exp\left(-\frac{n\varepsilon^2}{4p}\right) \leq \frac{\delta}{10}$. Putting it together, we have that for n large enough, we have

$$\mathbf{E}[B] \geq \left(1 - \frac{\delta}{4} - \frac{\delta}{10}\right)n\mathbb{H}(p)\left(1 - \frac{\delta}{10}\right) \geq (1-\delta)n\mathbb{H}(p),$$

as claimed.

As for the upper bound, observe that if an input sequence x has probability q, then the output sequence $y = \texttt{Ext}(x)$ has probability to be generated which is at least q. Now, all sequences of length $|y|$ have equal probability to be generated. Thus, we have the following (trivial) inequality $2^{|\texttt{Ext}(x)|}q \leq 2^{|\texttt{Ext}(x)|}\,\mathbf{Pr}[y = \texttt{Ext}(X)] \leq 1$, implying that $|\texttt{Ext}(x)| \leq \lg(1/q)$. Thus,

$$\mathbf{E}[B] = \sum_x \mathbf{Pr}[X = x]\,|\texttt{Ext}(x)| \leq \sum_x \mathbf{Pr}[X = x] \lg \frac{1}{\mathbf{Pr}[X = x]} = \mathbb{H}(X). \qquad \blacksquare$$

25.2. Bibliographical Notes

The presentation here follows [MU05, Sec. 9.1-Sec 9.3].

Bibliography

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.

Chapter 26

Entropy II

By Sariel Har-Peled, December 30, 2015[1]

> The memory of my father is wrapped up in white paper, like sandwiches taken for a day at work. Just as a magician takes towers and rabbits out of his hat, he drew love from his small body, and the rivers of his hands overflowed with good deeds.
>
> – – Yehuda Amichai, My Father..

26.1. Compression

In this section, we will consider the problem of how to compress a binary string. We will map each binary string, into a new string (which is hopefully shorter). In general, by using a simple counting argument, one can show that no such mapping can achieve real compression (when the inputs are adversarial). However, the hope is that there is an underling distribution on the inputs, such that some strings are considerably more common than others.

Definition 26.1.1. A compression function `Compress` takes as input a sequence of n coin flips, given as an element of $\{H,T\}^n$, and outputs a sequence of bits such that each input sequence of n flips yields a distinct output sequence.

The following is easy to verify.

Lemma 26.1.2. *If a sequence S_1 is more likely than S_2 then the compression function that minimizes the expected number of bits in the output assigns a bit sequence to S_2 which is at least as long as S_1.*

Note, that this is very weak. Usually, we would like the function to output a prefix code, like the Huffman code.

Theorem 26.1.3. *Consider a coin that comes up heads with probability $p > 1/2$. For any constant $\delta > 0$, when n is sufficiently large, the following holds.*

(i) *There exists a compression function* `Compress` *such that the expected number of bits output by* `Compress` *on an input sequence of n independent coin flips (each flip gets heads with probability p) is at most $(1+\delta)n\mathbb{H}(p)$; and*

(ii) *The expected number of bits output by any compression function on an input sequence of n independent coin flips is at least $(1-\delta)n\mathbb{H}(p)$.*

Proof: Let $\varepsilon > 0$ be a constant such that $p - \varepsilon > 1/2$. The first bit output by the compression procedure is '1' if the output string is just a copy of the input (using $n+1$ bits overall in the output), and '0' if it is compressed. We compress only if the number of ones in the input sequence, denoted by X is larger than $(p-\varepsilon)n$. By the Chernoff inequality, we know that $\mathbf{Pr}[X < (p-\varepsilon)n] \le \exp\left(-n\varepsilon^2/2p\right)$.

If there are more than $(p-\varepsilon)n$ ones in the input, and since $p - \varepsilon > 1/2$, we have that

$$\sum_{j=\lceil n(p-\varepsilon)\rceil}^{n} \binom{n}{j} \le \sum_{j=\lceil n(p-\varepsilon)\rceil}^{n} \binom{n}{\lceil n(p-\varepsilon)\rceil} \le \frac{n}{2} 2^{n\mathbb{H}(p-\varepsilon)},$$

by Corollary 26.2.1. As such, we can assign each such input sequence a number in the range $0 \ldots \frac{n}{2}2^{n\mathbb{H}(p-\varepsilon)}$, and this requires (with the flag bit) $1 + \lfloor \lg n + n\mathbb{H}(p-\varepsilon) \rfloor$ random bits.

Thus, the expected number of bits output is bounded by

$$(n+1)\exp\left(-n\varepsilon^2/2p\right) + (1 + \lfloor \lg n + n\mathbb{H}(p-\varepsilon)\rfloor) \le (1+\delta)n\mathbb{H}(p),$$

by carefully setting ε and n being sufficiently large. Establishing the upper bound.

As for the lower bound, observe that at least one of the sequences having exactly $\tau = \lfloor (p+\varepsilon)n \rfloor$ heads, must be compressed into a sequence having

$$\lg \binom{n}{\lfloor (p+\varepsilon)n \rfloor} - 1 \ge \lg \frac{2^{n\mathbb{H}(p+\varepsilon)}}{n+1} - 1 = n\mathbb{H}(p-\varepsilon) - \lg(n+1) - 1 = \mu,$$

by Corollary 26.2.1. Now, any input string with less than τ heads has lower probability to be generated. Indeed, for a specific strings with $\alpha < \tau$ ones the probability to generate them is $p^\alpha(1-p)^{n-\alpha}$ and $p^\tau(1-p)^{n-\tau}$, respectively. Now, observe that

$$p^\alpha(1-p)^{n-\alpha} = p^\tau(1-p)^{n-\tau} \cdot \frac{(1-p)^{\tau-\alpha}}{p^{\tau-\alpha}} = p^\tau(1-p)^{n-\tau}\left(\frac{1-p}{p}\right)^{\tau-\alpha} < p^\tau(1-p)^{n-\tau},$$

as $1 - p < 1/2 < p$ implies that $(1-p)/p < 1$.

As such, Lemma 26.1.2 implies that all the input strings with less than τ ones, must be compressed into strings of length at least μ, by an optimal compresser. Now, the Chenroff inequality implies that $\mathbf{Pr}[X \le \tau] \ge 1 - \exp\left(-n\varepsilon^2/12p\right)$. Implying that an optimal compresser outputs on average at least $\left(1 - \exp\left(-n\varepsilon^2/12p\right)\right)\mu$. Again, by carefully choosing ε and n sufficiently large, we have that the average output length of an optimal compressor is at least $(1-\delta)n\mathbb{H}(p)$. ∎

26.2. From previous lecture

Corollary 26.2.1. *We have:*

(i) $q \in [0, 1/2] \Rightarrow \binom{n}{\lfloor nq \rfloor} \le 2^{n\mathbb{H}(q)}$. *(ii)* $q \in [1/2, 1]$ $\binom{n}{\lceil nq \rceil} \le 2^{n\mathbb{H}(q)}$.

(iii) $q \in [1/2, 1] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \le \binom{n}{\lfloor nq \rfloor}$. *(iv)* $q \in [0, 1/2] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \le \binom{n}{\lceil nq \rceil}$.

26.3. Bibliographical Notes

The presentation here follows [MU05, Sec. 9.1-Sec 9.3].

Bibliography

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.

Chapter 27

Entropy III - Shannon's Theorem

By Sariel Har-Peled, December 30, 2015[1]

> The memory of my father is wrapped up in
> white paper, like sandwiches taken for a day at work.
>
> Just as a magician takes towers and rabbits
> out of his hat, he drew love from his small body,
>
> and the rivers of his hands
> overflowed with good deeds.
> – – Yehuda Amichai, My Father..

27.1. Coding: Shannon's Theorem

We are interested in the problem sending messages over a noisy channel. We will assume that the channel noise is "nicely" behaved.

Definition 27.1.1. The input to a *binary symmetric channel* with parameter p is a sequence of bits $x_1, x_2, \ldots$, and the output is a sequence of bits $y_1, y_2, \ldots$, such that $\mathbf{Pr}[x_i = y_i] = 1 - p$ independently for each i.

Translation: Every bit transmitted have the same probability to be flipped by the channel. The question is how much information can we send on the channel with this level of noise. Naturally, a channel would have some capacity constraints (say, at most 4,000 bits per second can be sent on the channel), and the question is how to send the largest amount of information, so that the receiver can recover the original information sent.

Now, its important to realize that noise handling is unavoidable in the real world. Furthermore, there are tradeoffs between channel capacity and noise levels (i.e., we might be able to send considerably more bits on the channel but the probability of flipping (i.e., p) might be much larger). In designing a communication protocol over this channel, we need to figure out where is the optimal choice as far as the amount of information sent.

Definition 27.1.2. A (k, n) *encoding function* $\mathtt{Enc} : \{0,1\}^k \to \{0,1\}^n$ takes as input a sequence of k bits and outputs a sequence of n bits. A (k, n) *decoding function* $\mathtt{Dec} : \{0,1\}^n \to \{0,1\}^k$ takes as input a sequence of n bits and outputs a sequence of k bits.

Thus, the sender would use the encoding function to send its message, and the decoder would use the received string (with the noise in it), to recover the sent message. Thus, the sender starts with a message with k bits, it blow it up to n bits, using the encoding function, to get some robustness to noise, it send it over the (noisy) channel to the receiver. The receiver, takes the given (noisy) message with n bits, and use the decoding function to recover the original k bits of the message.

Naturally, we would like k to be as large as possible (for a fixed n), so that we can send as much information as possible on the channel. Naturally, there might be some failure probability; that is, the receiver might be unable to recover the original string, or recover an incorrect string.

The following celebrated result of Shannon[2] in 1948 states exactly how much information can be sent on such a channel.

Theorem 27.1.3 (Shannon's theorem.). *For a binary symmetric channel with parameter $p < 1/2$ and for any constants $\delta, \gamma > 0$, where n is sufficiently large, the following holds:*

(i) For an $k \leq n(1 - \mathbb{H}(p) - \delta)$ there exists (k, n) encoding and decoding functions such that the probability the receiver fails to obtain the correct message is at most γ for every possible k-bit input messages.

(ii) There are no (k, n) encoding and decoding functions with $k \geq n(1 - \mathbb{H}(p) + \delta)$ such that the probability of decoding correctly is at least γ for a k-bit input message chosen uniformly at random.

27.2. Proof of Shannon's theorem

The proof is not hard, but requires some care, and we will break it into parts.

27.2.1. How to encode and decode efficiently

27.2.1.1. The scheme

Our scheme would be simple. Pick $k \leq n(1 - \mathbb{H}(p) - \delta)$. For any number $i = 0, \ldots, \widehat{K} = 2^{k+1} - 1$, randomly generate a binary string Y_i made out of n bits, each one chosen independently and uniformly. Let $Y_0, \ldots, Y_{\widehat{K}}$ denote these codewords.

For each of these codewords we will compute the probability that if we send this codeword, the receiver would fail. Let $X_0, \ldots, X_K$, where $K = 2^k - 1$, be the K codewords with the lowest probability of failure. We assign these words to the 2^k messages we need to encode in an arbitrary fashion. Specifically, for $i = 0, \ldots, 2^k - 1$, we encode i as the string X_i.

The decoding of a message w is done by going over all the codewords, and finding all the codewords that are in (Hamming) distance in the range $[p(1 - \varepsilon)n, p(1 + \varepsilon)n]$ from w. If there is only a single word X_i with this property, we return i as the decoded word. Otherwise, if there are no such word or there is more than one word then the decoder stops and report an error.

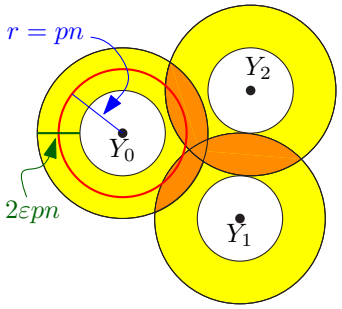
27.2.1.2. The proof

[2] Claude Elwood Shannon (April 30, 1916 - February 24, 2001), an American electrical engineer and mathematician, has been called "the father of information theory".

Intuition. Each code Y_i corresponds to a region that looks like a ring. The "ring" for Y_i is all the strings in Hamming distance between $(1-\varepsilon)r$ and $(1+\varepsilon)r$ from Y_i, where $r = pn$. Clearly, if we transmit a string Y_i, and the receiver gets a string inside the ring of Y_i, it is natural to try to recover the received string to the original code corresponding to Y_i. Naturally, there are two possible bad events here:

(A) The received string is outside the ring of Y_i, and

(B) The received string is contained in several rings of different Ys, and it is not clear which one should the receiver decode the string to. These bad regions are depicted as the darker regions in the figure on the right.



Let $S_i = \mathcal{S}(Y_i)$ be all the binary strings (of length n) such that if the receiver gets this word, it would decipher it to be the original string assigned to Y_i (here are still using the extended set of codewords $Y_0, \ldots, Y_{\widehat{K}}$). Note, that if we remove some codewords from consideration, the set $\mathcal{S}(Y_i)$ just increases in size (i.e., the bad region in the ring of Y_i that is covered multiple times shrinks). Let W_i be the probability that Y_i was sent, but it was not deciphered correctly. Formally, let r denote the received word. We have that

$$W_i = \sum_{r \notin S_i} \mathbf{Pr}[r \text{ was received when } Y_i \text{ was sent}]. \tag{27.1}$$

To bound this quantity, let $\Delta(x, y)$ denote the Hamming distance between the binary strings x and y. Clearly, if x was sent the probability that y was received is

$$w(x, y) = p^{\Delta(x,y)}(1-p)^{n-\Delta(x,y)}.$$

As such, we have

$$\mathbf{Pr}[r \text{ received when } Y_i \text{ was sent}] = w(Y_i, r).$$

Let $\overline{S_{i,r}}$ be an indicator variable which is 1 if $r \notin S_i$. We have that

$$W_i = \sum_{r \notin S_i} \mathbf{Pr}[r \text{ received when } Y_i \text{ was sent}] = \sum_{r \notin S_i} w(Y_i, r) = \sum_r \overline{S_{i,r}} w(Y_i, r). \tag{27.2}$$

The value of W_i is a random variable over the choice of $Y_0, \ldots, Y_{\widehat{K}}$. As such, its natural to ask what is the expected value of W_i.

Consider the ring

$$\text{ring}(r) = \left\{ x \in \{0,1\}^n \;\middle|\; (1-\varepsilon)np \le \Delta(x, r) \le (1+\varepsilon)np \right\},$$

where $\varepsilon > 0$ is a small enough constant. Observe that $x \in \text{ring}(y)$ if and only if $y \in \text{ring}(x)$. Suppose, that the code word Y_i was sent, and r was received. The decoder returns the original code associated with Y_i, if Y_i is the only codeword that falls inside $\text{ring}(r)$.

Lemma 27.2.1. *Given that Y_i was sent, and r was received and furthermore $r \in \text{ring}(Y_i)$, then the probability of the decoder failing, is*

$$\tau = \mathbf{Pr}\Big[r \notin S_i \;\Big|\; r \in \text{ring}(Y_i) \Big] \le \frac{\gamma}{8},$$

where γ is the parameter of Theorem 27.1.3.

Proof: The decoder fails here, only if $\mathrm{ring}(r)$ contains some other codeword Y_j ($j \neq i$) in it. As such,

$$\tau = \mathbf{Pr}\Big[r \notin S_i \;\Big|\; r \in \mathrm{ring}(Y_i)\Big] \leq \mathbf{Pr}\Big[Y_j \in \mathrm{ring}(r), \text{ for any } j \neq i\Big] \leq \sum_{j \neq i} \mathbf{Pr}\Big[Y_j \in \mathrm{ring}(r)\Big].$$

Now, we remind the reader that the Y_js are generated by picking each bit randomly and independently, with probability $1/2$. As such, we have

$$\mathbf{Pr}\Big[Y_j \in \mathrm{ring}(r)\Big] = \frac{\big|\mathrm{ring}(r)\big|}{|\{0,1\}^n|} = \sum_{m=(1-\varepsilon)np}^{(1+\varepsilon)np} \frac{\binom{n}{m}}{2^n} \leq \frac{n}{2^n}\binom{n}{\lfloor (1+\varepsilon)np \rfloor},$$

since $(1+\varepsilon)p < 1/2$ (for ε sufficiently small), and as such the last binomial coefficient in this summation is the largest. By Corollary 27.3.2 (i), we have

$$\mathbf{Pr}\Big[Y_j \in \mathrm{ring}(r)\Big] \leq \frac{n}{2^n}\binom{n}{\lfloor (1+\varepsilon)np \rfloor} \leq \frac{n}{2^n} 2^{n\mathbb{H}\left((1+\varepsilon)p\right)} = n2^{n\left(\mathbb{H}\left((1+\varepsilon)p\right)-1\right)}.$$

As such, we have

$$\begin{aligned}\tau = \mathbf{Pr}\Big[r \notin S_i \;\Big|\; r \in \mathrm{ring}(Y_i)\Big] &\leq \sum_{j \neq i} \mathbf{Pr}\Big[Y_j \in \mathrm{ring}(r)\Big] \leq \widehat{K}\,\mathbf{Pr}[Y_1 \in \mathrm{ring}(r)] \leq 2^{k+1} n 2^{n\left(\mathbb{H}\left((1+\varepsilon)p\right)-1\right)} \\ &\leq n2^{n\left(1-\mathbb{H}(p)-\delta\right)+1+n\left(\mathbb{H}\left((1+\varepsilon)p\right)-1\right)} \leq n2^{n\left(\mathbb{H}\left((1+\varepsilon)p\right)-\mathbb{H}(p)-\delta\right)+1}\end{aligned}$$

since $k \leq n(1-\mathbb{H}(p)-\delta)$. Now, we choose ε to be a small enough constant, so that the quantity $\mathbb{H}((1+\varepsilon)p) - \mathbb{H}(p) - \delta$ is equal to some (absolute) negative (constant), say $-\beta$, where $\beta > 0$. Then, $\tau \leq n2^{-\beta n+1}$, and choosing n large enough, we can make τ smaller than $\gamma/8$, as desired. As such, we just proved that

$$\tau = \mathbf{Pr}\Big[r \notin S_i \;\Big|\; r \in \mathrm{ring}(Y_i)\Big] \leq \frac{\gamma}{8}.$$ ■

Lemma 27.2.2. *Consider the situation where Y_i is sent, and the received string is r. We have that*

$$\mathbf{Pr}[r \notin \mathrm{ring}(Y_i)] = \sum_{r \notin \mathrm{ring}(Y_i)} w(Y_i, r) \leq \frac{\gamma}{8},$$

where γ is the parameter of Theorem 27.1.3.

Proof: This quantity, is the probability of sending Y_i when every bit is flipped with probability p, and receiving a string r such that more than $pn + \varepsilon pn$ bits where flipped (or less than $pn - \varepsilon pn$). But this quantity can be bounded using the Chernoff inequality. Indeed, let $Z = \Delta(Y_i, r)$, and observe that $\mathbf{E}[Z] = pn$, and it is the sum of n independent indicator variables. As such

$$\sum_{r \notin \mathrm{ring}(Y_i)} w(Y_i, r) = \mathbf{Pr}[|Z - \mathbf{E}[Z]| > \varepsilon pn] \leq 2\exp\left(-\frac{\varepsilon^2}{4} pn\right) < \frac{\gamma}{4},$$

since ε is a constant, and for n sufficiently large. ■

Lemma 27.2.3. *We have that $f(Y_i) = \sum_{r \notin \mathrm{ring}(Y_i)} \mathbf{E}\Big[\overline{S_{i,r}} w(Y_i, r)\Big] \leq \gamma/8$ (the expectation is over all the choices of the Ys excluding Y_i).*

Proof: Observe that $\overline{S_{i,r}} w(Y_i, r) \leq w(Y_i, r)$ and for fixed Y_i and r we have that $\mathbf{E}[w(Y_i, r)] = w(Y_i, r)$. As such, we have that

$$f(Y_i) = \sum_{r \notin \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r)\right] \leq \sum_{r \notin \text{ring}(Y_i)} \mathbf{E}[w(Y_i, r)] = \sum_{r \notin \text{ring}(Y_i)} w(Y_i, r) \leq \frac{\gamma}{8},$$

by Lemma 27.2.2. ■

Lemma 27.2.4. *We have that* $g(Y_i) = \sum_{r \in \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r)\right] \leq \gamma/8$ *(the expectation is over all the choices of the Ys excluding* Y_i*).*

Proof: We have that $\overline{S_{i,r}} w(Y_i, r) \leq \overline{S_{i,r}}$, as $0 \leq w(Y_i, r) \leq 1$. As such, we have that

$$\begin{aligned}
g(Y_i) = \sum_{r \in \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r)\right] &\leq \sum_{r \in \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}}\right] = \sum_{r \in \text{ring}(Y_i)} \mathbf{Pr}[r \notin S_i] \\
&= \sum_r \mathbf{Pr}[r \notin S_i \cap (r \in \text{ring}(Y_i))] \\
&= \sum_r \mathbf{Pr}\left[r \notin S_i \;\middle|\; r \in \text{ring}(Y_i)\right] \mathbf{Pr}[r \in \text{ring}(Y_i)] \\
&\leq \sum_r \frac{\gamma}{8} \mathbf{Pr}[r \in \text{ring}(Y_i)] \leq \frac{\gamma}{8},
\end{aligned}$$

by Lemma 27.2.1. ■

Lemma 27.2.5. *For any i, we have* $\mu = \mathbf{E}[W_i] \leq \gamma/4$*, where* γ *is the parameter of Theorem 27.1.3, where* W_i *is the probability of failure to recover* Y_i *if it was sent, see Eq. (27.1).*

Proof: We have by Eq. (27.2) that $W_i = \sum_r \overline{S_{i,r}} w(Y_i, r)$. For a fixed value of Y_i, we have by linearity of expectation, that

$$\begin{aligned}
\mathbf{E}\left[W_i \;\middle|\; Y_i\right] &= \mathbf{E}\left[\sum_r \overline{S_{i,r}} w(Y_i, r) \;\middle|\; Y_i\right] = \sum_r \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r) \;\middle|\; Y_i\right] \\
&= \sum_{r \in \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r) \;\middle|\; Y_i\right] + \sum_{r \notin \text{ring}(Y_i)} \mathbf{E}\left[\overline{S_{i,r}} w(Y_i, r) \;\middle|\; Y_i\right] = g(Y_i) + f(Y_i) \leq \frac{\gamma}{8} + \frac{\gamma}{8} = \frac{\gamma}{4},
\end{aligned}$$

by Lemma 27.2.3 and Lemma 27.2.4. Now $\mathbf{E}[W_i] = \mathbf{E}\left[\mathbf{E}\left[W_i \;\middle|\; Y_i\right]\right] \leq \mathbf{E}[\gamma/4] \leq \gamma/4$. ■

In the following, we need the following trivial (but surprisingly deep) observation.

Observation 27.2.6. *For a random variable X, if* $\mathbf{E}[X] \leq \psi$*, then there exists an event in the probability space, that assigns X a value* $\leq \psi$*.*

Lemma 27.2.7. *For the codewords* $X_0, \ldots, X_K$*, the probability of failure in recovering them when sending them over the noisy channel is at most* γ*.*

Proof: We just proved that when using $Y_0, \ldots, Y_{\widehat{K}}$, the expected probability of failure when sending Y_i, is $\mathbf{E}[W_i] \le \gamma/4$, where $\widehat{K} = 2^{k+1} - 1$. As such, the expected total probability of failure is

$$\mathbf{E}\left[\sum_{i=0}^{\widehat{K}} W_i\right] = \sum_{i=0}^{\widehat{K}} \mathbf{E}[W_i] \le \frac{\gamma}{4} 2^{k+1} \le \gamma 2^k,$$

by Lemma 27.2.5. As such, by Observation 27.2.6, there exist a choice of Y_is, such that

$$\sum_{i=0}^{\widehat{K}} W_i \le 2^k \gamma.$$

Now, we use a similar argument used in proving Markov's inequality. Indeed, the W_i are always positive, and it can not be that 2^k of them have value larger than γ, because in the summation, we will get that

$$\sum_{i=0}^{\widehat{K}} W_i > 2^k \gamma.$$

Which is a contradiction. As such, there are 2^k codewords with failure probability smaller than γ. We set the 2^k codewords $X_0, \ldots, X_K$ to be these words, where $K = 2^k - 1$. Since we picked only a subset of the codewords for our code, the probability of failure for each codeword shrinks, and is at most γ. ■

Lemma 27.2.7 concludes the proof of the constructive part of Shannon's theorem.

27.2.2. Lower bound on the message size

We omit the proof of this part. It follows similar argumentation showing that for every ring associated with a codewords it must be that most of it is covered only by this ring (otherwise, there is no hope for recovery). Then an easy packing argument implies the claim.

27.3. From previous lectures

Lemma 27.3.1. *Suppose that nq is integer in the range $[0, n]$. Then* $\dfrac{2^{n\mathbb{H}(q)}}{n+1} \le \binom{n}{nq} \le 2^{n\mathbb{H}(q)}$.

Lemma 27.3.1 can be extended to handle non-integer values of q. This is straightforward, and we omit the easy details.

Corollary 27.3.2. *We have:*
(i) $q \in [0, 1/2] \Rightarrow \binom{n}{\lfloor nq \rfloor} \le 2^{n\mathbb{H}(q)}$. *(ii)* $q \in [1/2, 1]$ $\binom{n}{\lceil nq \rceil} \le 2^{n\mathbb{H}(q)}$.
(iii) $q \in [1/2, 1] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \le \binom{n}{\lfloor nq \rfloor}$. *(iv)* $q \in [0, 1/2] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \le \binom{n}{\lceil nq \rceil}$.

Theorem 27.3.3. *Suppose that the value of a random variable X is chosen uniformly at random from the integers $\{0, \ldots, m-1\}$. Then there is an extraction function for X that outputs on average at least $\lfloor \lg m \rfloor - 1 = \lfloor \mathbb{H}(X) \rfloor - 1$ independent and unbiased bits.*

27.4. Bibliographical Notes

The presentation here follows [MU05, Sec. 9.1-Sec 9.3].

Bibliography

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.

Chapter 28

Low Dimensional Linear Programming

By Sariel Har-Peled, December 30, 2015[①]

> "Napoleon has not been conquered by man. He was greater than all of us. But god punished him because he relied on his own intelligence alone, until that prodigious instrument was strained to breaking point. Everything breaks in the end."
>
> – Carl XIV Johan, King of Sweden.

28.1. Linear programming in constant dimension ($d > 2$)

Let assume that we have a set H of n linear inequalities defined over d (d is a small constant) variables. Every inequality in H defines a closed half space in $\mathbb{R}^d$. Given a vector $\overrightarrow{c} = (c_1, \ldots, c_d)$ we want to find $p = (p_1, \ldots, p_d) \in \mathbb{R}^d$ which is in all the half spaces $h \in H$ and $f(p) = \sum_i c_i p_i$ is maximized. Formally:

LP in d dimensions:$(H, \overrightarrow{c})$
H - set of n closed half spaces in $\mathbb{R}^d$
$\overrightarrow{c}$ - vector in d dimensions
Find $p \in \mathbb{R}^d$ s.t. $\forall h \in H$ we have $p \in h$ and $f(p)$ is maximized.
Where $f(p) = \left\langle p, \overrightarrow{c} \right\rangle$.

A closed half space in d dimensions is defined by an inequality of the form

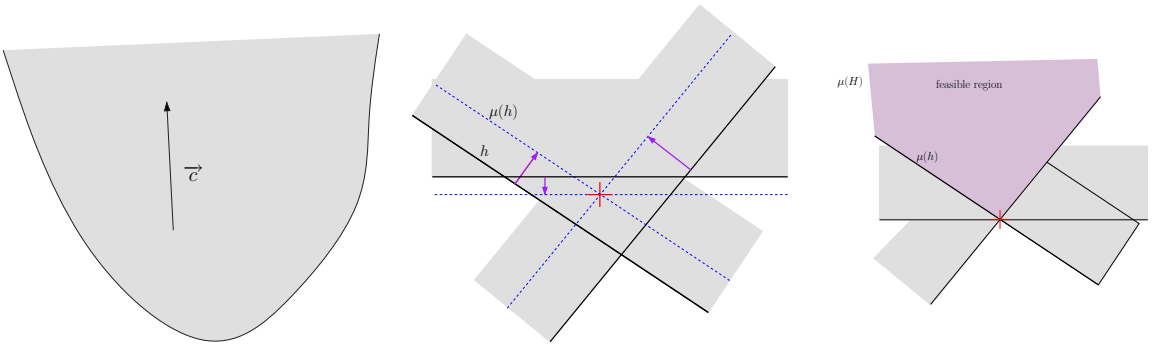
$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \leq b_n.$$

One difficulty that we ignored earlier, is that the optimal solution for the LP might be unbounded, see Figure 28.1.

Namely, we can find a solution with value ∞ to the target function.

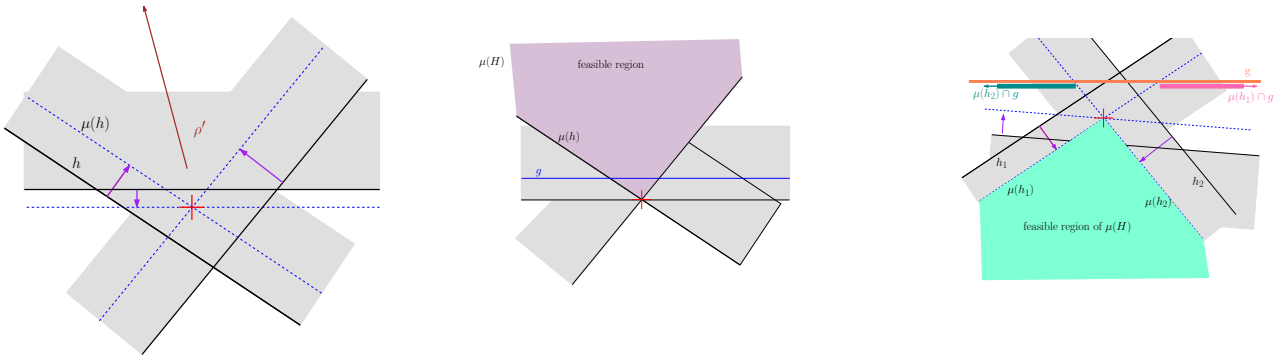
For a half space h let $\eta(h)$ denote the normal of h directed into the feasible region. Let $\mu(h)$ denote the closed half space, resulting from h by translating it so that it passes through the origin. Let $\mu(H)$ be the resulting set of half spaces from H. See Figure 28.1 (b).

The new set of constraints $\mu(H)$ is depicted in Figure 28.1 (c).



(a) (b) (c)

Figure 28.1: (a) Unbounded LP. (b). (c).



(a) (b) (c)

Figure 28.2: (a). (b). (c).

Lemma 28.1.1. *$(H, \overrightarrow{c})$ is unbounded if and only if $(\mu(H), \overrightarrow{c})$ is unbounded.*

Proof: Consider the ρ' the unbounded ray in the feasible region of $(H, \overrightarrow{c})$ such that the line that contain it passes through the origin. Clearly, ρ' is unbounded also in $(H, \overrightarrow{c})$, and this is if and only if. See Figure 28.2 (a). ■

Lemma 28.1.2. *Deciding if $(\mu(H), \overrightarrow{c})$ is bounded can be done by solving a $d-1$ dimensional LP. Furthermore, if it is bounded, then we have a set of d constraints, such that their intersection prove this.*

Furthermore, the corresponding set of d constraints in H testify that $(H, \overrightarrow{c})$ is bounded.

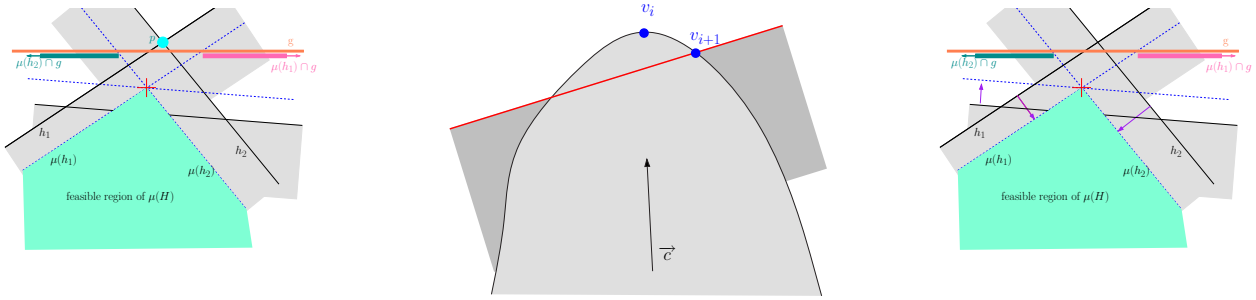
Proof: Rotate space, such that $\overrightarrow{c}$ is the vector $(0, 0, \ldots, 0, 1)$. And consider the hyperplane $g \equiv x_d = 1$. Clearly, $(\mu(H), \overrightarrow{c})$ is unbounded if and only if the region $g \cap \bigcap_{h \in \mu(H)} h$ is non-empty. By deciding if this region is unbounded, is equivalent to solving the following LP: $L' = (H', (1, 0, \ldots, 0))$ where

$$H' = \left\{ g \cap h \;\middle|\; h \in \mu(H) \right\}.$$

Let $h \equiv a_1 x_1 + \ldots + a_d x_d \leq 0$, the region corresponding to $g \cap h$ is $a_1 x_1 + \cdots + a_{d-1} x_{d-1} \leq -a_d$ which is a $d-1$ dimensional hyperplane. See Figure 28.2 (b).

But this is a $d-1$ dimensional LP, because everything happens on the hyperplane $x_d = 1$.

Notice that if $(\mu(H), \overrightarrow{c})$ is bounded (which happens if and only if $(H, \overrightarrow{c})$ is bounded), then L' is infeasible, and the LP L' would return us a set d constraints that their intersection is empty. Interpreting those constraints



(a) (b) (c)

Figure 28.3: (a). (b). (c).

in the original LP, results in a set of constraints that their intersection is bounded in the direction of $\overrightarrow{c}$. See Figure 28.2 (c).

(In the above example, $\mu(H) \cap g$ is infeasible because the intersection of $\mu(h_2) \cap g$ and $\mu(h_1) \cap g$ is empty, which implies that $h_1 \cap h_2$ is bounded in the direction $\overrightarrow{c}$ which we care about. The positive y direction in this figure.) ■

We are now ready to show the algorithm for the LP for $L = (H, \overrightarrow{c})$. By solving a $d-1$ dimensional LP we decide whether L is unbounded. If it is unbounded, we are done (we also found the unbounded solution, if you go carefully through the details).

See Figure 28.3 (a).

(in the above figure, we computed p.)

In fact, we just computed a set $h_1, \ldots, h_d$ s.t. their intersection is bounded in the direction of $\overrightarrow{c}$ (thats what the boundness check returned).

Let us randomly permute the remaining half spaces of H, and let $h_1, h_2, \ldots, h_d, h_{d+1}, \ldots, h_n$ be the resulting permutation.

Let v_i be the vertex realizing the optimal solution for the LP:

$$L_i = \left(\{h_1, \ldots, h_i\}, \overrightarrow{c}\right)$$

There are two possibilities:

1. $v_i = v_{i+1}$. This means that $v_i \in h_{i+1}$ and it can be checked in constant time.
2. $v_i \neq v_{i+1}$. It must be that $v_i \notin h_{i+1}$ but then, we must have... What is depicted in Figure 28.3 (b).

B - the set of d constraints that define v_{i+1}. If $h_{i+1} \notin B$ then $v_i = v_{i+1}$. As such, the probability of $v_i \neq v_{i+1}$ is roughly d/i because this is the probability that one of the elements of B is h_{i+1}. Indeed, fix the first $i+1$ elements, and observe that there are d elements that are marked (those are the elements of B). Thus, we are asking what is the probability of one of d marked elements to be the last one in a random permutation of $h_{d+1}, \ldots, h_{i+1}$, which is exactly $d/(i+1-d)$.

Note that if some of the elements of B is $h_1, \ldots, h_d$ than the above expression just decreases (as there are less marked elements).

Well, let us restrict our attention to ∂h_{i+1}. Clearly, the optimal solution to L_{i+1} on h_{i+1} is the required v_{i+1}. Namely, we solve the LP $L_{i+1} \cap h_{i+1}$ using recursion.

This takes $T(i+1, d-1)$ time. What is the probability that $v_{i+1} \neq v_i$?

Well, one of the d constraints defining v_{i+1} has to be h_{i+1}.The probability for that is ≤ 1 for $i \leq 2d-1$, and it is

$$\leq \frac{d}{i+1-d},$$

otherwise.

Summarizing everything, we have:

$$\begin{aligned} T(n,d) &= O(n) + T(n,d-1) + \sum_{i=d+1}^{2d} T(i,d-1) \\ &+ \sum_{i=2d+1}^{n} \frac{d}{i+1-d} T(i,d-1) \end{aligned}$$

What is the solution of this monster? Well, one essentially to guess the solution and verify it. To guess solution, let us "simplify" (incorrectly) the recursion to :

$$T(n,d) = O(n) + T(n,d-1) + d \sum_{i=2d+1}^{n} \frac{T(i,d-1)}{i+1-d}$$

So think about the recursion tree. Now, every element in the sum is going to contribute a near constant factor, because we divide it by (roughly) $i+1-d$ and also, we are guessing the the optimal solution is linear/near linear.

In every level of the recursion we are going to penalized by a multiplicative factor of d. Thus, it is natural, to conjecture that $T(n,d) \leq (3d)^{3d} n$.

Which can be verified by tedious substitution into the recurrence, and is left as exercise.

Theorem 28.1.3. *Given an d dimensional LP $(H, \overrightarrow{c})$,it can be solved in expected $O\Big((3d)^{3d} n\Big)$ time (the constant in the O is dim independent).*

BTW, we are being a bit conservative about the constant. In fact, one can prove that the running time is $d!n$. Which is still exponential in d.

```
SolveLP((H, c⃗))
    /* initialization */
    Rotate (H, c⃗) s.t. c⃗ = (0, ..., 1)
    Solve recursively the d − 1 dim LP:
            L′ ≡ μ(H) ∩ (x_d = 1)
    if L′ has a solution then
        return "Unbounded"

    Let g_1, ..., g_d be the set of constraints of L′ that testifies that L′ is infeasible
    Let h_1, ..., h_d be the hyperplanes of H corresponding to g_1, ..., g_d
    Permute H s.t. h_1, ..., h_d are first.
    v_d = ∂h_1 ∩ ∂h_2 ∩ ··· ∩ ∂h_d
    /*v_d is a vertex that testifies that (H, c⃗) is bounded */

    /* the algorithm itself */
    for i ← d + 1 to n do
        if v_{i−1} ∈ h_i then
            v_i ← v_{i−1}
        else
            v_i ← SolveLP((H_{i−1} ∩ ∂h_i, c⃗))      (*)
              where H_{i−1} = {h_1, ..., h_{i−1}}

    return v_n
```

28.2. Handling Infeasible Linear Programs

In the above discussion, we glossed over the question of how to handle LPs which are infeasible. This requires slightly modifying our algorithm to handle this case, and I am only describing the required modifications.

First, the simplest case, where we are given an LP L which is one dimensional (i.e., defined over one variable). Clearly, we can solve this LP in linear time (verify!), and furthermore, if there is no solution, we can return two input inequality $ax \leq b$ and $cx \geq d$ for which there is no solution together (i.e., those two inequalities [i.e., constraints] testifies that the LP is not satisfiable).

Next, assume that the algorithm `SolveLP` when called on a $d-1$ dimensional LP L', if L' is not feasible it return the d constraints of L' that together have non-empty intersection. Namely, those constraints are the witnesses that L' is infeasible.

So the only place, where we can get such answer, is when computing v_i (in the (*) line in the algorithm). Let $h'_1, \ldots, h'_d$ be the corresponding set of d constraints of H_{i-1} that testifies that $(H_{i-1} \cap \partial h_i, \overrightarrow{c})$ is an infeasible LP. Clearly, $h'_1, \ldots, h'_d, h_i$ must be a set of $d+1$ constraints that are together are infeasible, and that is what `SolveLP` returns.

28.3. References

The description in this class notes is loosely based on the description of low dimensional LP in the book of de Berg *et al.* [dBCKO08].

Bibliography

[dBCKO08] M. de Berg, O. Cheong, M. van Kreveld, and M. H. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Santa Clara, CA, USA, 3rd edition, 2008.

Chapter 29

Expanders I

By Sariel Har-Peled, December 30, 2015[①]

> "Mr. Matzerath has just seen fit to inform me that this partisan, unlike so many of them, was an authentic partisan. For - to quote the rest of my patient's lecture - there is no such thing as a part-time partisan. Real partisans are partisans always and as long as they live. They put fallen governments back in power and over throw governments that have just been put in power with the help of partisans. Mr. Matzerath contended - and this thesis struck me as perfectly plausible - that among all those who go in for politics your incorrigible partisan, who undermines what he has just set up, is closest to the artiest because he consistently rejects what he has just created."
>
> – Gunter Grass, The tin drum.

29.1. Preliminaries on expanders

29.1.1. Definitions

Let $\mathsf{G} = (V, E)$ Be an undirected graph, where $V = \{1, \ldots, n\}$. A ***d-regular graph*** is a graph where all vertices have degree d. A d-regular graph $\mathsf{G} = (V, E)$ is a δ-edge expander (or just, *δ-expander*) if for every set $S \subseteq V$ of size at most $|V|/2$, there are at least $\delta d |S|$ edges connecting S and $\overline{S} = V \setminus S$; that is

$$e(S, \overline{S}) \geq \delta d |S|, \tag{29.1}$$

where

$$e(X, Y) = \left| \left\{ uv \;\middle|\; u \in X, v \in Y \right\} \right|.$$

A graph is $[n, d, \delta]$***-expander*** if it is a n vertex, d-regular, δ-expander.

A (n, d)*-graph* G is a connected d-regular undirected (multi) graph. We will consider the set of vertices of such a graph to be the set $[\![n]\!] = \{1, \ldots, n\}$.

For a (multi) graph G with n nodes, its *adjacency matrix* is a $n \times n$ matrix M, where M_{ij} is the number of edges between i and j. It would be convenient to work the *transition matrix* Q associated with the random walk on G. If G is d-regular then $\mathsf{Q} = \mathsf{M}(\mathsf{G})/d$ and it is doubly stochastic.

A vector x is *eigenvector* of a matrix M with *eigenvalue* μ, if $x\mathsf{M} = \mu x$. In particular, by taking the dot product of both size by x, we get $\langle x\mathsf{M}, x\rangle = \langle \mu x, x\rangle$, which implies $\mu = \langle x\mathsf{M}, x\rangle / \langle x, x\rangle$. Since the adjacency

matrix M of G is symmetric, all its eigenvalues are real numbers (this is a special case of the spectral theorem from linear algebra). Two eigenvectors with different eigenvectors are orthogonal to each other.

We denote the eigenvalues of M by $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \cdots \widehat{\lambda_n}$, and the eigenvalues of Q by $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \cdots \widehat{\lambda_n}$. Note, that for a d-regular graph, the eigenvalues of Q are the eigenvalues of M scaled down by a factor of $1/d$; that is $\widehat{\lambda_i} = \widehat{\lambda_i}/d$.

Lemma 29.1.1. *Let G be an undirected graph, and let Δ denote the maximum degree in G. Then, $\left|\widehat{\lambda_1}(\mathsf{G})\right| = \left|\widehat{\lambda_1}(\mathsf{M})\right| = \Delta$ if and only one connected component of G is Δ-regular. The multiplicity of Δ as an eigenvector is the number of Δ-regular connected components. Furthermore, we have $\left|\widehat{\lambda_i}(\mathsf{G})\right| \leq \Delta$, for all i.*

Proof: The ith entry of $\mathsf{M}\mathbf{1}_n$ is the degree of the ith vertex v_i of G (i.e., $\mathsf{M}\mathbf{1}_n = d(v_i)$, where $\mathbf{1}_n = (1, 1, \ldots, 1) \in \mathbb{R}^n$. So, let x be an eigenvector of M with eigenvalue λ, and let $x_j \neq 0$ be the coordinate with the largest (absolute value) among all coordinates of x corresponding to a connect component H of G. We have that

$$|\lambda| \left|x_j\right| = \left|(\mathsf{M}x)_j\right| = \left|\sum_{v_i \in N(v_j)} x_i\right| \leq \Delta \left|x_j\right|,$$

where $N(v_j)$ are the neighbors of v_i in G. Thus, all the eigenvalues of G have $\left|\widehat{\lambda_i}\right| \leq \Delta$, for $i = 1, \ldots, n$. If $\lambda = \Delta$, then this implies that $x_i = x_j$ if $v_i \in N(v_j)$, and $d(v_j) = \Delta$. Applying this argument to the vertices of $N(v_j)$, implies that H must be Δ-regular, and furthermore, $x_j = x_i$, if $x_i \in V(H)$. Clearly, the dimension of the subspace with eigenvalue (in absolute value) Δ is exactly the number of such connected components. ■

The following is also known. We do not provide a proof since we do not need it in our argumentation.

Lemma 29.1.2. *If G is bipartite, then if λ is eigenvalue of $\mathsf{M}(\mathsf{G})$ with multiplicity k, then $-\lambda$ is also its eigenvalue also with multiplicity k.*

29.2. Tension and expansion

Let $\mathsf{G} = (V, E)$, where $V = \{1, \ldots, n\}$ and G is a d regular graph.

Definition 29.2.1. For a graph G, let $\gamma(\mathsf{G})$ denote the *tension* of G; that is, the smallest constant, such that for any function $f : V(\mathsf{G}) \to \mathbb{R}$, we have that

$$\mathop{\mathbf{E}}_{x,y \in V}\left[|f(x) - f(y)|^2\right] \leq \gamma(\mathsf{G}) \mathop{\mathbf{E}}_{xy \in E}\left[|f(x) - f(y)|^2\right]. \tag{29.2}$$

Intuitively, the tension captures how close is estimating the variance of a function defined over the vertices of G, by just considering the edges of G. Note, that a disconnected graph would have infinite tension, and the clique has tension 1.

Surprisingly, tension is directly related to expansion as the following lemma testifies.

Lemma 29.2.2. *Let $\mathsf{G} = (V, E)$ be a given connected d-regular graph with n vertices. Then, G is a δ-expander, where $\delta \geq \dfrac{1}{2\gamma(\mathsf{G})}$ and $\gamma(\mathsf{G})$ is the tension of G.*

Proof: Consider a set $S \subseteq V$, where $|S| \leq n/2$. Let $f_S(v)$ be the function assigning 1 if $v \in S$, and zero otherwise. Observe that if $(u,v) \in \left(S \times \overline{S}\right) \cup \left(\overline{S} \times S\right)$ then $|f_S(u) - f_S(v)| = 1$, and $|f_S(u) - f_S(v)| = 0$ otherwise. As such, we have

$$\frac{2|S|(n-|S|)}{n^2} = \mathop{\mathbf{E}}_{x,y\in V}\left[|f_S(x) - f_S(y)|^2\right] \leq \gamma(\mathsf{G}) \mathop{\mathbf{E}}_{xy\in E}\left[|f_S(x) - f_S(y)|^2\right] = \gamma(\mathsf{G})\frac{e\left(S,\overline{S}\right)}{|E|},$$

by Lemma 29.2.4. Now, since G is d-regular, we have that $|E| = nd/2$. Furthermore, $n - |S| \geq n/2$, which implies that

$$e\left(S,\overline{S}\right) \geq \frac{2|E| \cdot |S| \ (n-|S|)}{\gamma(\mathsf{G})n^2} = \frac{2(nd/2)(n/2)|S|}{\gamma(\mathsf{G})n^2} = \frac{1}{2\gamma(\mathsf{G})} d|S|.$$

which implies the claim (see Eq. (29.1)). ■

Now, a clique has tension 1, and it has the best expansion possible. As such, the smaller the tension of a graph, the better expander it is.

Definition 29.2.3. Given a random walk matrix Q associated with a d-regular graph, let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ denote the *orthonormal eigenvector basis* defined by Q. That is, $v_1, \ldots, v_n$ is an orthonormal basis for $\mathbb{R}^n$, where all these vectors are eigenvectors of Q and $v_1 = 1^n/\sqrt{n}$. Furthermore, let $\widehat{\lambda}_i$ denote the ith eigenvalue of Q, associated with the eigenvector v_i, such that $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n$.

Lemma 29.2.4. *Let $\mathsf{G} = (V,E)$ be a given connected d-regular graph with n vertices. Then $\gamma(\mathsf{G}) = \frac{1}{1-\widehat{\lambda}_2}$, where $\widehat{\lambda}_2 = \lambda_2/d$ is the second largest eigenvalue of Q.*

Proof: Let $f : V \to \mathbb{R}$. Since in Eq. (29.2), we only look on the difference between two values of f, we can add a constant to f, and would not change the quantities involved in Eq. (29.2). As such, we assume that $\mathbf{E}[f(x)] = 0$. As such, we have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{x,y\in V}\left[|f(x) - f(y)|^2\right] &= \mathop{\mathbf{E}}_{x,y\in V}\left[(f(x) - f(y))^2\right] = \mathop{\mathbf{E}}_{x,y\in V}\left[(f(x))^2 - 2f(x)f(y) + (f(y))^2\right] \\
&= \mathop{\mathbf{E}}_{x,y\in V}\left[(f(x))^2\right] - 2\mathop{\mathbf{E}}_{x,y\in V}[f(x)f(y)] + \mathop{\mathbf{E}}_{x,y\in V}\left[(f(y))^2\right] \\
&= \mathop{\mathbf{E}}_{x\in V}\left[(f(x))^2\right] - 2\mathop{\mathbf{E}}_{x\in V}[f(x)]\mathop{\mathbf{E}}_{y\in V}[f(y)] + \mathop{\mathbf{E}}_{y\in V}\left[(f(y))^2\right] = 2\mathop{\mathbf{E}}_{x\in V}\left[(f(x))^2\right].
\end{aligned} \tag{29.3}$$

Now, let $\mathcal{I}$ be the $n \times n$ identity matrix (i.e., one on its diagonal, and zero everywhere else). We have that

$$\begin{aligned}
\rho &= \frac{1}{d}\sum_{xy\in E}(f(x) - f(y))^2 = \frac{1}{d}\left(\sum_{x\in V} d(f(x))^2 - 2\sum_{xy\in E} f(x)f(y)\right) = \sum_{x\in V}(f(x))^2 - \frac{2}{d}\sum_{xy\in E} f(x)f(y) \\
&= \sum_{x,y\in V}(\mathcal{I} - \mathsf{Q})_{xy} f(x)f(y).
\end{aligned}$$

Note, that 1^n is an eigenvector of Q with eigenvalue 1, and this is the largest eigenvalue of Q. Let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ be the orthonormal eigenvector basis defined by Q, with eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n$, respectively. Write $f = \sum_{i=1}^n \alpha_i v_i$, and observe that

$$0 = \mathbf{E}[f(x)] = \sum_{i=1}^n \frac{f(i)}{n} = \left\langle f, \frac{v_1}{\sqrt{n}} \right\rangle = \left\langle \sum_i \alpha_i v_i, \frac{v_1}{\sqrt{n}} \right\rangle = \frac{1}{\sqrt{n}}\langle \alpha_1 v_1, v_1 \rangle = \frac{\alpha_1}{\sqrt{n}},$$

since $v_i \perp v_1$ for $i \geq 2$. Hence $\alpha_1 = 0$, and we have

$$\begin{aligned}\rho &= \sum_{x,y\in V} (\mathcal{I} - \mathsf{Q})_{xy} f(x) f(y) = \sum_{x,y\in V} (\mathcal{I} - \mathsf{Q})_{xy} \sum_{i=2}^{n} \alpha_{i=1}^{n} \alpha_i v_i(x) \sum_{j=1}^{n} \alpha_j v_j(y) \\ &= \sum_{i,j} \alpha_i \alpha_j \sum_{x\in V} v_i(x) \sum_{y\in V} (\mathcal{I} - \mathsf{Q})_{xy} v_j(y).\end{aligned}$$

Now, we have that

$$\sum_{y\in V} (\mathcal{I} - \mathsf{Q})_{xy} v_j(y) = \left\langle \begin{bmatrix} x\text{th row of} \\ (\mathcal{I} - \mathsf{Q}) \end{bmatrix}, v_j \right\rangle = \Big((\mathcal{I} - \mathsf{Q}) v_j\Big)(x) = \Big(\Big(1 - \widehat{\lambda_j}\Big) v_j\Big)(x) = \Big(1 - \widehat{\lambda_j}\Big) v_j(x),$$

since v_j is eigenvector of Q with eigenvalue $\widehat{\lambda_j}$. Since $v_1, \ldots, v_n$ is an orthonormal basis, and $f = \sum_{i=1}^{n} \alpha_i v_i$, we have that $\|f\|^2 = \sum_j \alpha_j^2$. Going back to ρ, we have that

$$\begin{aligned}\rho &= \sum_{i,j} \alpha_i \alpha_j \sum_{x\in V} v_i(x) \Big(1 - \widehat{\lambda_j}\Big) v_j(x) = \sum_{i,j} \alpha_i \alpha_j \Big(1 - \widehat{\lambda_j}\Big) \sum_{x\in V} v_i(x) v_j(x) \\ &= \sum_{i,j} \alpha_i \alpha_j \Big(1 - \widehat{\lambda_j}\Big) \big\langle v_i, v_j \big\rangle = \sum_{j=1}^{n} \alpha_j^2 \Big(1 - \widehat{\lambda_j}\Big) \big\langle v_j, v_j \big\rangle \\ &\geq \Big(1 - \widehat{\lambda_2}\Big) \sum_{j=2}^{n} \alpha_j^2 \sum_{x\in V} \Big(v_j(x)\Big)^2 = \Big(1 - \widehat{\lambda_2}\Big) \sum_{j=2}^{n} \alpha_j^2 = \Big(1 - \widehat{\lambda_2}\Big) \|f\|^2 = \Big(1 - \widehat{\lambda_2}\Big) \sum_{j=1}^{n} (f(x))^2 \qquad (29.4) \\ &= n\Big(1 - \widehat{\lambda_2}\Big) \mathop{\mathbf{E}}_{x\in V}\Big[(f(x))^2\Big],\end{aligned}$$

since $\alpha_1 = 0$ and $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \cdots \geq \widehat{\lambda_n}$.

We are now ready for the kill. Indeed, by Eq. (29.3), and the above, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{x,y\in V}\Big[|f(x) - f(y)|^2\Big] = 2 \mathop{\mathbf{E}}_{x\in V}\Big[(f(x))^2\Big] \;&\leq\; \frac{2}{n\Big(1 - \widehat{\lambda_2}\Big)} \rho = \frac{2}{dn\Big(1 - \widehat{\lambda_2}\Big)} \sum_{xy\in E} (f(x) - f(y))^2 \\ &= \frac{1}{1 - \widehat{\lambda_2}} \cdot \frac{1}{|E|} \sum_{xy\in E} (f(x) - f(y))^2 = \frac{1}{1 - \widehat{\lambda_2}} \mathop{\mathbf{E}}_{xy\in E}\Big[|f(x) - f(y)|^2\Big].\end{aligned}$$

This implies that $\gamma(\mathsf{G}) \leq \dfrac{1}{1 - \widehat{\lambda_2}}$. Observe, that the inequality in our analysis, had risen from Eq. (29.4), but if we take $f = v_2$, then the inequality there holds with equality, which implies that $\gamma(\mathsf{G}) \geq \frac{1}{1 - \widehat{\lambda_2}}$, which implies the claim. ■

Lemma 29.2.2 together with the above lemma, implies that the expansion δ of a d-regular graph G is at least $\delta = 1/2\gamma(\mathsf{G}) = (1 - \lambda_2/d)/2$, where λ_2 is the second eigenvalue of the adjacency matrix of G. Since the tension of a graph is direct function of its second eigenvalue, we could either argue about the tension of a graph or its second eigenvalue when bounding the graph expansion.

Chapter 30

Expanders II

By Sariel Har-Peled, December 30, 2015[①]

> Be that as it may, it is to night school that I owe what education I possess; I am the first to own that it doesn't amount to much, though there is something rather grandiose about the gaps in it.
>
> – Gunter Grass, The tin drum.

30.1. Bi-tension

Our construction of good expanders, would use the idea of composing graphs together. To this end, in our analysis, we will need the notion of bi-tension. Let $\widetilde{\mathsf{E}}(\mathsf{G})$ be the set of *directed* edges of G; that is, every edge $xy \in E(\mathsf{G})$ appears twice as $(x \to y)$ and $(y \to x)$ in $\widetilde{\mathsf{E}}$.

Definition 30.1.1. For a graph G, let $\gamma_2(\mathsf{G})$ denote the *bi-tension* of G; that is, the smallest constant, such that for any two function $f, g : V(\mathsf{G}) \to \mathbb{R}$, we have that

$$\mathop{\mathbf{E}}_{x,y\in V}\left[|f(x)-g(y)|^2\right] \le \gamma_2(\mathsf{G}) \mathop{\mathbf{E}}_{(x\to y)\in\widetilde{\mathsf{E}}}\left[|f(x)-g(y)|^2\right]. \tag{30.1}$$

The proof of the following lemma is similar to the proof of Lemma 30.3.1. The proof is provided for the sake of completeness, but there is little new in it.

Lemma 30.1.2. *Let $\mathsf{G} = (V, E)$ be a connected d-regular graph with n vertices. Then $\gamma_2(\mathsf{G}) = \dfrac{1}{1-\widehat{\lambda}}$, where $\widehat{\lambda} = \widehat{\lambda}(\mathsf{G})$, where $\widehat{\lambda}(\mathsf{G}) = \max\left(\widehat{\lambda}_2, -\widehat{\lambda}_n\right)$, where $\widehat{\lambda}_i$ is the ith largest eigenvalue of the random walk matrix associated with G.*

Proof: We can assume that $\mathbf{E}[f(x)] = 0$. As such, we have that

$$\mathop{\mathbf{E}}_{x,y\in V}\left[|f(x)-g(y)|^2\right] = \mathop{\mathbf{E}}_{x,y\in V}\left[(f(x))^2\right] - 2\mathop{\mathbf{E}}_{x,y\in V}[f(x)g(y)] + \mathop{\mathbf{E}}_{y\in V}\left[(g(y))^2\right] = \mathop{\mathbf{E}}_{x,y\in V}\left[(f(x))^2\right] + \mathop{\mathbf{E}}_{y\in V}\left[(g(y))^2\right]. \tag{30.2}$$

Let Q be the matrix associated with the random walk on G (each entry is either zero or $1/d$), we have

$$\begin{aligned}\rho = \mathop{\mathbf{E}}_{(x\to y)\in\widetilde{\mathsf{E}}}\left[|f(x)-g(y)|^2\right] &= \frac{1}{nd}\sum_{(x\to y)\in\widetilde{\mathsf{E}}}(f(x)-g(y))^2 = \frac{1}{n}\sum_{x,y\in V}\mathsf{Q}_{xy}(f(x)-g(y))^2 \\ &= \frac{1}{n}\sum_{x\in V}\Big((f(x))^2+(g(x))^2\Big) - \frac{2}{n}\sum_{x,y\in V}\mathsf{Q}_{xy}f(x)g(y).\end{aligned}$$

Let $\mathcal{B}(\mathsf{Q}) = \langle v_1,\ldots,v_n\rangle$ be the orthonormal eigenvector basis defined by Q (see Definition 30.3.3), with eigenvalues $\widehat{\lambda_1}\geq\widehat{\lambda_2}\geq\cdots\geq\widehat{\lambda_n}$, respectively. Write $f=\sum_{i=1}^n\alpha_i v_i$ and $g=\sum_{i=1}^n\beta_i v_i$. Since $\mathbf{E}[f(x)]=0$, we have that $\alpha_1=0$. Now, $\mathsf{Q}_{xy}=\mathsf{Q}_{yx}$, and we have

$$\begin{aligned}\sum_{x,y\in V}\mathsf{Q}_{xy}f(x)g(y) &= \sum_{x,y\in V}\mathsf{Q}_{yx}\left(\sum_i\alpha_i v_i(x)\right)\left(\sum_j\beta_j v_j(y)\right) = \sum_{i,j}\alpha_i\beta_j\sum_{y\in V}v_j(y)\sum_{x\in V}\mathsf{Q}_{yx}v_i(x) \\ &= \sum_{i,j}\alpha_i\beta_j\sum_{y\in V}v_j(y)\Big(\widehat{\lambda_i}v_i(y)\Big) = \sum_{i,j}\alpha_i\beta_j\widehat{\lambda_i}\left\langle v_j,v_i\right\rangle = \sum_{i=2}^n\alpha_i\beta_i\widehat{\lambda_i}\sum_{y\in V}(v_i(y))^2 \\ &\leq \widehat{\lambda}\sum_{i=2}^n\frac{\alpha_i^2+\beta_i^2}{2}\sum_{y\in V}(v_i(y))^2 \leq \frac{\widehat{\lambda}}{2}\sum_{i=1}^n\sum_{y\in V}\Big((\alpha_i v_i(y))^2+(\beta_i v_i(y))^2\Big) \\ &= \frac{\widehat{\lambda}}{2}\sum_{y\in V}\Big((f(y))^2+(g(y))^2\Big)\end{aligned}$$

As such,

$$\begin{aligned}\mathop{\mathbf{E}}_{(x\to y)\in\widetilde{\mathsf{E}}}\left[|f(x)-g(y)|^2\right] &= \frac{1}{nd}\sum_{(x\to y)\in\widetilde{\mathsf{E}}}|f(x)-g(y)|^2 = \frac{1}{n}\sum_{y\in V}\Big((f(y))^2+(g(y))^2\Big) - \frac{1}{n}\sum_{x,y\in V}\frac{2f(x)g(y)}{d} \\ &= \frac{1}{n}\sum_{y\in V}\Big((f(y))^2+(g(y))^2\Big) - \frac{2}{n}\sum_{x,y\in V}\mathsf{Q}_{xy}f(x)g(y) \\ &\geq \left(\frac{1}{n}-\frac{2}{n}\cdot\frac{\widehat{\lambda}}{2}\right)\sum_{y\in V}\Big((f(y))^2+(g(y))^2\Big) = \Big(1-\widehat{\lambda}\Big)\left(\mathop{\mathbf{E}}_{y\in V}\Big[(f(y))^2\Big]+\mathop{\mathbf{E}}_{y\in V}\Big[(g(y))^2\Big]\right) \\ &= \Big(1-\widehat{\lambda}\Big)\mathop{\mathbf{E}}_{x,y\in V}\Big[|f(x)-g(y)|^2\Big],\end{aligned}$$

by Eq. (30.2). This implies that $\gamma_2(\mathsf{G})\leq 1/\Big(1-\widehat{\lambda}\Big)$. Again, by trying either $f=g=v_2$ or $f=v_n$ and $g=-v_n$, we get that the inequality above holds with equality, which implies $\gamma_2(\mathsf{G})\geq 1/\Big(1-\widehat{\lambda}\Big)$. Together, the claim now follows. ■

30.2. Explicit construction

For a set $U\subseteq V$ of vertices, its *characteristic vector*, denoted by $x=\chi_U$, is the n dimensional vector, where $x_i=1$ if and only if $i\in U$.

The following is an easy consequence of Lemma 30.3.2.

Lemma 30.2.1. *For a d-regular graph G the vector $\mathbf{1}^n=(1,1,\ldots,1)$ is the only eigenvector with eigenvalue d (of the adjacency matrix $\mathsf{M}(\mathsf{G})$), if and only if G is connected. Furthermore, we have $|\lambda_i|\leq d$, for all i.*

Our main interest would be in the second largest eigenvalue of M. Formally, let

$$\lambda_2(\mathsf{G}) = \max_{x \perp \mathbf{1}^n, x \neq 0} \left| \frac{\langle x\mathsf{M}, x \rangle}{\langle x, x \rangle} \right|.$$

We state the following result but do not prove it since we do not need it for our nafarious purposes (however, we did prove the left side of the inequality).

Theorem 30.2.2. *Let G be a δ-expander with adjacency matrix M and let $\lambda_2 = \lambda_2(\mathsf{G})$ be the second-largest eigenvalue of M. Then*

$$\frac{1}{2}\left(1 - \frac{\lambda_2}{d}\right) \leq \delta \leq \sqrt{2\left(1 - \frac{\lambda_2}{d}\right)}.$$

What the above theorem says, is that the expansion of a $[n, d, \delta]$-expander is a function of how far is its second eigenvalue (i.e., λ_2) from its first eigenvalue (i.e., d). This is usually referred to as the *spectral gap*.

We will start by explicitly constructing an expander that has "many" edges, and then we will show to reduce its degree till it become a constant degree expander.

30.2.1. Explicit construction of a small expander

30.2.1.1. A quicky reminder of fields

A *field* is a set $\mathbb{F}$ together with two operations, called addition and multiplication, and denoted by $+$ and $\cdot$, respectively, such that the following axioms hold:

(i) Closure: $\forall x, y \in \mathbb{F}$, we have $x + y \in \mathbb{F}$ and $x \cdot y \in \mathbb{F}$.

(ii) Associativity: $\forall x, y, z \in \mathbb{F}$, we have $x + (y + z) = (x + y) + z$ and $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.

(iii) Commutativity: $\forall x, y \in \mathbb{F}$, we have $x + y = y + x$ and $x \cdot y = y \cdot x$.

(iv) Identity: There exists two distinct special elements $0, 1 \in \mathbb{F}$. We have that $\forall x \in \mathbb{F}$ it holds $x + 0 = a$ and $x \cdot 1 = x$.

(v) Inverse: There exists two distinct special elements $0, 1 \in \mathbb{F}$, and we have that $\forall x \in \mathbb{F}$ there exists an element $-x \in \mathbb{F}$, such that $x + (-x) = 0$.

Similarly, $\forall x \in \mathbb{F}$, $x \neq 0$, there exists an element $y = x^{-1} = 1/x \in \mathbb{F}$ such that $x \cdot y = 1$.

(vi) Distributivity: $\forall x, y, z \in \mathbb{F}$ we have $x \cdot (y + z) = x \cdot y + x \cdot z$.

Let $q = 2^t$, and $r > 0$ be an integer. Consider the finite field $\mathbb{F}_q$. It is the field of polynomials of degree at most $t - 1$, where the coefficients are over $\mathbb{Z}_2$ (i.e., all calculations are done modulus 2). Formally, consider the polynomial

$$p(x) = x^t + x + 1.$$

It it irreducible over $\mathbb{F}_2 = \{0, 1\}$ (i.e., $p(0) = p(1) \neq 0$). We can now do polynomial arithmetic over polynomials (with coefficients from $\mathbb{F}_2$), where we do the calculations modulus $p(x)$. Note, that any irreducible polynomial of degree n yields the same field up to isomorphism. Intuitively, we are introducing the n distinct roots of $p(x)$ into $\mathbb{F}$ by creating an extension field of $\mathbb{F}$ with those roots.

An element of $\mathbb{F}_q = \mathbb{F}_{2^t}$ can be interpreted as a binary string $b = b_0 b_1 \ldots, b_{t-1}$ of length t, where the corresponding polynomial is

$$\mathrm{poly}(b) = \sum_{i=0}^{t-1} b_i x^i.$$

The nice property of $\mathbb{F}_q$ is that addition can be interpreted as a `xor` operation. That is, for any $x, y \in \mathbb{F}_q$, we have that $x + y + y = x$ and $x - y - y = x$. The key properties of $\mathbb{F}_q$ we need is that multiplications and addition can be computed in it in polynomial time in t, and it is a field (i.e., each non-zero element has a unique inverse).

30.2.1.1.1. Computing multiplication in $\mathbb{F}_q$. Consider two elements $\alpha, \beta \in \mathbb{F}_q$. Multiply the two polynomials $\mathrm{poly}(\alpha)$ by $\mathrm{poly}(\beta)$, let $\mathrm{poly}(\gamma)$ be the resulting polynomial (of degree at most $2t-2$), and compute the remainder $\mathrm{poly}(\beta)$ when dividing it by the irreducible polynomial $p(x)$. For this remainder polynomial, normalize the coefficients by computing their modules base 2. The resulting polynomial is the product of α and β.

For more details on this field, see any standard text on abstract algebra.

30.2.1.2. The construction

Let $q = 2^t$, and $r > 0$ be an integer. Consider the linear space $\mathbb{G} = \mathbb{F}_q^r$. Here, a member $\alpha = (\alpha_0, \ldots, \alpha_r) \in \mathbb{G}$ can be thought of as being a string (of length $r+1$) over $\mathbb{F}_q$, or alternatively, as a binary string of length $n = t(r+1)$.

For $\alpha = (\alpha_0, \ldots, \alpha_r) \in \mathbb{G}$, and $x, y \in \mathbb{F}_q$, define the operator

$$\rho(\alpha, x, y) = \alpha + y \cdot \left(1, x, x^2, \ldots, x^r\right) = \left(\alpha_0 + y, \alpha_1 + yx, \alpha_2 + yx^2, \ldots, \alpha_r + yx^r\right) \in \mathbb{G}.$$

Since addition over $\mathbb{F}_q$ is equivalent to a xor operation we have that

$$\begin{aligned}\rho(\rho(\alpha, x, y), x, y) &= \left(\alpha_0 + y + y, \alpha_1 + yx + yx, \alpha_2 + yx^2 + yx^2, \ldots, \alpha_r + yx^r + yx^r\right) \\ &= (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_r) = \alpha.\end{aligned}$$

Furthermore, if $(x, y) \neq (x', y')$ then $\rho(\alpha, x, y) \neq \rho(\alpha, x', y')$.

We now define a graph $\mathsf{LD}(q, r) = (\mathbb{G}, E)$, where

$$E = \left\{ \alpha\beta \;\middle|\; \begin{array}{l} \alpha \in \mathbb{G}, x, y \in \mathbb{F}_q \\ \beta = \rho(\alpha, x, y) \end{array} \right\}$$

Note, that this graph is well defined, as $\rho(\beta, x, y) = \alpha$. The degree of a vertex of $\mathsf{LD}(q, r)$ is $\left|\mathbb{F}_q\right|^2 = q^2$, and $\mathsf{LD}(q, r)$ has $N = |\mathbb{G}| = q^{r+1} = 2^{t(r+1)} = 2^n$ vertices.

Theorem 30.2.3. *For any $t > 0, r > 0$ and $q = 2^t$, where $r < q$, we have that $\mathsf{LD}(q, r)$ is a graph with q^{r+1} vertices. Furthermore, $\lambda_1(\mathsf{LD}(q, r)) = q^2$, and $\lambda_i(\mathsf{LD}(q, r)) \leq rq$, for $i = 2, \ldots, n$.*

In particular, if $r \leq q/2$, then $\mathsf{LD}(q, r)$ is a $\left[q^{r+1}, q^2, \frac{1}{4}\right]$-expander.

Proof: Let M be the $N \times N$ adjacency matrix of $\mathsf{LD}(q, r)$. Let $L : \mathbb{F}_q \to \{0, 1\}$ be a linear map which is onto. It is easy to verify that $\left|L^{-1}(0)\right| = \left|L^{-1}(1)\right|$[2]

We are interested in the eigenvalues of the matrix M. To this end, we consider vectors in $\mathbb{R}^N$. The ith row an ith column of M is associated with a unique element $b_i \in \mathbb{G}$. As such, for a vector $v \in \mathbb{R}^N$, we denote by

[2] Indeed, if $Z = L^{-1}(0)$, and $L(x) = 1$, then $L(y) = 1$, for all $y \in U = \left\{x + z \;\middle|\; z \in Z\right\}$. Now, its clear that $|Z| = |U|$.

$v[b_i]$ the ith coordinate of v. In particular, for $\alpha = (\alpha_0, \ldots, \alpha_r) \in \mathbb{G}$, let $v_\alpha \in \mathbb{R}^N$ denote the vector, where its $\beta = (\beta_0, \ldots, \beta_r) \in \mathbb{G}$ coordinate is

$$v_\alpha[\beta] = (-1)^{L\left(\sum_{i=0}^{r} \alpha_i \beta_i\right)}.$$

Let $V = \left\{ v_\alpha \;\middle|\; \alpha \in \mathbb{G} \right\}$. For $\alpha \neq \alpha' \in V$, observe that

$$\langle v_\alpha, v_{\alpha'} \rangle = \sum_{\beta \in \mathbb{G}} (-1)^{L\left(\sum_{i=0}^{r} \alpha_i \beta_i\right)} \cdot (-1)^{L\left(\sum_{i=0}^{r} \alpha'_i \beta_i\right)} = \sum_{\beta \in \mathbb{G}} (-1)^{L\left(\sum_{i=0}^{r} \left(\alpha_i + \alpha'_i\right) \beta_i\right)} = \sum_{\beta \in \mathbb{G}} v_{\alpha+\alpha'}[\beta] \,.$$

So, consider $\psi = \alpha + \alpha' \neq 0$. Assume, for the simplicity of exposition that all the coordinates of ψ are non-zero. We have, by the linearity of L that

$$\langle v_\alpha, v_{\alpha'} \rangle = \sum_{\beta \in \mathbb{G}} (-1)^{L\left(\sum_{i=0}^{r} \alpha_i \beta_i\right)} = \sum_{\beta_0 \in \mathbb{F}_q, \ldots, \beta_{r-1} \in \mathbb{F}_q} (-1)^{L(\psi_0 \beta_0 + \cdots + \psi_{r-1} \beta_{r-1})} \sum_{\beta_r \in \mathbb{F}_q} (-1)^{L(\psi_r \beta_r)}.$$

However, since $\psi_r \neq 0$, the quantity $\left\{ \psi_r \beta_r \;\middle|\; \beta_r \in \mathbb{F}_q \right\} = \mathbb{F}_q$. Thus, the summation $\sum_{\beta_r \in \mathbb{F}_q} (-1)^{L(\psi_r \beta_r)}$ gets $\left| L^{-1}(0) \right|$ terms that are 1, and $\left| L^{-1}(0) \right|$ terms that are -1. As such, this summation is zero, implying that $\langle v_\alpha, v_{\alpha'} \rangle = 0$. Namely, the vectors of V are orthogonal.

Observe, that for $\alpha, \beta, \psi \in \mathbb{G}$, we have $v_\alpha[\beta + \psi] = v_\alpha[\beta]\, v_\alpha[\psi]$. For $\alpha \in \mathbb{G}$, consider the vector $\mathsf{M} v_\alpha$. We have, for $\beta \in \mathbb{G}$, that

$$\begin{aligned}
(\mathsf{M} v_\alpha)[\beta] = \sum_{\psi \in \mathbb{G}} \mathsf{M}_{\beta\psi} \cdot v_\alpha[\psi] = \sum_{\substack{x,y \,\in\, \mathbb{F}_q \\ \psi = \rho(\beta, x, y)}} v_\alpha[\psi] = \sum_{x,y \,\in\, \mathbb{F}_q} v_\alpha[\beta + y(1, x, \ldots, x^r)] \\
= \left(\sum_{x,y \,\in\, \mathbb{F}_q} v_\alpha[y(1, x, \ldots, x^r)] \right) \cdot v_\alpha[\beta] \,.
\end{aligned}$$

Thus, setting $\lambda(\alpha) = \sum_{x,y \,\in\, \mathbb{F}_q} v_\alpha[y(1, x, \ldots, x^r)] \in \mathbb{R}$, we have that $\mathsf{M} v_\alpha = \lambda(\alpha) \cdot v_\alpha$. Namely, v_α is an eigenvector, with eigenvalue $\lambda(\alpha)$.

Let $p_\alpha(x) = \sum_{i=0}^{r} \alpha_i x^i$, and let

$$\begin{aligned}
\lambda(\alpha) = \sum_{x,y \,\in\, \mathbb{F}_q} v_\alpha[y(1, x, \ldots, x^r)] \in \mathbb{R} = \sum_{x,y \in \mathbb{F}_q} (-1)^{L(y p_\alpha(x))} \\
= \sum_{\substack{x,y \in \mathbb{F}_q \\ p_\alpha(x) = 0}} (-1)^{L(y\ p_\alpha(x))} + \sum_{\substack{x,y \in \mathbb{F}_q \\ p_\alpha(x) \neq 0}} (-1)^{L(y\ p_\alpha(x))}.
\end{aligned}$$

If $p_\alpha(x) = 0$ then $(-1)^{L(y\ p_\alpha(x))} = 1$, for all y. As such, each such x contributes q to $\lambda(\alpha)$.

If $p_\alpha(x) \neq 0$ then $y\ p_\alpha(x)$ takes all the values of $\mathbb{F}_q$, and as such, $L(y\ p_\alpha(x))$ is 0 for half of these values, and 1 for the other half. Implying that these kind terms contribute 0 to $\lambda(\alpha)$. But $p_\alpha(x)$ is a polynomial of degree r, and as such there could be at most r values of x for which the first term is taken. As such, if $\alpha \neq 0$ then $\lambda(\alpha) \leq rq$. If $\alpha = 0$ then $\lambda(\alpha) = q^2$, which implies the theorem. ■

This construction provides an expander with constant degree only if the number of vertices is a constant. Indeed, if we want an expander with constant degree, we have to take q to be as small as possible. We get the relation $n = q^{r+1} \leq q^q$, since $r \leq r$, which implies that $q = \Omega(\log n / \log\log n)$. Now, the expander of Theorem 30.2.3 is q^2-regular, which means that it is not going to provide us with a constant degree expander.

However, we are going to use it as our building block in a construction that would start with this expander and would inflate it up to the desired size.

30.3. From previous lectures

Lemma 30.3.1. *Let* $\mathsf{G} = (V, E)$ *be a given connected* d*-regular graph with* n *vertices. Then* $\gamma(\mathsf{G}) = \frac{1}{1-\widehat{\lambda}_2}$*, where* $\widehat{\lambda}_2 = \lambda_2/d$ *is the second largest eigenvalue of* Q*.*

Lemma 30.3.2. *Let* G *be an undirected graph, and let* Δ *denote the maximum degree in* G*. Then,* $\left|\widehat{\lambda}_1(\mathsf{G})\right| = \left|\widehat{\lambda}_1(\mathsf{M})\right| = \Delta$ *if and only one connected component of* G *is* Δ*-regular. The multiplicity of* Δ *as an eigenvector is the number of* Δ*-regular connected components. Furthermore, we have* $\left|\widehat{\lambda}_i(\mathsf{G})\right| \leq \Delta$*, for all* i*.*

Definition 30.3.3. Given a random walk matrix Q associated with a d-regular graph, let $\mathcal{B}(\mathsf{Q}) = \langle v_1, \ldots, v_n \rangle$ denote the *orthonormal eigenvector basis* defined by Q. That is, $v_1, \ldots, v_n$ is an orthonormal basis for $\mathbb{R}^n$, where all these vectors are eigenvectors of Q and $v_1 = 1^n/\sqrt{n}$. Furthermore, let $\widehat{\lambda}_i$ denote the ith eigenvalue of Q, associated with the eigenvector v_i, such that $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n$.

Chapter 31

Expanders III - The Zig Zag Product

By Sariel Har-Peled, December 30, 2015[①]

> Gradually, but not as gradually as it seemed to some parts of his brain, he began to infuse his tones with a sarcastic wounding bitterness. Nobody outside a madhouse, he tried to imply, could take seriously a single phrase of this conjectural, nugatory, deluded, tedious rubbish. Within quite a short time he was contriving to sound like an unusually fanatical Nazi trooper in charge of a book-burning reading out to the crowd excerpts from a pamphlet written by a pacifist, Jewish, literate Communist. A growing mutter, half-amused, half-indignant, arose about him, but he closed his ears to it and read on. Almost unconsciously he began to adopt an unnameable foreign accent and to read faster and faster, his head spinning. As if in a dream he heard Welch stirring, then whispering, then talking at his side. he began punctuating his discourse with smothered snorts of derision. He read on, spitting out the syllables like curses, leaving mispronunciations, omissions, spoonerisms uncorrected, turning over the pages of his script like a score-reader following a presto movement, raising his voice higher and higher. At last he found his final paragraph confronting him, stopped, and look at his audience.
>
> – Kingsley Amis, Lucky Jim.

31.1. Building a large expander with constant degree

31.1.1. Notations

For a vertex $v \in V(\mathsf{G})$, we will denote by $v_{\mathsf{G}}[i] = v[i]$ the ith neighbor of v in the graph G (we order the neighbors of a vertex in an arbitrary order).

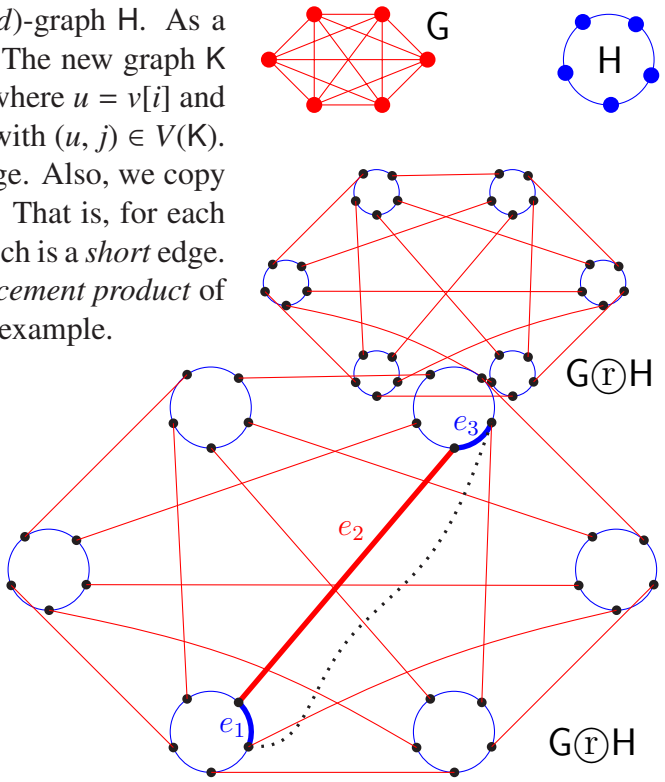
The regular graphs we next discuss have *consistent labeling*. That is, for a regular graph G (we assume here that G is regular). This means that if u is the ith neighbor v then v is the ith neighbor of u. Formally, this means that $v[i][i] = v$, for all v and i. This is a non-trivial property, but its easy to verify that the low quality expander of Theorem 31.4.3 has this property. It is also easy to verify that the complete graph can be easily be made into having consistent labeling (exercise). These two graphs would be sufficient for our construction.

31.1.2. The Zig-Zag product

At this point, we know how to construct a good "small" expander. The question is how to build a large expander (i.e., large number of vertices) and with constant degree.

The intuition of the construction is the following: It is easy to improve the expansion qualities of a graph by squaring it. The problem is that the resulting graph G has a degree which is too large. To overcome this, we will replace every vertex in G by a copy of a small graph that is connected and has low degree. For example, we could replace every vertex of degree d in G by a path having d vertices. Every such vertex is now in charge of original edge of the graph. Naturally, such a replacement operation reduces the quality of the expansion of the resulting graph. In this case, replacing a vertex with a path is a potential "disaster", since every such subpath increases the lengths of the paths of the original graph by a factor of d (and intuitively, a good expander have "short" paths between any pair of vertices).

Consider a "large" (n, D)-graph G and a "small" (D, d)-graph H. As a first stage, we replace every vertex of G by a copy of H. The new graph K has $[\![n]\!] \times [\![D]\!]$ as a vertex set. Here, the edge $vu \in V(\mathsf{G})$, where $u = v[i]$ and $v = u[j]$, is replaced by the edge connecting $(v, i) \in V(\mathsf{K})$ with $(u, j) \in V(\mathsf{K})$. We will refer to this resulting edge $(v, i)(u, j)$ as a *long* edge. Also, we copy all the edges of the small graph to each one of its copies. That is, for each $i \in [\![n]\!]$, and $uv \in E(\mathsf{H})$, we add the edge $(i, u)(i, v)$ to K, which is a *short* edge. We will refer to K, which is a $(nD, d + 1)$-graph, as a *replacement product* of G and H, denoted by $\mathsf{G} \textcircled{r} \mathsf{H}$. See figure on the right for an example.



Again, intuitively, we are losing because the expansion of the resulting graph had deteriorated too much. To overcome this problem, we will perform local shortcuts to shorten the paths in the resulting graph (and thus improve its expansion properties). A *zig-zag-zig path* in the replacement product graph K, is a three edge path $e_1e_2e_3$, where e_1 and e_3 are short edges, and the middle edge e_2 is a long edge. That is, if $e_1 = (i, u)(i, v)$, $e_2 = (i, v)(j, v')$, and $e_3 = (j, v')(j, u')$, then $e_1, e_2, e_3 \in E(\mathsf{K})$, $ij \in E(\mathsf{G})$, $uv \in E(\mathsf{H})$ and $v'u' \in E(\mathsf{H})$. Intuitively, you can think about e_1 as a small "zig" step in H, e_2 is a long "zag" step in G, and finally e_3 is a "zig" step in H.

Another way of representing a zig-zag-zig path $v_1v_2v_3v_4$ starting at the vertex $v_1 = (i, v) \in V(\mathsf{F})$, is to parameterize it by two integers $\ell, \ell' \in [\![d]\!]$, where

$$v_1 = (i, v), \quad v_2 = (i, v_{\mathsf{H}}[\ell]) \quad v_3 = (i_{\mathsf{G}}[v_{\mathsf{H}}[\ell]], v_{\mathsf{H}}[\ell]) \quad v_4 = (i_{\mathsf{G}}[v_{\mathsf{H}}[\ell]] \ , \ (v_{\mathsf{H}}[\ell])_{\mathsf{H}}[\ell']).$$

Let Z be the set of all (unordered) pairs of vertices of K connected by such a zig-zag-zig path. Note, that every vertex (i, v) of K has d^2 paths having (i, v) as an end point. Consider the graph $\mathsf{F} = (V(\mathsf{K}), Z)$. The graph F has nD vertices, and it is d^2 regular. Furthermore, since we shortcut all these zig-zag-zig paths in K, the graph F is a much better expander (intuitively) than K. We will refer to the graph F as the *zig-zag* product of G and H.

Definition 31.1.1. The *zig-zag product* of (n, D)-graph G and a (D, d)-graph H, is the (nD, d^2) graph $\mathsf{F} = \mathsf{G} \textcircled{z} \mathsf{H}$, where the set of vertices is $[\![n]\!] \times [\![D]\!]$ and for any $v \in [\![n]\!]$, $i \in [\![D]\!]$, and $\ell, \ell' \in [\![d]\!]$ we have in F the edge connecting the vertex (i, v) with the vertex $(i_{\mathsf{G}}[v_{\mathsf{H}}[\ell]] \ , \ (v_{\mathsf{H}}[\ell])_{\mathsf{H}}[\ell'])$.

Remark 31.1.2. We need the resulting zig-zag graph to have consistent labeling. For the sake of simplicity of exposition, we are just going to assume this property.

We next bound the tension of the zig-zag product graph.

Theorem 31.1.3. *We have* $\gamma(\mathsf{G} \circledcirc \mathsf{H}) \leq \gamma_2(\mathsf{G})(\gamma_2(\mathsf{H}))^2$. *and* $\gamma_2(\mathsf{G} \circledcirc \mathsf{H}) \leq \gamma_2(\mathsf{G})(\gamma_2(\mathsf{H}))^2$.

Proof: Let $\mathsf{G} = (\llbracket n \rrbracket, E)$ be a (n, D)-graph and $\mathsf{H} = (\llbracket D \rrbracket, E')$ be a (D, d)-graph. Fix any function $f : \llbracket n \rrbracket \times \llbracket D \rrbracket \to \mathbb{R}$, and observe that

$$\begin{aligned}
\psi &= \mathop{\mathbf{E}}_{\substack{u,v\in\llbracket n\rrbracket \\ k,\ell\in\llbracket D\rrbracket}} \left[|f(u,k) - f(v,\ell)|^2 \right] = \mathop{\mathbf{E}}_{k,\ell\in\llbracket D\rrbracket} \left[\mathop{\mathbf{E}}_{u,v\in\llbracket n\rrbracket} \left[|f(u,k) - f(v,\ell)|^2 \right] \right] \\
&\leq \mathop{\mathbf{E}}_{k,\ell\in\llbracket D\rrbracket} \left[\gamma_2(\mathsf{G}) \mathop{\mathbf{E}}_{uv\in E(\mathsf{G})} \left[|f(u,k) - f(v,\ell)|^2 \right] \right] = \gamma_2(\mathsf{G}) \underbrace{\mathop{\mathbf{E}}_{k,\ell\in\llbracket D\rrbracket} \left[\mathop{\mathbf{E}}_{\substack{u\in\llbracket n\rrbracket \\ p\in\llbracket D\rrbracket}} \left[\left| f(u,k) - f(u[p],\ell) \right|^2 \right] \right]}_{=\Delta_1}.
\end{aligned}$$

Now,

$$\begin{aligned}
\Delta_1 &= \mathop{\mathbf{E}}_{\substack{u\in\llbracket n\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\mathop{\mathbf{E}}_{k,p\in\llbracket D\rrbracket} \left[\left| f(u,k) - f(u[p],\ell) \right|^2 \right] \right] \leq \mathop{\mathbf{E}}_{\substack{u\in\llbracket n\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\gamma_2(\mathsf{H}) \mathop{\mathbf{E}}_{kp\in E(\mathsf{H})} \left[\left| f(u,k) - f(u[p],\ell) \right|^2 \right] \right] \\
&= \gamma_2(\mathsf{H}) \underbrace{\mathop{\mathbf{E}}_{\substack{u\in\llbracket n\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\mathop{\mathbf{E}}_{\substack{p\in\llbracket D\rrbracket \\ j\in\llbracket d\rrbracket}} \left[\left| f(u,p[j]) - f(u[p],\ell) \right|^2 \right] \right]}_{=\Delta_2}.
\end{aligned}$$

Now,

$$\begin{aligned}
\Delta_2 &= \mathop{\mathbf{E}}_{\substack{j\in\llbracket d\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\mathop{\mathbf{E}}_{\substack{u\in\llbracket n\rrbracket \\ p\in\llbracket D\rrbracket}} \left[\left| f(u,p[j]) - f(u[p],\ell) \right|^2 \right] \right] = \mathop{\mathbf{E}}_{\substack{j\in\llbracket d\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\mathop{\mathbf{E}}_{\substack{v\in\llbracket n\rrbracket \\ p\in\llbracket D\rrbracket}} \left[\left| f(v[p],p[j]) - f(v,\ell) \right|^2 \right] \right] \\
&= \mathop{\mathbf{E}}_{\substack{j\in\llbracket d\rrbracket \\ v\in\llbracket n\rrbracket}} \left[\mathop{\mathbf{E}}_{\substack{p\in\llbracket D\rrbracket \\ \ell\in\llbracket D\rrbracket}} \left[\left| f(v[p],p[j]) - f(v,\ell) \right|^2 \right] \right] \\
&= \gamma_2(\mathsf{H}) \underbrace{\mathop{\mathbf{E}}_{\substack{j\in\llbracket d\rrbracket \\ v\in\llbracket n\rrbracket}} \left[\mathop{\mathbf{E}}_{p\ell\in E(\mathsf{H})} \left[\left| f(v[p],p[j]) - f(v,\ell) \right|^2 \right] \right]}_{=\Delta_3}.
\end{aligned}$$

Now, we have

$$\Delta_3 = \mathop{\mathbf{E}}_{\substack{j\in\llbracket d\rrbracket \\ v\in\llbracket n\rrbracket}} \left[\mathop{\mathbf{E}}_{\substack{p\in\llbracket D\rrbracket \\ i\in\llbracket d\rrbracket}} \left[\left| f(v[p],p[j]) - f(v,p[i]) \right|^2 \right] \right] = \mathop{\mathbf{E}}_{(u,k)(\ell,v)\in E(\mathsf{G} \circledcirc \mathsf{H})} \left[|f(u,k) - f(\ell,v)| \right],$$

as $(v[p], p[j])$ is adjacent to $(v[p], p)$ (a short edge), which is in turn adjacent to (v, p) (a long edge), which is adjacent to $(v, p[i])$ (a short edge). Namely, $(v[p], p[j])$ and $(v, p[i])$ form the endpoints of a zig-zag path in the replacement product of G and H. That is, these two endpoints are connected by an edge in the zig-zag product graph. Furthermore, it is easy to verify that each zig-zag edge get accounted for in this representation exactly once, implying the above inequality. Thus, we have $\psi \leq \gamma_2(\mathsf{G})(\gamma_2(\mathsf{H}))^2\Delta_3$, which implies the claim.

The second claim follows by similar argumentation. ∎

31.1.3. Squaring

The last component in our construction, is *squaring*squaring!graph a graph. Given a (n,d)-graph G, consider the multigraph G^2 formed by connecting any vertices connected in G by a path of length 2. Clearly, if M is the adjacency matrix of G, then the adjacency matrix of G^2 is the matrix M^2. Note, that $\left(\mathsf{M}^2\right)_{ij}$ is the number of distinct paths of length 2 in G from i to j. Note, that the new graph might have self loops, which does not effect our analysis, so we keep them in.

Lemma 31.1.4. *Let G be a (n,d)-graph. The graph G^2 is a (n,d^2)-graph. Furthermore $\gamma_2\left(\mathsf{G}^2\right) = \frac{(\gamma_2(\mathsf{G}))^2}{2\gamma_2(\mathsf{G})-1}$.*

Proof: The graph G^2 has eigenvalues $\left(\widehat{\lambda}_1(\mathsf{G})\right)^2, \ldots, \left(\widehat{\lambda}_1(\mathsf{G})\right)^2$ for its matrix Q^2. As such, we have that

$$\widehat{\lambda}\left(\mathsf{G}^2\right) = \max\left(\widehat{\lambda}_2\left(\mathsf{G}^2\right), -\widehat{\lambda}_n\left(\mathsf{G}^2\right)\right).$$

Now, $\widehat{\lambda}_1\left(\mathsf{G}^2\right) = 1$. Now, if $\widehat{\lambda}_2(\mathsf{G}) \geq \left|\widehat{\lambda}_n(\mathsf{G})\right| < 1$ then $\widehat{\lambda}\left(\mathsf{G}^2\right) = \widehat{\lambda}_2\left(\mathsf{G}^2\right) = \left(\widehat{\lambda}_2(\mathsf{G})\right)^2 = \left(\widehat{\lambda}(\mathsf{G})\right)^2$.

If $\widehat{\lambda}_2(\mathsf{G}) < \left|\widehat{\lambda}_n(\mathsf{G})\right|$ then $\widehat{\lambda}\left(\mathsf{G}^2\right) = \widehat{\lambda}_2\left(\mathsf{G}^2\right) = \left(\widehat{\lambda}_n(\mathsf{G})\right)^2 = \left(\widehat{\lambda}(\mathsf{G})\right)^2$..

Thus, in either case $\widehat{\lambda}\left(\mathsf{G}^2\right) = \left(\widehat{\lambda}(\mathsf{G})\right)^2$. Now, By Lemma 31.4.1 $\gamma_2(\mathsf{G}) = \frac{1}{1-\widehat{\lambda}(\mathsf{G})}$, which implies that $\widehat{\lambda}(\mathsf{G}) = 1 - 1/\gamma_2(\mathsf{G})$, and thus

$$\gamma_2\left(\mathsf{G}^2\right) = \frac{1}{1-\widehat{\lambda}(\mathsf{G}^2)} = \frac{1}{1-\left(\widehat{\lambda}(\mathsf{G})\right)^2} = \frac{1}{1-\left(1-\frac{1}{\gamma_2(\mathsf{G})}\right)^2} = \frac{\gamma_2(\mathsf{G})}{2-\frac{1}{\gamma_2(\mathsf{G})}} = \frac{(\gamma_2(\mathsf{G}))^2}{2\gamma_2(\mathsf{G})-1}.$$

■

31.1.4. The construction

So, let build an expander using Theorem 31.4.3, with parameters $r = 7$ $q = 2^4 = 32$. Let $d = q^2 = 256$. The resulting graph H has $N = q^{r+1} = d^4$ vertices, and it is $d = q^2$ regular. Furthermore, $\widehat{\lambda}_i \leq r/q = 7/32$, for all $i \geq 2$. As such, we have

$$\gamma(\mathsf{H}) = \gamma_2(\mathsf{H}) = \frac{1}{1-7/32} = \frac{32}{25}.$$

Let G_0 be any graph that its square is the complete graph over $n_0 = N+1$ vertices. Observe that G_0^2 is d^4-regular. Set $\mathsf{G}_i = \left(\mathsf{G}_{i-1}^2 \circledcirc \mathsf{H}\right)$, Clearly, the graph G_i has

$$n_i = n_{i-1}N$$

vertices. The graph $\mathsf{G}_{i-1}^2 \circledcirc \mathsf{H}$ is d^2 regular. As far as the bi-tension, let $\alpha_i = \gamma_2(\mathsf{G}_i)$. We have that

$$\alpha_i = \frac{\alpha_{i-1}^2}{2\alpha_{i-1}-1}(\gamma_2(\mathsf{H}))^2 = \frac{\alpha_{i-1}^2}{2\alpha_{i-1}-1}\left(\frac{32}{25}\right)^2 \leq 1.64\frac{\alpha_{i-1}^2}{2\alpha_{i-1}-1}.$$

It is now easy to verify, that α_i can not be bigger than 5.

Theorem 31.1.5. *For any $i \geq 0$, one can compute deterministically a graph G_i with $n_i = (d^4+1)d^{4i}$ vertices, which is d^2 regular, where $d = 256$. The graph G_i is a $(1/10)$-expander.*

Proof: The construction is described above. As for the expansion, since the bi-tension bounds the tension of a graph, we have that $\gamma(\mathsf{G}_i) \leq \gamma_2(\mathsf{G}_i) \leq 5$. Now, by Lemma 31.4.2, we have that G_i is a δ-expander, where $\delta \geq 1/(2\gamma(\mathsf{G}_i)) \geq 1/10$. ■

31.2. Bibliographical notes

A good survey on expanders is the monograph by Hoory *et al.* [HLW06]. The small expander construction is from the paper by Alon *et al.* [ASS08] (but its originally from the work by Along and Roichman [AR94]). The work by Alon *et al.* [ASS08] contains a construction of an expander that is constant degree, which is of similar complexity to the one we presented here. Instead, we used the zig-zag expander construction from the influential work of Reingold *et al.* [RVW02]. Our analysis however, is from an upcoming paper by Mendel and Naor [MN08]. This analysis is arguably reasonably simple (as simplicity is in the eye of the beholder, we will avoid claim that its the simplest), and (even better) provide a good intuition and a systematic approach to analyzing the expansion.

We took a creative freedom in naming notations, and the name tension and bi-tension are the author's own invention.

31.3. Exercises

Exercise 31.3.1 (Expanders made easy.). By considering a random bipartite three-regular graph on $2n$ vertices obtained by picking three random permutations between the two sides of the bipartite graph, prove that there is a $c > 0$ such that for every n there exits a $(2n, 3, c)$-expander. (What is the value of c in your construction?)

Exercise 31.3.2 (Is your consistency in vain?). In the construction, we assumed that the graphs we are dealing with when building expanders have consistent labeling. This can be enforced by working with bipartite graphs, which implies modifying the construction slightly.

(A) Prove that a d-regular bipartite graph always has a consistent labeling (hint: consider matchings in this graph).

(B) Prove that if G is bipartite so is the graph G^3 (the cubed graph).

(C) Let G be a (n, D)-graph and let H be a (D, d)-graph. Prove that if G is bipartite then $\mathsf{GG}ⓩ\mathsf{H}$ is bipartite.

(D) Describe in detail a construction of an expander that is: (i) bipartite, and (ii) has consistent labeling at every stage of the construction (prove this property if necessary). For the ith graph in your series, what is its vertex degree, how many vertices it has, and what is the quality of expansion it provides?

Exercise 31.3.3 (Tension and bi-tension.). [**30 points**]

Disprove (i.e., give a counter example) that there exists a universal constant c, such that for any connected graph G, we have that $\gamma(\mathsf{G}) \leq \gamma_2(\mathsf{G}) \leq c\gamma(\mathsf{G})$.

Acknowledgements

Much of the presentation was followed suggestions by Manor Mendel. He also contributed some of the figures.

31.4. From previous lectures

Lemma 31.4.1. *Let $\mathsf{G} = (V, E)$ be a connected d-regular graph with n vertices. Then $\gamma_2(\mathsf{G}) = \dfrac{1}{1-\widehat{\lambda}}$, where $\widehat{\lambda} = \widehat{\lambda}(\mathsf{G})$, where $\widehat{\lambda}(\mathsf{G}) = \max\left(\widehat{\lambda}_2, -\widehat{\lambda}_n\right)$, where $\widehat{\lambda}_i$ is the ith largest eigenvalue of the random walk matrix associated with G.*

Lemma 31.4.2. *Let* $\mathsf{G} = (V, E)$ *be a given connected* d*-regular graph with* n *vertices. Then,* G *is a* δ*-expander, where* $\delta \geq \dfrac{1}{2\gamma(\mathsf{G})}$ *and* $\gamma(\mathsf{G})$ *is the tension of* G*.*

Theorem 31.4.3. *For any* $t > 0, r > 0$ *and* $q = 2^t$*, where* $r < q$*, we have that* $\mathsf{LD}(q, r)$ *is a graph with* q^{r+1} *vertices. Furthermore,* $\lambda_1(\mathsf{LD}(q, r)) = q^2$*, and* $\lambda_i(\mathsf{LD}(q, r)) \leq rq$*, for* $i = 2, \ldots, n$*.*

In particular, if $r \leq q/2$*, then* $\mathsf{LD}(q, r)$ *is a* $\left[q^{r+1}, q^2, \frac{1}{4}\right]$*-expander.*

Bibliography

[AR94] N. Alon and Y. Roichman. Random cayley graphs and expanders. *Random Struct. Algorithms*, 5(2):271–285, 1994.

[ASS08] N. Alon, O. Schwartz, and A. Shapira. An elementary construction of constant-degree expanders. *Combin. Probab. Comput.*, 17(3):319–327, 2008.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin Amer. Math. Soc.*, 43:439–561, 2006.

[MN08] M. Mendel and A. Naor. Towards a calculus for non-linear spectral gaps. manuscript, 2008.

[RVW02] O. Reingold, S. Vadhan, and A. Wigderson. Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors. *Annals Math.*, 155(1):157–187, 2002.

Chapter 32

Miscellaneous Prerequisite

By Sariel Har-Peled, December 30, 2015[①]

> Be that as it may, it is to night school that I owe what education I possess; I am the first to own that it doesn't amount to much, though there is something rather grandiose about the gaps in it.
>
> – The tin drum, Gunter Grass.

The purpose of this chapter is to remind the reader (and the author) about some basic definitions and results in mathematics used in the text. The reader should refer to standard texts for further details.

32.1. Geometry and linear algebra

A set X in $\mathbb{R}^d$ is *closed*, if any sequence of converging points of X converges to a point that is inside X. A set $X \subseteq \mathbb{R}^d$ is *compact* if it is closed and bounded; namely; there exists a constant c, such that for all $\mathsf{p} \in X$, $\|\mathsf{p}\| \leq c$.

Definition 32.1.1 (Convex hull). The *convex hull* of a set $\mathsf{R} \subseteq \mathbb{R}^d$ is the set of all convex combinations of points of R; that is,

$$\mathcal{CH}(\mathsf{R}) = \left\{ \sum_{i=0}^{m} \alpha_i \mathsf{r}_i \;\middle|\; \forall i \;\; \mathsf{r}_i \in \mathsf{R}, \alpha_i \geq 0, \text{ and } \sum_{j=1}^{m} \alpha_i = 1 \right\}.$$

In the following, we cover some material from linear algebra. Proofs of these facts can be found in any text on linear algebra, for example [Leo98].

For a matrix M, let M^T denote the transposed matrix. We remind the reader that for two matrices M and B, we have $(\mathsf{MB})^T = \mathsf{B}^T\mathsf{M}^T$. Furthermore, for any three matrices M, B, and C, we have $(\mathsf{MB})\mathsf{C} = \mathsf{M}(\mathsf{BC})$.

A matrix $\mathsf{M} \in \mathbb{R}^{n \times n}$ is ***symmetric*** if $\mathsf{M}^T = \mathsf{M}$. All the eigenvalues of a symmetric matrix are real numbers. A symmetric matrix M is ***positive definite*** if $x^T\mathsf{M}x > 0$, for all $x \in \mathbb{R}^n$. Among other things this implies that M is non-singular. If M is symmetric, then it is positive definite if and only if all its eigenvalues are positive numbers.

In particular, if M is symmetric positive definite, then $\det(\mathsf{M}) > 0$. Since all the eigenvalues of a positive definite matrix are positive real numbers, the following holds, as can be easily verified.

Claim 32.1.2. *A symmetric matrix* M *is positive definite if and only if there exists a matrix* B *such that* $\mathsf{M} = \mathsf{B}^T\mathsf{B}$ *and* B *is not singular.*

For two vectors $u, v \in \mathbb{R}^n$, let $\langle u, v\rangle = u^T v$ denote their dot product.

Lemma 32.1.3. *Given a simplex* $\triangle$ *in* $\mathbb{R}^d$ *with vertices* $v_1, \ldots, v_d, v_{d+1}$ *(or equivalently* $\triangle = \mathcal{CH}(v_1, \ldots, v_{d+1})$*), the* **volume** *of this simplex is the absolute value of* $(1/d!)|C|$*, where* C *is the value of the determinant* $C = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ v_1 & v_2 & \ldots & v_{d+1} \end{vmatrix}$*. In particular, for a triangle with vertices at* (x, y)*,* (x', y')*, and* (x'', y'') *its area is the absolute value of* $\frac{1}{2}\begin{vmatrix} 1 & x & y \\ 1 & x' & y' \\ 1 & x'' & y'' \end{vmatrix}$.

32.1.1. Linear and affine subspaces

Definition 32.1.4. The *linear subspace* spanned by a set of vectors $\mathcal{V} \subseteq \mathbb{R}^d$ is the set $\text{linear}(\mathcal{V}) = \left\{ \sum_i \alpha_i \overrightarrow{\mathsf{v}}_i \;\middle|\; \alpha_i \in \mathbb{R}, \overrightarrow{\mathsf{v}}_i \in \mathcal{V} \right.$

An *affine combination* of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a linear combination $\sum_{i=1}^n \alpha_i \cdot \mathbf{v}_i = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$ in which the sum of the coefficients is 1; thus, $\sum_{i=1}^n \alpha_i = 1$. The maximum dimension of the affine subspace in such a case is $(n-1)$-dimensions.

Definition 32.1.5. The *affine subspace* spanned by a set $\mathcal{V} \subseteq \mathbb{R}^d$ is

$$\text{affine}(\mathcal{V}) = \left\{ \sum_i \alpha_i \overrightarrow{\mathsf{v}}_i \;\middle|\; \alpha_i \in \mathbb{R}, \overrightarrow{\mathsf{v}}_i \in \mathcal{V}, \text{ and } \sum_i \alpha_i = 1 \right\}.$$

For any vector $\overrightarrow{\mathsf{v}} \in \mathcal{V}$, we have that $\text{affine}(\mathcal{V}) = \overrightarrow{\mathsf{v}} + \text{linear}\left(\mathcal{V} - \overrightarrow{\mathsf{v}}\right)$, where $\mathcal{V} - \overrightarrow{\mathsf{v}} = \left\{ \overrightarrow{\mathsf{v}}' - \overrightarrow{\mathsf{v}} \;\middle|\; \overrightarrow{\mathsf{v}}' \in \mathcal{V} \right\}$.

32.1.2. Computational geometry

The following are standard results in computational geometry; see [dBCKO08] for more details.

Lemma 32.1.6. *The convex hull of* n *points in the plane can be computed in* $O(n \log n)$ *time.*

Lemma 32.1.7. *The lower and upper envelopes of* n *lines in the plane can be computed in* $O(n \log n)$ *time.*

Proof: Use duality and the algorithm of Lemma 32.1.6. ■

32.2. Calculus

Lemma 32.2.1. *For* $x \in (-1, 1)$*, we have* $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i}$.

Lemma 32.2.2. *The following hold:*
(A) For all $x \in \mathbb{R}$*,* $1 + x \leq \exp(x)$.
(B) For $x \geq 0$*,* $1 - x \leq \exp(-x)$.

(C) For $0 \leq x \leq 1$, $\exp(x) \leq 1 + 2x$.
(D) For $x \in [0, 1/2]$, $\exp(-2x) \leq 1 - x$.

Proof: (A) Let $f(x) = 1 + x$ and $g(x) = \exp(x)$. Observe that $f(0) = g(0) = 1$. Now, for $x \geq 0$, we have that $f'(x) = 1$ and $g'(x) = \exp(x) \geq 1$. As such $f(x) \leq g(x)$ for $x \geq 0$. Similarly, for $x < 0$, we have $g'(x) = \exp(x) < 1$, which implies that $f(x) \leq g(x)$.

(B) This is immediate from (A).

(C) Observe that $\exp(1) \leq 1 + 2 \cdot 1$ and $\exp(0) = 1 + 2 \cdot 0$. By the convexity of $1 + 2x$, it follows that $\exp(x) \leq 1 + 2x$ for all $x \in [0, 1]$.

(D) Observe that (i) $\exp(-2(1/2)) = 1/e \leq 1/2 = 1 - (1/2)$, (ii) $\exp(-2 \cdot 0) = 1 \leq 1 - 0$, (iii) $\exp(-2x)' = -2\exp(-2x)$, and (iv) $\exp(-2x)'' = 4\exp(-2x) \geq 0$ for all x. As such, $\exp(-2x)$ is a convex function and the claim follows. ■

Lemma 32.2.3. *For* $1 > \varepsilon > 0$ *and* $y \geq 1$, *we have that* $\dfrac{\ln y}{\varepsilon} \leq \log_{1+\varepsilon} y \leq 2\dfrac{\ln y}{\varepsilon}$.

Proof: By Lemma 32.2.2, $1 + x \leq \exp(x) \leq 1 + 2x$ for $x \in [0, 1]$. This implies that $\ln(1 + x) \leq x \leq \ln(1 + 2x)$. As such, $\log_{1+\varepsilon} y = \dfrac{\ln y}{\ln(1+\varepsilon)} = \dfrac{\ln y}{\ln(1+2(\varepsilon/2))} \leq \dfrac{\ln y}{\varepsilon/2}$. The other inequality follows in a similar fashion. ■

Bibliography

[dBCKO08] M. de Berg, O. Cheong, M. van Kreveld, and M. H. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Santa Clara, CA, USA, 3rd edition, 2008.

[Leo98] S. J. Leon. *Linear Algebra with Applications*. Prentice Hall, 5th edition, 1998.