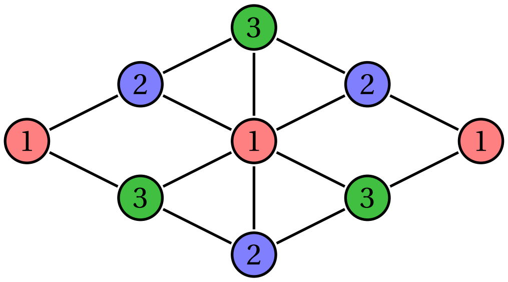
Probabilistic Combinatorics

Winter 2019

Thomas Rothvoss



UNIVERSITY *of*
WASHINGTON

Last changes: March 15, 2019

Contents

Chapter 1

Introduction

The *probabilistic method* was spearheaded by Paul Erdős to an extend that it is sometimes called the "Erdős method". By now it is one of the standard techniques in combinatorics and other areas of discrete mathematics as well as theoretical computer science. Simply phrased, the idea is to prove the statement of a theorem or prove the existence of an object using probability. The goal of these lecture notes is to give an introduction into the probabilistic method and the involved techniques, where we have a preference for elegant solutions rather than intricate calculations. In particular we will rarely care about the exact constants in order to keep the exposition as clean as possible. Parts of this text will follow the excellent textbook of Alon and Spencer [AS16], but we will also see applications found elsewhere.

1.1 Ramsey Graphs

While there have been earlier applications, probably one result by Erdős from 1947 popularized the probabilistic method. The question back in 1947 was whether there are undirected graphs $G = (V, E)$ that have neither a large clique, nor a large anti-clique. Here a *clique* is a set $S \subseteq V$ so that the induced subgraph $G[S]$ is complete, while S is an *anti-clique* if $G[S]$ contains no edges. Recall that the induced subgraph $G[S] = (S, \{\{u, v\} \in E \mid u, v \in S\})$ is the graph on nodes S that "inherits" exactly the edges contained in S. Also recall that $N(u) := \{v \in V \mid \{u, v\} \in E\}$ is the *neighborhood* of a node $u \in V$. And the "first theorem" of *Ramsey Theory* shows that there has to be at least a clique or anti-clique of logarithmic size inany graph.

Lemma 1.1. *Any n-node graph contains either a clique or anti-clique of size $\frac{1}{2}\log_2(n)$.*

Proof. We prove the following claim by induction over $k + \ell$:

Claim. *Any graph on $n \geq 2^{k+\ell}$ nodes contains either a k-clique or an ℓ-anti-clique.*

Proof of claim. Fix a node v. If $|N(v)| \geq n/2$ then $G[N(v)]$ has at least $2^{(k-1)+\ell}$ nodes so it either contains a size-ℓ anti-clique or it contains a $(k-1)$-clique which we can extend to a k-clique by adding v. The other case is that $|V \setminus (\{v\} \cup N(v))| \geq n/2$ in which we can similarly argue that there is either a k clique or an $(\ell - 1)$ anti-clique not incident to v. □

In particular a 2^{2k}-node graph must contain either a k-clique or k-anticlique, which then gives the claim. □

Usually one defines $R(k,\ell)$ as the minimum integer so that every graph with at least $R(k,\ell)$ nodes contains either a k-clique or a ℓ-anti clique. For example $R(3,3) = 6$, which is often quoted as the fact that at every party with at least 6 people, there are either 3 people who all know each other or 3 people who all do not know each other.

Somewhat surprisingly there are indeed graphs without a $\omega(\log n)$ clique or anti-clique.

Theorem 1.2. *For any n there is a graph without a $2\log_2(n) + O(1)$ clique or anti-clique.*

Proof. We pick a graph $G = ([n], E)$ at random where every possible edge $\{u, v\}$ is inserted into the graph independently with probability $1/2$. Fix $k := 2\log_2(n) + C$ for a big constant C. By symmetry it suffices to show that the probability that a k-clique exists is less than $1/2$. And we can bound that probability by the expected number of k-cliques:

$$\begin{aligned}\Pr[\exists k\text{-clique in } G] &\leq \sum_{S \subseteq V : |S| = k} \Pr[G[S] \text{ is complete}] \leq \binom{n}{k} \cdot 2^{-\binom{k}{2}} \\ &\leq n^k 2^{-k^2/2+k} = 2^{\log_2(n)\cdot k - k^2/2 + k} < \frac{1}{2}\end{aligned}$$

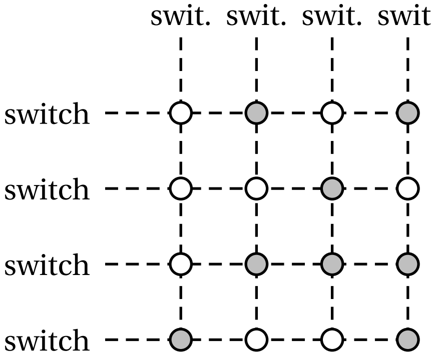
for our choice of k. □

The proof was essentially trivial. But it is surprisingly hard to come up with an inherently different construction of a graph without a large clique or anti-clique. In fact no non-probabilistic construction of a graph without $O(\log n)$-clique or anti-clique is known! Apparently this is a deeper problem of constructing a random-like object without the use of randomness. The best explicit construction by Barak, Rao, Shaltiel and Wigderson [BRSW06] provides an n-node graph without a clique or anti-clique of size $2^{2^{\log^{1-\varepsilon}(\log(n))}}$.

As a second remark, the line of arguments where we set up a random experiment and then reason using the expectation (here the expected number of cliques/anticliques) is also called the *First Moment Method*. Often these types of proofs are the easiest probabilistic proofs.

1.2 Balancing lights

We want to study another application where it will be advantageous to construct the desired object by a mix ob randomization and deterministic choice. Suppose we have an $n \times n$ array of lights in some initial state where each light is either on or off. We have $2n$ switches, one for each horizontal line and one for each vertical line that switches the whole line. The question is: given any adversarial initial state of the lights, how many lights can be guaranteed to be turned on? In particular how much more than just half the lights can be switched on?



Lemma 1.3. *One can always turn switches so that* $\frac{n^2}{2} + \Theta(n^{3/2})$ *many lights are on.*

Proof. We can formalize the claim as follows: Given a matrix $\boldsymbol{A} \in \{-1,1\}^{n\times n}$, show that there are $\boldsymbol{x}, \boldsymbol{y} \in \{-1,1\}^n$ so that $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{y} \geq \Omega(n^{3/2})$. We forget about the signs $\boldsymbol{x}$ for the moment and pick only $\boldsymbol{y} \in \{-1,1\}^n$ uniformly at random. Then the inner product $\langle \boldsymbol{A}_i, \boldsymbol{y}\rangle$ is the sum of n uniform random elements from $\{-1,1\}$. In particular $\mathbb{E}[|\langle \boldsymbol{A}_i, \boldsymbol{y}\rangle|] = \Theta(\sqrt{n})$ (we will fill out details in the exercises). Then $\sum_{i=1}^n |\langle \boldsymbol{A}_i, \boldsymbol{y}\rangle| = \Theta(n^{3/2})$. Now we pick $x_i := \text{sign}(\langle \boldsymbol{A}_i, \boldsymbol{y}\rangle)$ and

$$\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{y} = \sum_{i=1}^n x_i \cdot \langle \boldsymbol{A}_i, \boldsymbol{y}\rangle = \Theta(n^{3/2})$$

as desired. □

1.3 On the number of disjoint pairs

Suppose that we have a set family $\mathcal{F} \subseteq 2^{\{1,\ldots,n\}}$. Let

$$d(\mathcal{F}) := \{\{F, F'\} \mid F, F' \in \mathcal{F} \text{ with } F \cap F' = \emptyset\}$$

be the number of *disjoint pairs* of sets in $\mathcal{F}$. The question that Daykin and ErdḦ os where wondering is, how large can $|\mathcal{F}|$ be so that still a good fraction of pairs is disjoint. For example one could let $\mathcal{F}$ be all the subsets of $\{1, \ldots, \frac{n}{2}\}$ plus all the subsets of $\{\frac{n}{2}+1, \ldots, n\}$. Then at least half the pairs is disjoint and $|\mathcal{F}| = \Theta(1) \cdot 2^{n/2}$. But what happens beyond the threshold of $2^{n/2}$ many sets? Daykin and Erdős conjectured that as soon as $|\mathcal{F}| \geq 2^{(\frac{1}{2}+\delta)n}$ for some constant $\delta > 0$ one would have $d(\mathcal{F}) \leq o(|\mathcal{F}|^2)$. And indeed this is true, as was proven by Alon and Frankl.

Theorem 1.4 (Alon, Frank 1985)**.** *Let $\mathcal{F} \subseteq 2^{[n]}$ be a family of $|\mathcal{F}| = 2^{(\frac{1}{2}+\frac{1}{t})n}$ sets where $t \in \mathbb{N}$. Then $d(\mathcal{F}) \leq |\mathcal{F}|^{2-\Theta(1/t^2)}$.*

Proof. We sample independently $t+1$ members $A_1, \ldots, A_{t+1} \in \mathcal{F}$ uniformly from the set family. Then we estimate that

$$\begin{aligned}
\Pr\left[|A_1 \cup \ldots \cup A_{t+1}| \leq \frac{n}{2}\right] &\overset{\text{union bound}}{\leq} \sum_{S \subseteq [n]:|S|=\frac{n}{2}} \Pr[A_1, \ldots, A_{t+1} \subseteq S] \qquad (*) \\
&\overset{\text{independence}}{=} \sum_{S \subseteq [n]:|S|=\frac{n}{2}} \Big(\underbrace{\Pr[A_1 \subseteq S]}_{\leq 2^{n/2}/|\mathcal{F}|}\Big)^{t+1} \leq 2^n \cdot \left(\frac{2^{n/2}}{2^{(\frac{1}{2}+\frac{1}{t})n}}\right)^{t+1} \\
&= 2^{n \cdot (1-\frac{t+1}{t})} = 2^{-n/t}
\end{aligned}$$

Note that if we only had $2^{n/2}$ many sets and they are subsets of either the first $n/2$ or the second $n/2$ elements, then this probability would have been only $2^{-\Theta(t)}$. Hence we are indeed using the assumption that we have a lot more sets. From this estimate we can already quickly see that in $2 \cdot (t+1)$ samples we would very likely see at least some collisions.

Now we get a more precise analysis. Still, we consider the random experiment where sets $A_1, \ldots, A_{t+1} \in \mathcal{F}$ are drawn at random. Moreover, let

$$Y := |\{B \in \mathcal{F} \mid B \cap (A_1 \cup \ldots \cup A_{t+1}) = \emptyset\}|$$

be the random variable that gives the number of sets disjoint to all of the $t+1$ samples. The correct intuition will be that $\mathbb{E}[Y]$ is going to be a good proxy for $d(\mathcal{F})^{t+1}$.

First of all, the bound in $(*)$ implies an upper bound on the expected value of Y:

$$\begin{aligned} \mathbb{E}[Y] &\leq \underbrace{\Pr\left[|A_1 \cup \ldots \cup A_{t+1}| \leq \frac{n}{2}\right]}_{\leq 2^{-n/t}} \cdot \underbrace{\mathbb{E}\left[Y \mid |A_1 \cup \ldots \cup A_{t+1}| \leq \frac{n}{2}\right]}_{\leq |\mathcal{F}|} + \underbrace{\mathbb{E}\left[Y \mid |A_1 \cup \ldots \cup A_{t+1}| > \frac{n}{2}\right]}_{\leq 2^{n/2}} \\ &\leq \underbrace{2^{-n/t}|\mathcal{F}|}_{=2^{n/2}} + 2^{n/2} = 2 \cdot 2^{-n/t} \cdot |\mathcal{F}| \end{aligned}$$

It will be convenient to also define $\text{disj}(B) := |\{A \in \mathcal{F} \mid A \cap B = \emptyset\}|$ as the number of sets in the family disjoint to B. We can upper bound

$$\mathbb{E}[Y] = \sum_{B \in \mathcal{F}} \underbrace{\left(\frac{\text{disj}(B)}{|\mathcal{F}|}\right)^{t+1}}_{\Pr[B \text{ disj. to } A_1,\ldots,A_{t+1}]} \geq |\mathcal{F}| \cdot \left(\frac{1}{|\mathcal{F}|^2} \overbrace{\sum_{B \in \mathcal{F}} \text{disj}(B)}^{=2d(\mathcal{F})}\right)^{t+1} = \frac{1}{|\mathcal{F}|^{2t+1}} \cdot (2d(\mathcal{F}))^{t+1}$$

using Jensen's inequality and the convexity of $x \mapsto x^{t+1}$. Combining this we have shown that

$$\frac{2^{t+1}}{|\mathcal{F}|^{2t+1}} \cdot d(\mathcal{F})^{t+1} \leq \mathbb{E}[Y] \leq 2 \cdot 2^{-n/t} |\mathcal{F}|$$

which can be rearranged to

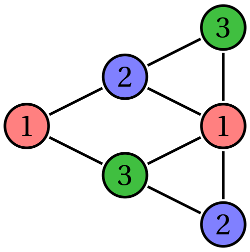
$$d(\mathcal{F}) \leq O(1) \cdot 2^{-n/(t \cdot (t+1))} \cdot |\mathcal{F}|^2$$

which gives the claimed bound. □

This result falls into a large category of applications of the probabilistic method, where it's not about the existence of some object that is the outcome of a random experiment, but about some inequality of deterministic quantities. Often inequalities can be proven by doing a random experiment that relate the involved quantities.

1.4 Graphs with high chromatic number and high girth

For an undirected graph $G = (V, E)$, a *coloring* with k colors is a map $c : V \to \{1, \ldots, k\}$ so that $c(i) \neq c(j)$ for all $\{i, j\} \in E$. We denote $\chi(G)$ as the minimum number o colors that are needed to color G. It's easy to see that if G contains a K_k as a subgraph, then $\chi(G) \geq k$. For example below, we see a 3-coloring of a graph. Clearly there will not be a 2-coloring as the graph contains a 3-clique.



In general it is **NP**-hard to determine $\chi(G)$, but one might wonder what other obstructions for good colorings there might be. In particular, one might believe that a graph has a coloring with few colors as long as there are no *short cycles*. However, it turns out that this is false. But it is quite non-trivial to construct example graphs showing this. Let $\text{girth}(G)$ denote the smallest number of edges in any cycle in G and let $\alpha(G)$ denote the size of the largest independent set.

Theorem 1.5 (Erdős 1959)**.** *For any $k \in \mathbb{N}$, there is family of graphs that have $\text{girth}(G) > k$ and $\chi(G) > \Omega(n^{1/(4k)})$.*

Proof. Let n be large enough, compared to k. Set $d := n^{1/(2k)}$. We pick a *random graph* $G = (V,E)$ on n vertices by inserting each edge independently, say with probability $\frac{d}{n}$. In other words, we a random graph that has about d. In order to prove that $\chi(G) \geq \sqrt{d}$, we show that there is not even an independent set of size $\frac{n}{\sqrt{d}}$ in such a graph. And in fact we can even do so by counting the expected number of subsets S with $|S| = \frac{n}{\sqrt{d}}$ that do not include an edge:

$$\begin{aligned}
\Pr\left[\alpha(G) > \frac{n}{\sqrt{d}}\right] &\leq \binom{n}{n/\sqrt{d}} \cdot \left(1 - \frac{d}{n}\right)^{\binom{n/\sqrt{d}}{2}} \\
&\leq n^{n/\sqrt{d}} \cdot \exp\left(-\frac{d}{n} \cdot \frac{1}{4}(n/\sqrt{d})^2\right) \\
&\leq \exp\left(\ln(n) \cdot \frac{n}{\sqrt{d}} - \frac{1}{4} \cdot n\right) \leq o(1)
\end{aligned}$$

That means $\chi(G) \geq \sqrt{d} = n^{1/(4k)}$ with probability $1 - o(1)$. Now we would love to show that G will not contain short cycles. But there is a problem here. For example there are about $\Theta(n^k)$ candidate cycles of length k and each particular one exists with probability $(\frac{d}{n})^k$. Then the expected number of length-k cycles is of order $\Theta(d^k)$.

That means G will contain a ton of length-k cycles. But there is a way to fix this. Let X be the number of cycles of length at most k. If we can show that $X \leq o(n)$, then we can delete one edge from every cycle and end up with a graph with $\text{girth}(G) > k$. Deleting edges might decrease the chromatic number, but there will be, say $n/2$ many nodes U that did not have any incident edge deleted. That subgraph $G[U]$ will satisfy the claim. Back to our estimates on the number of cycles:

$$\mathbb{E}[X] \leq \sum_{\ell=3}^{k} n^{\ell} \cdot \left(\frac{d}{n}\right)^{\ell} \leq k \cdot d^{k} = k \cdot \sqrt{n} \leq o(n)$$

In particular $\Pr[X \geq \frac{n}{4}] \leq o(1)$ if $n \gg k$. □

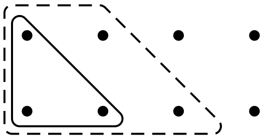
The line of arguments that we have seen here is also called the *method of alterations* where in general one sets up a random experiment that gives an object that does not quite satisfy the desired requirements. Then one has a 2nd round in which the object is modified.

1.5 The Rödl Nibble

How many matchings does it take to cover all the vertices in a complete n-node graph? Trivially $\lceil n/2 \rceil$ many. How many triangles does it take to cover all the edges in a complete graph? This already requires some thoughts. The number must be at least $\frac{1}{3} \cdot \binom{n}{2} \approx \frac{1}{6} n^2$, but it is not fully trivial whether this bound can be achieved.

And in fact, we want to consider this question in more generality. Let $\mathcal{H}_{n,r} = ([n], \mathcal{E})$ be the complete r-uniform hypergraph, meaning that the hyperedges are $\mathcal{E} = \binom{[n]}{r}$. We define

$$M(n,k,\ell) := \#\text{ min edges of } \mathcal{H}_{n,k} \text{ needed to cover all edges in } \mathcal{H}_{n,\ell}$$



Example: size $\ell = 3$ edge covered by a size $k = 5$ edge

Note that $\mathcal{H}_{n,\ell}$ has $\binom{n}{\ell}$ many edges and every edge of $\mathcal{H}_{n,k}$ can cover at most $\binom{k}{\ell}$ of these, so

$$M(n,k,\ell) \geq \frac{\binom{n}{\ell}}{\binom{k}{\ell}}$$

Erdős and Hanani conjectured in 1963 that this bound can be achieved asymptotically. It took two decades until this was proven by Rödl:

Theorem 1.6 (Rödl 1985)**.** *For all* $2 \leq \ell < k$ *one has*

$$\frac{M(n,k,\ell)}{\binom{n}{\ell}/\binom{k}{\ell}} \xrightarrow{n\to\infty} 1.$$

By now there are also algebraic constructions known, but Rödl's proof technique is robust and quite useful in other settings. We prove the more general statement that guarantees the existence of near perfect coverings in a hypergraph. Later we will argue how it implies Rödl's Theorem. For a hypergraph $H = (V, \mathcal{E})$ we denote $d_H(i)$ as the *degree* of a vertex $i \in V$. Moreover, for a pair of nodes $i, j \in V$ we let $d_H(i,j)$ be the number of edges containing both nodes. We will omit the index H if it is clear from context. A *cover* is a set of edges $\mathcal{F} \subseteq \mathcal{E}$ with $\bigcup_{e\in\mathcal{F}} e = [n]$. Rödl proved that under some assumptions a hypergraph contains an almost perfect cover (here the order of the quantifiers should be understood as $\forall r \in \mathbb{Z}_{\geq 2}, K \geq 1, \delta > 0 \; \exists D_0, \varepsilon$):

Theorem 1.7 (Existence of almost perfect covers in hypergraphs)**.** *Fix arbitrary constants* $r \in \mathbb{Z}_{\geq 2}$ *and* $K \geq 1$*. Let* $H = ([n], \mathcal{E})$ *be an* r*-uniform hypergraph with* $n \geq D \geq D_0$ *satisfying*

1. *Every vertex* $i \in [n]$ *but at most* εn *many of them one has* $d(i) = (1 \pm \varepsilon) \cdot D$*.*
2. *For all* $i \in V$ *one has at least* $1 \leq d(i) \leq K \cdot D$*.*
3. *For any distinct nodes* $i, j \in [n]$ *one has* $d(i,j) \leq \varepsilon D$*.*

Then there is a cover of $(1+\delta) \cdot \frac{n}{r}$ *edges where* $(\varepsilon \to 0$ *and* $D_0 \to \infty) \Rightarrow \delta \to 0$*.*

We should first convince ourselfs that the naive proof strategy must fail. Suppose we sample each hyper edge of with probability $\frac{1}{D}$ so that each node is covered in expectation once, but the probability of being covered is only at least $1 - \frac{1}{e}$. The solution is that we only take small "bites" or "nibbles" of hyperedges in the sense that we only sample a very small fraction of edges. The lemma characterizing a successful "bite" is as follows (again the order of the quantifiers should be understood as $\forall r \in \mathbb{Z}_{\geq 2}, K \geq 1, \delta > 0, \alpha > 0 \; \exists D_0, \varepsilon$):

Lemma 1.8. *Fix arbitrary constants* $r \in \mathbb{Z}_{\geq 2}$*,* $K \geq 1$ *and* $\alpha > 0$*. Let* $H = ([n], \mathcal{E})$ *be an* r*-uniform hypergraph with* $n \geq D \geq D_0$ *satisfying the following:*

i) *All vertices* $i \in V$ *except at most* εn *of them have* $d(i) = (1 \pm \varepsilon)D$*.*

ii) *For all* $i \in [n]$ *one has* $d(i) \leq K \cdot D$*.*

iii) *Any two vertices* i, j *satisfy* $d(i,j) \leq \varepsilon D$*.*

Then there is a set $\mathcal{E}' \subseteq \mathcal{E}$ *of hyperedges so that*

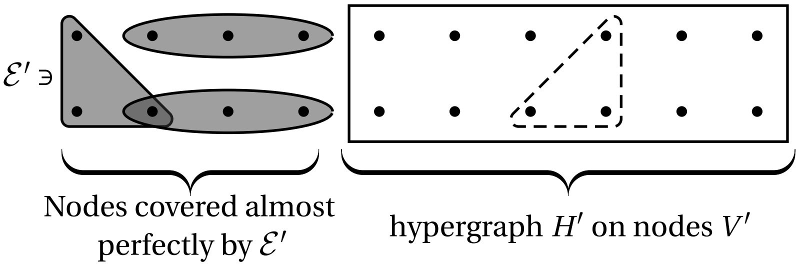
(A) $|\mathcal{E}'| = \alpha \cdot \frac{n}{r} \cdot (1 \pm \delta)$.

(B) The set $V' := V \setminus \bigcup_{e \in \mathcal{E}'} e$ *of uncovered nodes has size* $|V'| = n \cdot e^{-\alpha} \cdot (1 \pm \delta)$.

(C) For all uncovered vertices $i \in V'$ *except* $\delta |V'|$ *of them, the degree* $d'(i)$ *in the induced hypergraph* $H' = (V', \{e \in \mathcal{E} : e \subseteq V'\})$ *is* $d'(i) = D \cdot e^{-\alpha(r-1)} \cdot (1 \pm \delta)$.

Again one has $(\varepsilon \to 0$ *and* $D_0 \to \infty) \Rightarrow \delta \to 0$.

It is important that after an application of Lemma 1.8 we can still guarantee the same regularity for the remaining hypergraph H' that we had before so that Lemma 1.8 can be applied again. We can visualize Lemma 1.8 as follows:



$\mathcal{E}' \ni$
Nodes covered almost perfectly by $\mathcal{E}'$
hypergraph H' on nodes V'

Proof of Lemma 1.8. We pick a random subset $\mathcal{E}' \subseteq \mathcal{E}$ that contains every edge independently with probability $\frac{\alpha}{D}$. In order to keep the notation simple we will write $K = O(1)$ and we write $o(1)$ instead of introducing a sequence of various constants depending on ε that all tend to 0. The way to interpret this is that as we send $\varepsilon \to 0$ and $D_0 \to \infty$, also the expression hidden by $o(1)$ will go to 0.

Note that the assumptions on uniformity and degree imply that $|\mathcal{E}| = \frac{nD}{r} \cdot (1 \pm o(1))$ and hence $\mathbb{E}[|\mathcal{E}'|] = (1 \pm o(1)) \cdot \frac{\alpha n}{r}$. Then (A) follows from concentration bounds like Chernov. We call a node i *good* if it satisfies the degree bound $d(i) = (1 \pm \varepsilon) \cdot D$ from i). Let

$$I_i := \begin{cases} 1 & \text{if } i \notin \bigcup_{e \in \mathcal{E}'} e \\ 0 & \text{otherwise} \end{cases}$$

be the indicator variable telling whether i is uncovered. In fact, the probability that i is uncovered is

$$\Pr[I_i = 1] = \Big(1 - \frac{\alpha}{D}\Big)^{d(i)} \overset{\text{if } i \text{ is good}}{=} e^{-\alpha} \cdot (1 \pm o(1)).$$

Since most vertices are good anyway, this implies that $\mathbb{E}[|V'|] = ne^{-\alpha}(1 \pm o(1))$. However, the expectation will not be enough to control the behavior of $|V'|$. We will also bound the variance of the random variable $|V'|$. First, for distinct nodes $i, j \in [n]$ we have

$$\begin{aligned} \mathrm{Cov}[I_i, I_j] &= \mathbb{E}[I_i I_j] - \mathbb{E}[I_i]\,\mathbb{E}[I_j] = \Big(1-\frac{\alpha}{D}\Big)^{d(i)+d(j)-d(i,j)} - \Big(1-\frac{\alpha}{D}\Big)^{d(i)+d(j)} \\ &\leq \Big(1-\frac{\alpha}{D}\Big)^{-d(i,j)} - 1 \overset{d(i,j)\leq o(D)}{\leq} o(1) \end{aligned}$$

In particular we have used that $\mathbb{E}[I_i I_j]$ is the probability that both nodes are uncovered, which requires that none of the $d(i)+d(j)-d(i,j)$ many incident hyperedges is sampled. Then we can bound the variance of the number of uncovered nodes by

$$\mathrm{Var}[|V'|] = \sum_{i=1}^n \underbrace{\mathrm{Var}[I_i]}_{\leq 1} + \sum_{i,j\in[n]:i\neq j} \mathrm{Cov}[I_i, I_j] \leq n + o(n^2) \leq o(n^2)$$

Then Chebychev's Inequality[1] tells us that $|V'| = (1 \pm o(1)) \cdot \mathbb{E}[|V'|]$ with probability at least 0.99 and we have (B).

It remains to prove that the degrees of most nodes in V' are as claimed in (C). Note that the hypergraph H' inherits only the hyperedges completely contained in V'. First note that all but $o(n)$ many nodes $i \in [n]$ satisfy

(I) $d(i) = (1 \pm o(1)) \cdot D$.

(II) All but at most $o(D)$ many edges $e \in \delta_H(i)$ satisfy $|\{f \in \mathcal{E} : i \notin f, f \cap e \neq \emptyset\}| = (1 \pm o(1)) \cdot (r-1) \cdot D$.

Here (I) is one of the assumptions. For (II) note that (ignoring the outliers) each of the $r-1$ vertices in $e \setminus \{i\}$ have degree $D \cdot (1 \pm o(1))$ and there are only $o(D)$ many edges containing i and one or more other nodes in $e \setminus \{i\}$. Consider a node i satisfying (I) and (II). We call an edge $e \in \delta_H(i)$ *good* if it satisfies the condition in (II). For such a good edge, the chance that it stays in H' is

$$\Pr[e \subseteq V' \mid i \in V'] = \Big(1-\frac{\alpha}{D}\Big)^{(1\pm o(1))\cdot(r-1)D} \overset{D \text{ large}}{=} e^{-(1\pm o(1))\cdot\alpha\cdot(r-1)}$$

The degree of i is mostly controlled by the number of good edges as there are $o(D)$ bad ones and so

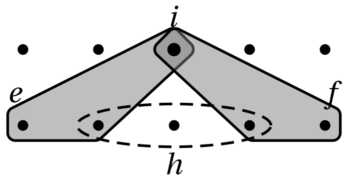
$$\mathbb{E}[d'(i) \mid i \in V'] = D \cdot \exp(-(1 \pm o(1)) \cdot \alpha(r-1))$$

[1] Recall that *Chebychev's Inequality* says that for any random variable X and any $\lambda > 0$ one has $\Pr[|X - \mathbb{E}[X]| \geq \lambda \cdot \sqrt{\mathrm{Var}[X]}] \leq \frac{1}{\lambda^2}$. Also recall that the *variance* is $\mathrm{Var}[X] := \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Also it is useful to remember that if $X = X_1 + \ldots + X_n$ is the sum of (not necessarily independent) random variables, then $\mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i] + \sum_{i\neq j} \mathrm{Cov}[X_i, X_j]$ where $\mathrm{Cov}[X_i, X_j] := \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ is the *covariance* of the pair (X_i, X_j).

Again, we need to show that $d'(i)$ is also close to its expecation with high probability and again we will estimate the variance for that purpose. Let I_e be the indicator random variable for the event $e \subseteq V'$. Then

$$\begin{aligned}\mathrm{Var}[d'(i) \mid i \in V'] \;\le\; & \underbrace{\mathbb{E}[d'(i) \mid i \in V']}_{\le (1+o(1))\cdot D} + \underbrace{\sum_{\substack{e,f\in\delta_H(i)\\ |e\cap f|>1}}}_{\le o(D^2)} \underbrace{\mathrm{Cov}[I_e, I_f \mid i \in V']}_{\le 1} \\ & + \sum_{\substack{e,f\in\delta_H(i)\\ e\cap f=\{i\}}} \underbrace{\mathrm{Cov}[I_e, I_f \mid i \in V']}_{\le o(1) \text{ by } (*)} \le o(D^2)\end{aligned}$$

It remains to argue why $(*)$ holds. Fix two edges $e, f \in \delta_H(i)$ with $e \cap f = \{i\}$. The events $\{e \subseteq V' \mid i \in V'\}$ and $\{f \subseteq V' \mid i \in V'\}$ are not necessarily independent as there can be edges h that overlap both e and f. Let $t(e,f) := |\{h \in \mathcal{E} \mid h \cap e \neq \emptyset, h \cap f \neq \emptyset, i \notin h\}|$ be the number of such intersecting edges h.



The number of such edges is $t(e,f) \le (r-1)^2 \cdot o(D) \le o(D)$ and using a similar estimate as earlier we can write $\mathrm{Cov}[I_e, I_f \mid i \in V'] \le (1 - \frac{\alpha}{D})^{-t(e,f)} - 1 \le o(1)$. Then again $\Pr[d'(i) \notin (1 \pm o(1)) \cdot e^{-\alpha(r-1)}] \le o(1)$ and we have proven all necessary claims. □

Now we can show the main Theorem.

Proof of Theorem 1.7. Suppose the goal is a cover of an r-uniform hypergraph with only $(1+\delta)\frac{n}{r}$ many edges. We apply the previous lemma where the "size of the bite" is a tiny constant $\alpha := \frac{\delta}{2}$ and we apply Lemma 1.8 $t := \frac{1}{\alpha}\ln(\frac{2r}{\alpha})$ many times. We consider r and α as fixed and have $\varepsilon = o(1)$ and $N, D \to \infty$. We obtain a sequence of smaller and smaller hypergraphs $H_i = (V_i, \mathcal{E}_i)$ with approximate degree D_i. The number of non-covered nodes after t iterations is

$$|V_t| \le |V| \cdot (e^{-\alpha(1\pm o(1))})^t \le n \cdot \frac{\alpha}{2r} \cdot (1 \pm o(1)) < \frac{\alpha n}{r}$$

In every of the t iterations, we can compare the number of nodes that are covered with the number of sampled edges and see that the ratio always satisfies

$$\frac{\text{edges sampled in iteration}}{\text{nodes covered in iteration}} = \frac{\frac{\alpha}{r} \cdot (1 \pm o(1))}{1 - e^{-\alpha(1\pm o(1))}} < \frac{1+\alpha}{r}$$

In particular that means we are sampling $(1+\alpha)\frac{n}{r}$ many edges in total to cover $n\cdot(1-\frac{\alpha}{r})$ many nodes. We should mention that some number of $o(n)$ nodes in the original hypergraph might start with degree $D\cdot(1\pm o(1))$ and remain uncovered while the degree does not go down in a controlled way. But the original degree of D is of the form $O(D_i)$ which is sufficient.

Finally we cover the remaining nodes with one private edge per node. That provides a cover of size $(1+\alpha)\frac{n}{r}+\frac{\alpha}{r}\cdot n=(1+2\alpha)\frac{n}{r}$. This shows the claim. □

If the use of the nibble technique is still arcane to the reader, it maybe helpful that the following sampling method is "morally equivalent": take the approximately regular hypergraph $H=(V,\mathcal{E})$ and for each hyperedge $e\in\mathcal{E}$ pick a uniform random number $s_e\in[0,1]$. Now sort the edges $\mathcal{E}=\{e_1,\dots,e_m\}$ so that $0<s_{e_1}<s_{e_2}<\dots<s_{e_m}<1$. We create a matching $\mathcal{E}'\subseteq\mathcal{E}$ as follows. Starting with $\mathcal{E}':=\emptyset$ and consider the indices $i=1,\dots,m$ in increasing order. Add e_i to $\mathcal{E}'$ if e_i does not overlap any edge that was previously added to $\mathcal{E}'$. Then similarly $\mathcal{E}'$ should end up being a matching that covers a $1-o(1)$ fraction of nodes.

Now we can prove the Erdős-Hanani Conjecture:

Theorem 1.9 (Rödl 1985)**.** *The complete ℓ-uniform hypergraph $\mathcal{H}_{n,\ell}$ can be covered with $(1\pm o(1))\cdot\binom{n}{\ell}/\binom{k}{\ell}$ edges from $\mathcal{H}_{n,k}$ as $n\to\infty$.*

Proof. We define a hypergraph $H=(V,E)$ with nodes $V:=\binom{[n]}{\ell}$ and edges $E=\{\binom{S}{\ell}\mid S\in\binom{[n]}{k}\}$. Observe that this graph has $|V|=\binom{n}{\ell}$ vertices and is $\binom{k}{\ell}$-uniform. The degree of the vertices is $D:=\binom{n-\ell}{k-\ell}$ as for every ℓ-tuple $S_1\in V$ an edge is obtained by picking $S_2\in\binom{[n]\setminus S_1}{k-\ell}$ and taking the hyperedge corresponding to subsets of $S_1\dot\cup S_2$. Two distinct vertices $S_1,S_2\in\binom{[n]}{\ell}$ lie in at most $\binom{n-\ell-1}{k-\ell-1}=o(D)$ many joint hyperedges as $n\to\infty$. Hence there is a cover of H with only $(1+o(1))\cdot\frac{|V|}{\binom{k}{\ell}}$ hyperedges. □

1.6 Independent Sets in Locally Sparse Graphs

Consider an undirected graph $G=(V,E)$ with $|V|=n$ nodes and maximum degree d. One might wonder what size of an independent set one can guarantee, only depending on those parameters. It is an easy exercise to find an independent set of size $\frac{n}{d+1}$ (even in polynomial time). And this bound is tight if the graph consists of disjoint unions of $(d+1)$-size cliques. But maybe if the graph is *locally sparse* one could do better? A result of Ajtai, Komlós and Szemerédi [AKS81] showed that in a *triangle-free* graph, there is always an independent set of size $\Omega(\frac{n}{d}\log(d))$. Shearer [She83] later found a simpler and quite elegant proof, see also Chapter 6 in the textbook of Tao and Vu [TV10]. From the proofs it quickly becomes clear that the key property is that neighborhoods $N(v)$ need to contain large independent sets. We will see here a generalization of the result by Alon [Alo96], telling that a graph where each neighborhood $N(v)$ is $O(1)$-colorable contains an independent set of size $\Omega(\frac{n}{d}\log(d))$. Note that this is indeed a generalization as a graph is triangle-free if and only if each neighborhood $N(v)$ is 1-colorable.

Theorem 1.10. *Let $G = (V,E)$ be a graph with maximum degree d where each neighborhood $N(v)$ is r-colorable. Then G contains an independent set of size $\Omega(\frac{n}{d}\cdot\frac{\log(d)}{\log(r)})$.*

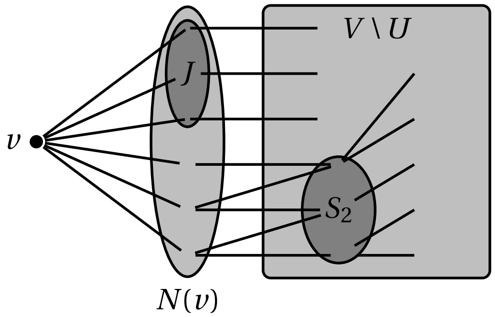
Proof. Let $\mathcal{I} := \{S \subseteq V \mid S \text{ is independent set}\}$ be the family of all independent sets in the graph. We consider the random experiment where we draw $S \sim \mathcal{I}$ uniformly at random. Our goal is to somehow argue that S is large in expectation. We consider the random variable

$$X_v := d\cdot|S\cap\{v\}| + |S\cap N(v)|$$

for each node $v \in V$. The sum over those random variables is a good proxy for the size of S since

$$\mathbb{E}\Big[\sum_{v\in V} X_v\Big] = d\cdot\mathbb{E}\Big[\underbrace{\sum_{v\in V}|S\cap\{v\}|}_{=|S|}\Big] + \mathbb{E}\Big[\underbrace{\sum_{v\in V}|S\cap N(v)|}_{\leq d|S|}\Big] \leq 2d\cdot\mathbb{E}[|S|]$$

We will prove that indeed $\mathbb{E}[X_v] \geq \Omega(\frac{\log(d)}{\log(r)})$ for each node, which then completes the claim. The trick is to lower bound $\mathbb{E}[X_v \mid ..]$ where we condition on what happens outside of v's neighborhood.



Claim I. *For $v \in V$, abbreviate $U := \{v\}\cup N(v)$ and fix any independent set $S_2 \subseteq V\setminus U$.* Then

$$\mathbb{E}[X_v \mid S\cap(V\setminus U) = S_2] \geq \Omega\Big(\frac{\log d}{\log r}\Big)$$

Proof of claim. Let $J := \{u \in N(v) \mid u \text{ not incident to } S_2\}$. Define $\mathcal{I}_1 := \{S_1 \subseteq J \mid S_1 \text{ is independent set}\}$. For each $S_1 \in \mathcal{I}_1 \cup \{v\}$, also $S_1 \dot\cup S_2$ is an independent set and they have to be sampled with uniform probability. In particular that means that $S \cap U$ (conditioned on S_2) produces a uniform sample from $\mathcal{I}_1 \cup \{v\}$. Then

$$\begin{aligned}\mathbb{E}[X_v] &= d \cdot \underbrace{\mathbb{E}[|S \cap \{v\}|]}_{=\frac{1}{|\mathcal{I}_1|+1}} + \mathbb{E}[|S \cap N(v)|] = \frac{d}{|\mathcal{I}_1|+1} + \underbrace{\frac{|\mathcal{I}_1|}{|\mathcal{I}_1|+1}}_{\geq 1/2} \underbrace{\mathop{\mathbb{E}}_{S_1 \sim \mathcal{I}_1}[|S_1|]}_{\geq \Omega(\frac{\log(|\mathcal{I}_1|)}{\log(r)})\ (*)} \\ &\geq \frac{d}{|\mathcal{I}_1|+1} + \Omega\Big(\frac{\log|\mathcal{I}_1|}{\log(r)}\Big) \overset{(**)}{\geq} \Omega\Big(\frac{\log(d)}{\log(r)}\Big)\end{aligned}$$

For (**) we use that if $|\mathcal{I}_1| \leq \sqrt{d}$, then the first terms is already $\Omega(\sqrt{d})$ and if $|\mathcal{I}_1| \geq \sqrt{d}$ then the 2nd term is large.

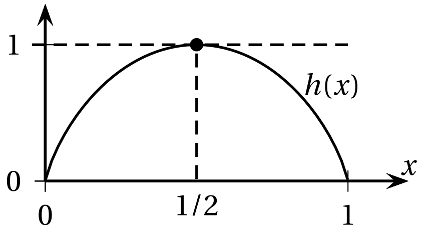
It still remains to argue $(*)$. Let us define a parameter $0 \leq \delta \leq 1$ so that $|\mathcal{I}_1| = 2^{\delta \cdot |J|}$. As $N(v)$ is r-colorable, $\mathcal{I}_1$ must contain at least all subsets of a $|J|/r$-size independent set. That implies that $2^{|J|/r} \leq |\mathcal{I}_1| = 2^{\delta |J|}$ and consequently $\delta \geq \frac{1}{r}$. Intuitively, this means that $\mathcal{I}_1$ is a fairly dense family of subsets of J. In particular we can use the estimate that we are about to prove as Claim II to get:

$$\mathop{\mathbb{E}}_{S_1 \sim \mathcal{I}_1}[|S_1|] \overset{\text{Claim II}}{\geq} \Omega\Big(\frac{|J| \cdot \delta}{\log(1/\delta)}\Big) \overset{\delta \geq \frac{1}{r}, |J|\delta = \log_2(\mathcal{I}_1)}{\geq} \Omega\Big(\frac{\log(|\mathcal{I}_1|)}{\log(r)}\Big)$$

that will finish the proof. □

Claim II. *Let $\mathcal{F} \subseteq 2^{[m]}$ be a family of $|\mathcal{F}| = 2^{\delta m}$ many subsets. Then $\mathbb{E}_{S \sim \mathcal{F}}[|S|] \geq \Omega(\frac{\delta m}{\log(1/\delta)})$.*

Proof of claim. Let us define the *binary entropy function* $h : [0,1] \to [0,1]$ by $h(x) := x \log_2(\frac{1}{x}) + (1-x)\log_2(\frac{1}{1-x})$.



For a random variable X, one defines the *entropy* as $H(X) = \sum_x \Pr[X = x] \cdot \log_2 \frac{1}{\Pr[X=x]}$ where the sum ranges over all events. For example $h(x)$ gives the entropy of a coin that gives head with probability x and tail with probability $1-x$. One useful property of entropy is that it is *subadditive*. In particular if $\boldsymbol{Y} \in \mathbb{R}^m$ is a random vector, then $H(\boldsymbol{Y}) \leq \sum_{i=1}^m H(Y_i)$, meaning that the total entropy is at most the sum of the entropy of the coordinates. Another useful fact is that the uniform distribution over N elements has entropy $\log_2(N)$.

Now sample a uniform element $S \sim \mathcal{F}$ and let $\boldsymbol{Y} \in \{0,1\}^m$ be the characteristic vector of S. We abbreviate $\alpha := \frac{\mathbb{E}[|S|]}{m}$. Then we can estimate that

$$\begin{aligned}\delta m \quad &= \quad \log_2(|\mathcal{F}|) = H(\boldsymbol{Y}) \overset{\text{subadditivity}}{\leq} \sum_{i=1}^{m} H(Y_i) = \sum_{i=1}^{m} h(\mathbb{E}[Y_i]) \\ &\overset{\text{concavity}}{\leq} \quad m \cdot h\Big(\underbrace{\frac{\mathbb{E}[|S|]}{m}}_{=\alpha}\Big) \overset{h(x)\leq 2x\log(\frac{1}{x})}{\leq} 2\alpha m \log\Big(\frac{1}{\alpha}\Big)\end{aligned}$$

Here we use Jensen's inequality with the concavity of h. The inequality can be rearranged to $\alpha \geq \Theta(\frac{\delta}{\log(1/\delta)})$. □

For a better intuition, let us revisit the arguments of the proof for $r = \Theta(1)$. If an adversary picks the set S_2 so that there are less than $\sqrt{d}$ many independent sets in $\{v\} \cup N(v)$, then the chance that v is picked is at least $\frac{1}{\sqrt{d}}$ and hence $X_v \geq \Theta(\sqrt{d})$. On the other hand, suppose S_2 is picked so that there are a lot of independent sets so that v's contribution is not enough. Say we have $|\mathcal{I}_1| = d$. But if these are all the independent sets contained in J and J is $O(1)$-colorable, then we know that $|J| \leq O(\log d)$. In particular the average set from $\mathcal{I}_1$ must quite large, at least $|S| \geq \Omega(\log d)$.

We also want to comment on the bound behind Claim II. The inequality indeed gives the right asymptotics. For example if $|\mathcal{F}| = 2^{\Theta(m)}$ it is rather intuitive that the average set in F must have size $\Theta(m)$. For the other end of the spectrum suppose that $|\mathcal{F}| = \text{poly}(m) = 2^{\delta \cdot m}$ with $\delta = \frac{O(\log m)}{m}$ and the average size is at least $\Theta(\frac{\delta m}{\log(1/\delta)}) = \Theta(\frac{\log m}{\log(m)}) = \Theta(1)$ as it should be.

1.7 Open problems

The exact range of Ramsey numbers is still unknown. The best estimates are

$$\Theta(k) \cdot 2^{k/2} \leq R(k,k) \leq 2^{2k-\Theta(\log(k)^2/\log\log k)},$$

which still leaves a significant gap [CFS15]. As of today also smaller Ramsey numbers such as $R(5,5)$ are unknown. A related open problem is the *Erdős-Hajnal Conjecture* that for every fixed graph H the following holds: A graph $G = (V,E)$ on n-vertices that does not have H as an induced subgraph must have a clique or independent set of size $n^{c(H)}$, where $c(H) > 0$ is a constant only depending on H.

1.8 Exercises

Exercise 1.1. Prove that for all $k, r \in \mathbb{N}$, there exists a constant $N(k,r)$ so that the following holds: Let $K_n = ([n], E)$ be the complete graph on $n \geq N(k,r)$ nodes and let $\chi : E \to \{1, \ldots, k\}$ be a coloring of the edges with k colors. Then there is a monochromatic subgraph with at least r nodes.

Exercise 1.2. Let $X = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n \in \{-1, 1\}$ are independent random variables with $\Pr[X_i = 1] = \Pr[X_i = -1] = \frac{1}{2}$. Prove that $\mathbb{E}[|X|] = \Theta(\sqrt{n})$.
Remark. There are certainly many ways to prove this fact. Can you come up with a short elementary argument?

Exercise 1.3. We call a set $A \subseteq \mathbb{Z} \setminus \{0\}$ *sum-free* if there are no distinct $a_1, a_2, a_3 \in A$ so that $a_1 + a_2 = a_3$. Our goal is to prove that any set $B = \{b_1, \ldots, b_n\} \subseteq \mathbb{Z} \setminus \{0\}$ contains a subset $A \subseteq B$ that is sum-free and has size $|A| \geq n/3$.

i) Pick $p = 3k+2$ as a large prime so that $p > 2|b_i|$ for all i and $k \in \mathbb{N}$. Consider the middle third of elements $C := \{k+1, \ldots, 2k+1\}$. Prove that there is an $r \in \{1, \ldots, p-1\}$ so that $|\{r \cdot b \mod p \mid b \in B\} \cap C| \geq \frac{n}{3}$.

ii) Take the choice of r from $i)$ and define $A := \{b \in B \mid (r \cdot b \mod p) \in C\}$. Prove that A is sum-free.

Exercise 1.4 (From Alon & Spencer [AS16])**.** Let $\mathcal{F} \subseteq 2^{[n]}$ be a family of sets that is *inclusion-free* meaning that there are no $A, B \in \mathcal{F}$ with $A \subset B$. Prove that $|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$.
Hint. Pick a uniform random permutation $\pi : [n] \to [n]$ and consider the random variable $X := |\{i \in [n] \mid \{\pi(1), \ldots, \pi(i)\} \in \mathcal{F}\}|$.

Exercise 1.5 (From Alon & Spencer [AS16])**.** Let $G = (A\dot{\cup}B, E)$ be a bipartite graph with n vertices in total. Each vertex $v \in A\dot{\cup}B$ has a list $S(v) \subseteq \{1, \ldots, k\}$ of $|S(v)| > \log_2(n)$ many colors. Prove that there is a proper coloring $\chi : A\dot{\cup}B \to \{1, \ldots, k\}$ where each node $v \in A\dot{\cup}B$ receives a color $\chi(v) \in S(v)$ from its list.

Exercise 1.6. Provide a family of graphs $G = (V, E)$ which is triangle-free and in which every node has degree $\Theta(d)$. Moreover $\mathcal{I} := \{S \subseteq V \mid S \text{ is independent set}\}$ should satisfy the following properties:

1. There is a node $u \in V$ so that $\Pr_{S \sim \mathcal{I}}[u \in S] \leq o(\frac{1}{d})$ if $d \to \infty$.
2. There is a node $v \in V$ so that $\mathbb{E}_{S \sim \mathcal{I}}[|N(v) \cap S|] \leq o(1)$.

Exercise 1.7. A d-regular graph $G = (V, E)$ is called a *β-expander* for $\beta > 0$ if

$$|\delta(S)| \geq \beta \cdot d \cdot |S| \quad \forall S \subseteq V : 1 \leq |S| \leq \frac{|V|}{2}$$

Pick 3 uniform random perfect matchings M_1, M_2, M_3 on nodes $[n]$ with n even. Prove that the graph $G = ([n], M_1 \cup M_2 \cup M_3)$ is a 3-regular $\Omega(1)$-expander with good probability (here we count multi-edges with multiplicity).

Exercise 1.8. Recall that for an undirected graph $G = (V, E)$, a matching $M \subseteq E$ is a set of edges that do not share any vertices. Also recall that G is *d-regular* if all degrees are d. Moreover we consider graphs without multi-edges and self-loops.

i) Prove the following statement: For every $\varepsilon > 0$ there is a $d_0 \in \mathbb{N}$ so that every d-regular graph $G = (V, E)$ with $d \geq d_0$ has a matching with at least $(\frac{1}{2} - \varepsilon) \cdot |V|$ many edges.

ii) Prove that at least approximate regularity is needed by providing a graph $G = (V, E)$ where all degrees are in $\{d, 2d\}$ but no matching has more than $|V|/3$ many edges.

Chapter 2

Concentration inequalities

Concentration of measure is a phenomenom that is particularly useful in probabilistic combinatorics. In this chapter, we want to present a variety of powerful concentration inequalities and theirs proofs. Throughout this chapter we do not aim to optimize constants but rather focus on a clean exposition of the key ideas.

2.1 Chernov bounds

The most basic case of concentration can be studied for the sum of independent random variables. The proof that we will see here follows classical work of Bernstein and Chernov. The idea is to first estimate the quantity $\mathbb{E}[\exp(tX)]$ for a suitable parameter $t > 0$, and then apply Markov's inequality.

Theorem 2.1 (Chernov Bound I)**.** *Suppose that $X_1, \dots, X_n$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq a_i$ where $\boldsymbol{a} \in \mathbb{R}^n_{\geq 0}$. Then for any $\lambda \geq 0$, the sum $X := X_1 + \dots + X_n$ satisfies $\Pr[|X| \geq \lambda \|\boldsymbol{a}\|_2] \leq 2e^{-\lambda^2/4}$.*

Proof. Let $t > 0$ be a parameter that we choose later. First suppose that $|t \cdot a_i| \leq 1$ for each i. Then

$$\begin{aligned}
\mathbb{E}[\exp(tX)] &\overset{\text{independence}}{=} \prod_{i=1}^n \mathbb{E}[\exp(tX_i)] \overset{e^x \leq 1+x+x^2 \forall |x| \leq 1}{\leq} \prod_{i=1}^n \mathbb{E}\left[1 + tX_i + t^2 X_i^2\right] \\
&= \prod_{i=1}^n \Big(1 + t\underbrace{\mathbb{E}[X_i]}_{=0} + t^2 \underbrace{\mathbb{E}[X_i^2]}_{\leq a_i^2}\Big) \leq \prod_{i=1}^n \left(1 + t^2 a_i^2\right) \overset{1+x \leq e^x}{\leq} \exp(t^2 \|\boldsymbol{a}\|_2^2)
\end{aligned}$$

Note that if we have some indices i where $ta_i > 1$, we can simply replace X_i by it's maximum value and get that $\mathbb{E}[\exp(tX_i)] \leq e^{ta_i} \leq e^{t^2 a_i^2}$. In other words, the estimate above still remains true. Either way,

$$\begin{aligned}\Pr[X > \lambda\|\boldsymbol{a}\|_2] &\overset{\text{monotonicity}}{=} \Pr\big[\exp(tX) > \exp(t\lambda\|\boldsymbol{a}\|_2)\big] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}[\exp(tX)]}{\exp(t\lambda\|\boldsymbol{a}\|_2)} \\ &\leq \exp(t^2\|\boldsymbol{a}\|_2^2 - t\lambda\|\boldsymbol{a}\|_2) \overset{t:=\frac{\lambda}{2\|\boldsymbol{a}\|_2}}{\leq} \exp\Big(-\frac{1}{4}\lambda^2\Big)\end{aligned}$$

□

Often one has $\{0,1\}$-events and then a different form of Chernov bound is useful. We will skip the proof that just uses different parameter settings.

Theorem 2.2 (Chernoff Bound II)**.** *Let* $X_1,\dots,X_n \in \{0,1\}$ *independent random variables with* $X := X_1 + \ldots + X_n$*. Then for* $0 < \delta < 1$ *one has*

$$\mathbb{E}\big[|X - \mathbb{E}[X]| \geq \delta\,\mathbb{E}[X]\big] \leq 2\exp\Big(-\frac{\delta^2}{3}\,\mathbb{E}[X]\Big).$$

An there is one more form that can be useful sometimes:

Theorem 2.3 (Chernoff Bound III)**.** *Let* $X_1,\dots,X_n \in \{0,1\}$ *independent random variables with* $X := X_1 + \ldots + X_n$*. Then for* $\delta > 0$ *one has*

$$\Pr[X > (1+\delta)\,\mathbb{E}[X]] < \Big(\frac{e^\delta}{(1+\delta)^{1+\delta}}\Big)^{\mathbb{E}[X]}$$

This expression is slightly more arcane. For all $\delta \geq 2$ then Chernov Bound III can be simplified to as

$$\Pr[X > \delta \cdot \mathbb{E}[X]] \leq \exp\Big(-\frac{1}{4}\ln(\delta)\cdot\delta\cdot\mathbb{E}[X]\Big).$$

2.1.1 Qualitative difference between the inequalities

The difference between the Chernov bounds appears rather cosmetic / technical on first glance. But we want to point out a crucial qualitative difference. For that sake, we reparameterize the 2nd Chernov bound. Suppose we have independent random variables $X_1,\dots,X_n \in \{0,1\}$ with $\Pr[X_i = 1] = \frac{\sigma^2}{n}$. Then it is easy to see that $\mathrm{Var}[X_i] = \Theta(\frac{\sigma^2}{n})$ and hence $\mathrm{Var}[X] = \Theta(\sigma^2)$, while also $\mathbb{E}[X] = n\cdot\frac{\sigma^2}{n} = \sigma^2$. Then we pick $0 < \lambda < \sigma$ so that $\delta = \frac{\lambda}{\sigma}$. Then we can rephrase the Chernov Bound II as

$$\Pr\Big[|X - \mathbb{E}[X]| \geq \underbrace{\underbrace{\delta}_{=\frac{\lambda}{\sigma}}\underbrace{\mathbb{E}[X]}_{=\sigma^2}}_{=\lambda\cdot\sigma}\Big] \leq 2\exp\Big(-\frac{1}{3}\underbrace{\delta^2}_{=\frac{\lambda^2}{\sigma^2}}\underbrace{\mathbb{E}[X]}_{=\sigma^2}\Big) = 2\exp\Big(-\frac{\lambda^2}{3}\Big)$$

which recovers the 1st Chernov bound (ignoring the different constants). But strangely, in Chernov II we are restricted to have $\delta < 1$ (or $\lambda < \sigma$), while this was not needed for Chernov I. On the other hand, if we reparameterize Chernov III, then we get $\Pr[X > \mathbb{E}[X] + \lambda\sigma] \leq \exp(-\Theta(\ln(\frac{\lambda}{\sigma})) \cdot \lambda \cdot \sigma)$, which gives a weaker exponential decay in λ.

The explanation is that in the setting of Chernov I, the deviation of the individual terms X_i is fully controlled by the upper bound of a_i^2 on its variance and even the individual terms themselfs would satisfy a concentration inequality of the form $\Pr[|X_i - \mathbb{E}[X_i]| > \lambda \cdot \sqrt{\text{Var}[X_i]}] \leq 2\exp(-\lambda^2/4)$ for all $\lambda \geq 0$.

Now consider the setting of Chernov II and III — say with $\Pr[X_i = 1] = p_i$ being small. Then an individual terms gives a deviation of approximately 1 from the expectation with probability p_i. If the deviation parameter δ (or λ) is small enough, the the sum X behaves like a Gaussian in terms of concentration (see Chernov II), but if the parameters δ (or λ) are getting too large then the rather unpredictable behavior of the individual terms dominates.

2.2 Martingale concentration

The proof of the Chernov bounds from above seemed to crucially use independence. But it turns out that a weaker concept suffices to get the same strong concentration effect. For a more detailed reading we refer to Chapter 7 of [AS16].

Consider a sequence $X_0, \dots, X_n$ of real-valued random variables that satisfy $\mathbb{E}[X_i \mid X_1, \dots, X_{i-1}] = X_{i-1}$ for $i = 1, \dots, n$. Such a sequence is called a *Martingale*. In particular, a random variable X_i is allowed to depend on the outcomes of the previous random variables $X_1, \dots, X_{i-1}$, but in expectation X_i needs to coincide with the previous value X_{i-1}. In particular $\mathbb{E}[X_i] = X_0$ for all i. A classical example is a *gambler* in a casino that only offers "fair" games in the sense that in expectation the gambler neither wins nor looses money, no matter how he plays. The gambler may switch between games depending on whether he has a winning streak or not. But no matter what strategy he uses, his expectation is always 0 and the probability to deviate significantly is tiny. The simplest form of a Martingale concentration result is the following:

Theorem 2.4 (Azuma's Inequality). *Let $0 = X_0, \dots, X_n$ be a Martingale with $|X_t - X_{t-1}| \leq 1$ for all $t = 1, \dots, n$. Then for any $\lambda \geq 0$ one has*

$$\Pr[|X_n| > \lambda\sqrt{n}] \leq 2\exp(-\lambda^2/4).$$

Actually we will even prove a more general result as we explain later. For the proof, it will be notationally more convenient to work with the *increments*

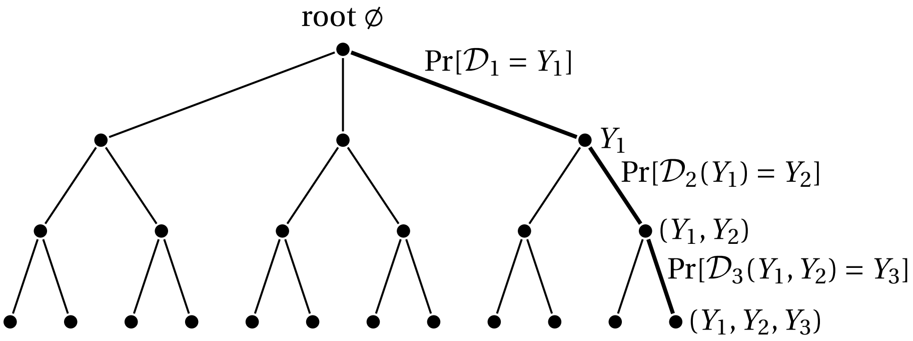
$Y_1,\dots,Y_n$ instead of the *summands* $X_t = Y_1 + \dots + Y_t$. We will also be lazy and assume that the sampling space is finite, so that we can talk about probabilities instead of densities (this does not affect correctness of the concentration results).

We imagine that we sample first $Y_1 \sim \mathcal{D}_1$, then $Y_2 \sim \mathcal{D}_2(Y_1)$ depending on the previous outcome. Then the tth number is sampled as $Y_t \sim \mathcal{D}_t(Y_1,\dots,Y_{t-1})$. As before, we assume that the distributions are *balanced*, that means

$$\mathop{\mathbb{E}}_{Y_t\sim\mathcal{D}_t(Y_1,\dots,Y_{t-1})}[Y_t] = 0 \quad \forall Y_1,\dots,Y_{t-1}$$

We also assume that $|Y_t| \leq 1$ for any possible outcome of Y_t. Our goal will be to give a concentration result for the sum $\sum_{i=1}^n Y_i$ of those increments.

There is a concrete way to interpret this random process. Consider a tree $T = (V,E)$ with a node $(Y_1,\dots,Y_t)$ for each possible outcome and each $t \in \{0,\dots,n\}$. We insert edges between $(Y_1,\dots,Y_{t-1})$ and $(Y_1,\dots,Y_t)$ that we label with probability $\Pr[\mathcal{D}_t(Y_1,\dots,Y_{t-1}) = Y_t]$. We also label nodes $(Y_1,\dots,Y_{t-1})$ with the variance $\mathrm{Var}[\mathcal{D}_t(Y_1,\dots,Y_{t-1})]$. Note that the root corresponds to the empty string $\emptyset$ and the leafs correspond to "fully determined" vectors $(Y_1,\dots,Y_n)$. A path from the root down to a leaf then corresponds to a *sample path*. For each node $(Y_1,\dots,Y_t)$ we define $V(Y_1,\dots,Y_t) := \sum_{i=1}^t \mathrm{Var}[\mathcal{D}_i(Y_1,\dots,Y_{i-1})]$ as the sum of the variances suffered on the sample path from the root down to that node.


root $\emptyset$
$\Pr[\mathcal{D}_1 = Y_1]$
Y_1
$\Pr[\mathcal{D}_2(Y_1) = Y_2]$
(Y_1, Y_2)
$\Pr[\mathcal{D}_3(Y_1, Y_2) = Y_3]$
(Y_1, Y_2, Y_3)

sampling tree for $n = 3$

The interesting point about Martingales is that the variance of the random process may vary a lot for different sample paths.

Theorem 2.5 (Freedman's Martingale Concentration Inequality)**.** *Let* $0 = X_0,\dots,X_n$ *be a Martingale with* $X_i := Y_1 + \dots + Y_i$ *so that* $|Y_i| \leq 1$*. Let* $V[Y_1,\dots,Y_n]$ *be the sum of the variances on the sample path leading to node* $(Y_1,\dots,Y_n)$*. Then for any* $\lambda \geq 0$ *and* $\sigma \geq 0$ *one has*

$$\Pr\Big[\big|X_n\big| \geq \lambda\cdot\sigma \text{ and } V[Y_1,\dots,Y_n] \leq \sigma^2\Big] \leq 2\exp\Big(-\frac{1}{8}\min\big\{\lambda^2, \lambda\cdot\sigma\big\}\Big)$$

Proof. We only show the upper bound. Let $0 \le \alpha \le \frac{1}{2}$ be a parameter that we determine later.

Claim. Let $\Phi_t := \exp(\alpha X_t - 2\alpha^2 V[Y_1,\dots,Y_t])$. Then $\mathbb{E}[\Phi_t] \le 1$ for $t \in \{0,\dots,n\}$.

Proof of claim. Note that the claim says that for every level t in the sampling tree, the average of Φ_t over nodes in that level is at most 1. Intuitively, Φ_t is the exponential moment of the Martingale, but we are discounting a factor that is proportional to the suffered variance until that point (and the variance is not necessarily uniform). The claim is easily proven by inducton. The proof will simply show that for any node in the tree, if we move to a random child, the expression defining Φ_t does not increase in expectation.

Now the formal proof. Clearly $\Phi_0 = 1$. For a general level $t \ge 1$, fix any $Y_1,\dots,Y_{t-1}$ — that means Φ_{t-1} is already determined — and draw $Y_t \sim \mathcal{D}_t(Y_1,\dots,Y_{t-1})$. Then

$$\begin{aligned}
\mathbb{E}[\Phi_t \mid Y_1,\dots,Y_{t-1}] &= \mathop{\mathbb{E}}_{Y_t}\Big[\exp\Big(\alpha\underbrace{(X_{t-1}+Y_t)}_{=X_t} - 2\alpha^2\underbrace{(V[Y_1,\dots,Y_{t-1}]+\mathrm{Var}[Y_t])}_{=V[Y_1,\dots,Y_t]}\Big)\Big] \\
&= \underbrace{\exp\Big(\alpha X_{t-1} - 2\alpha^2 V(Y_1,\dots,Y_{t-1})\Big)}_{=\Phi_{t-1}} \cdot \mathop{\mathbb{E}}_{Y_t}\big[\exp\big(\alpha Y_t - 2\alpha^2\mathrm{Var}[Y_t]\big)\big] \\
&\le \Phi_{t-1}\cdot\mathbb{E}\Big[1+\alpha\underbrace{\mathbb{E}[Y_t]}_{=0} + \underbrace{\alpha^2\,\mathbb{E}[Y_t^2] - \frac{2\alpha^2}{2}\cdot\mathrm{Var}[Y_t]}_{=0}\Big] = \Phi_{t-1}
\end{aligned}$$

Here we use the convenient fact that $\exp(x-y) \le 1+x+x^2-\frac{y}{2}$ for $-1 \le x \le 1$ and $0 \le y \le 1$. Here we crucially use that $|Y_t| \le 1$ and $\mathrm{Var}[V_t] \le 1$ and $\alpha \le \frac{1}{2}$. □

Now we can apply the trick of exponentiating and applying Markov's inequality that we have used before. Making the choice of $\alpha := \min\{\frac{\lambda}{4\sigma}, \frac{1}{2}\}$ then gives

$$\begin{aligned}
\Pr\Big[\underbrace{X_n - 2\alpha V[Y_1,\dots,Y_n] \ge \frac{\lambda\sigma}{2}}_{(**)}\Big] &\overset{(***)}{=} \Pr\Big[\underbrace{\exp\Big(\alpha X_n - 2\alpha^2 V[Y_1,\dots,Y_n]\Big)}_{\mathbb{E}[..]\le 1} \ge \exp\Big(\alpha\frac{\lambda\sigma}{2}\Big)\Big] \\
&\overset{\text{Markov}}{\le} \exp\Big(-\alpha\frac{\lambda\sigma}{2}\Big) = \begin{cases}\exp(-\frac{1}{4}\lambda\cdot\sigma) & \text{if } \lambda \ge 2\sigma \\ \exp(-\frac{1}{8}\lambda^2) & \text{if } \lambda < 2\sigma\end{cases}
\end{aligned}$$

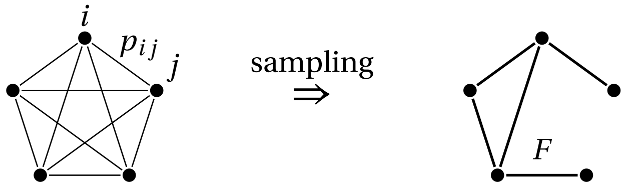
Here we use monotonicity of $x \mapsto \exp(\alpha x)$ in $(***)$. If the event in $(**)$ does not occur and additionally $V[Y_1,\dots,Y_n] \le \sigma^2$, then indeed

$$X_n \le \frac{\lambda\sigma}{2} + 2\underbrace{\alpha}_{\le\frac{\lambda}{4\sigma}}\underbrace{V[Y_1,\dots,Y_n]}_{\le\sigma^2} \le \lambda\sigma$$

as desired. □

2.2.1 An Application to the Size of Independent Sets

We want to demonstrate the power and usefulness of Martingales with an example application. Consider an undirected complete graph $G = (V, E)$ with nodes $|V| = n$ and suppose every edge $e \in E$ is labelled with a probability p_e. Independently every edge e "materializes" with the given probability p_e. Let $F \subseteq E$ be the obtained random sample of edges. We are interested in the quantity $\alpha(F) := \max\{|S| : S \subseteq V \text{ is an independent set w.r.t. } F\}$.



In particular we want to know whether $\alpha(F)$ is well concentrated. But there are several issues; in particular we have no closed formula for $\mathbb{E}[\alpha(F)]$. In fact, even if $p_e \in \{0,1\}$, determining $\mathbb{E}[\alpha(F)]$ within a factor of $n^{1-\varepsilon}$ is an **NP**-hard problem for any fixed $\varepsilon > 0$. So, how can be show a quantity that we cannot determine is well concentrated? This is where Martingales game into play:

Lemma 2.6. *For any probability vector $\boldsymbol{p} \in [0,1]^E$ and any $\lambda \geq 0$ one has* $\Pr[|\alpha(F) - \mathbb{E}[\alpha(F)]| > \lambda\sqrt{n}] \leq 2\exp(-\lambda^2/8)$.

Proof. Let $V = \{1, \ldots, n\}$ be the vertices in their natural ordering. Let $E_k := \{(i,k) \in E \mid i < k\}$ be all the edges between $\{1, \ldots, k-1\}$ and node k. Clearly $E = E_1 \dot\cup \ldots \dot\cup E_n$. We can also write $F_k := E_k \cap F$ and imagine that we sample the sets $F_1, \ldots, F_n$ one after the other in order to determine F. We abbreviate $F_{\leq k} := F_1 \cup \ldots \cup F_k$ and similarly we define $F_{\geq k} := F_k \cup \ldots \cup F_n$. We define a random variable

$$X_k := \mathop{\mathbb{E}}_{F_{\geq k+1}} [\alpha(F) \mid F_1, \ldots, F_k]$$

In other words, X_0 is deterministically the number $\mathbb{E}[\alpha(F)]$ and X_n is equal to the random variable $\alpha(F)$. Phrased differently we arrive at the random variable $\alpha(F)$ by revealing the F_k's one after the other.

Now suppose that $F_1, \ldots, F_k$ and hence X_k have been decided. Then

$$\mathop{\mathbb{E}}_{F_{k+1}} [X_{k+1} \mid F_{\leq k}] \overset{\text{Def } X_{k+1}}{=} \mathop{\mathbb{E}}_{F_{k+1}} \Big[\mathop{\mathbb{E}}_{F_{\geq k+2}} \big[\alpha(F) \mid F_{\leq k+1}\big] \mid F_{\leq k} \Big] = \mathop{\mathbb{E}}_{F_{\geq k+1}} \big[\alpha(F) \mid F_{\leq k}\big] = X_k.$$

That means $X_0, \ldots, X_k$ is Martingale. The only thing that remains to be checked is that the difference of the intermediate random variables is bounded and the

overall claim will follow:
Claim. *One always has* $|X_{k+1} - X_k| \leq 1$.
Proof of claim. We fix any outcome for $F_1, \dots, F_k$. We can be generous and even fix $F_{k+2}, \dots, F_n$. Let $S_{\max}$ be the maximum independent set if $F_k = \emptyset$ and let $S_{\min}$ be the maximum independent set if $F_k = E_k$. In other words, $S_{\max}$ is the largest possible outcome and $S_{\min}$ is the smallest possible outcome. But $||S_{\min}| - |S_{\max}|| \leq 1$ which is easy to see as simply dropping node k from $S_{\max}$ (if it was even in there) gives an independent set even if all edges in E_k materialize. □

For more applications and background on Martingales we refer to the excellent treatment in Alon and Spencer [AS16].

2.3 Gaussian Concentration

The content of this chapter is largely taken from the lecture notes "Concentration Inequalities" by Lalley[1]. Recall that $N(0,1)$ is the 1-dimensional *Gaussian distribution* with *density* $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. More generally, $N^n(0,1)$ is the Gaussian distribution over vectors $\boldsymbol{x} \in \mathbb{R}^n$ with density $\frac{1}{(2\pi)^{n/2}} e^{-\|\boldsymbol{x}\|_2^2/2}$. The Gaussian distribution has many useful properties (that can be derived straightforwardly from the density function):

- One can get a sample $\boldsymbol{x} \sim N^n(0,1)$ also by sampling each coordinate $x_i \sim N(0,1)$ independently.
- If $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are orthogonal unit vectors and $\boldsymbol{x} \sim N^n(0,1)$, then $\langle \boldsymbol{a}, \boldsymbol{x} \rangle, \langle \boldsymbol{b}, \boldsymbol{x} \rangle \sim N(0,1)$ are independent random variables.
- For any vector $\boldsymbol{a} \in \mathbb{R}^n$ with $\|\boldsymbol{a}\|_2 = 1$ and $x_1, \dots, x_n \sim N(0,1)$, one has $\sum_{i=1}^n a_i x_i \sim N(0,1)$.

We call a function $F : \mathbb{R}^n \to \mathbb{R}$ *Lipschitz* if $|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. A natural Lipschitz function is of course $F(\boldsymbol{x}) := \|\boldsymbol{x}\|_2$ and we could in fact use the machinery that we have seen so far to derive a concentration inequality $\Pr[\|\boldsymbol{x}\|_2 > \sqrt{n} + \lambda] \leq \exp(-\Theta(\lambda^2))$ by using the fact that $\|\boldsymbol{x}\|_2^2$ is a sum of independent random variables. But it turns out that one can prove such remarkable concentration inequality even for "unstructured" functions as long as they are Lipschitz:

[1] See `https://galton.uchicago.edu/~lalley/Courses/386/Concentration.pdf`

Theorem 2.7. *For any Lipschitz-function $F : \mathbb{R}^n \to \mathbb{R}$ with mean $\mu := \mathbb{E}_{\boldsymbol{x}\sim N^n(0,1)}[F(\boldsymbol{x})]$ and any $\lambda \geq 0$ one has*

$$\Pr_{\boldsymbol{x}\sim N^n(0,1)}\Big[|F(\boldsymbol{x}) - \mu| > \lambda\Big] \leq 2e^{-\lambda^2/2}.$$

For example if $\boldsymbol{x}$ is a random Gaussian, then $\|\boldsymbol{x}\|_2 = \sqrt{n} \pm O(1)$ with probability 99.99% — that means the standard deviation of the length is only a constant.

2.3.1 Exponential moments of Gaussians

Before we start, we want to discuss a couple of useful facts. First we need a well-known result about the *exponential moment* of a Gaussian. This can be easily obtained by integrating, but our approach might be more intuitive in explaining why the value is what it is:

Lemma 2.8. *For any $\lambda \in \mathbb{R}$ one has $\mathbb{E}_{x\sim N(0,1)}[\exp(\lambda x)] = \exp(\lambda^2/2)$.*

Proof. We will use only one property of Gaussians that we mentioned earlier: namely that if we generate independent random Gaussians $g_1, \ldots, g_k \sim N(0,1)$, then $x := \frac{1}{\sqrt{k}}(g_1 + \ldots + g_k) \sim N(0,1)$. Note that the 2nd degree Taylor polynomial of $\exp(x)$ is $1 + x + \frac{1}{2}x^2$, which we use in $(*)$. Now fix a value $\lambda \in \mathbb{R}$. In the following estimate we will write $\approx$ when ever we make a lower order error that goes to 0 if we send $k \to \infty$. Then

$$\begin{aligned}\mathbb{E}[\exp(\lambda x)] &= \mathbb{E}\Big[\exp\Big(\sum_{i=1}^{k} \frac{\lambda}{\sqrt{k}} g_i\Big)\Big] \overset{\text{independence}}{=} \prod_{i=1}^{k} \mathbb{E}\Big[\exp\Big(\frac{\lambda}{\sqrt{k}} g_i\Big)\Big] \\ &\overset{(*)}{\approx} \prod_{i=1}^{k} \mathbb{E}\Big[1 + \frac{\lambda}{\sqrt{k}} g_i + \frac{1}{2}\cdot\Big(\frac{\lambda}{\sqrt{k}} g_i\Big)^2\Big] = \Big(1 + \frac{\lambda^2/2}{k}\Big)^k \approx \exp(\lambda^2/2).\end{aligned}$$

In $(*)$ we make an error of $O((\frac{1}{\sqrt{k}})^3)$ in each factor by using the 2nd degree Taylor approximation. The claim follows. □

By applying the previous lemma with $\lambda' := \lambda\|\boldsymbol{a}\|_2$ we get:

Lemma 2.9. *For any $\lambda \in 0$ and $\boldsymbol{a} \in \mathbb{R}^n$ one has $\mathbb{E}_{\boldsymbol{x}\sim N^n(0,1)}[\exp(\lambda\langle \boldsymbol{a}, \boldsymbol{x}\rangle)] \leq e^{-\frac{1}{2}\lambda^2\|\boldsymbol{a}\|_2^2}$.*

Any Lipschitz function can be easily "smoothened" so that it changes by at most ε and the smoothened function is differentiable everywhere. Moreover, any differentiable function that is Lipschitz has a gradient that is bounded:

Lemma 2.10. *Suppose $F : \mathbb{R}^n \to \mathbb{R}$ is differentiable and Lipschitz. Then $\|\nabla F(\boldsymbol{x})\|_2 \leq 1$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Proof. Fix a point $\boldsymbol{x} \in \mathbb{R}^n$, then for all vectors $\boldsymbol{h}$ small enough, there is a constant $C > 0$ so that $|F(\boldsymbol{x}) - F(\boldsymbol{x}+\boldsymbol{h})| = |\langle \boldsymbol{h}, \nabla F(\boldsymbol{x})\rangle| \pm C\|\boldsymbol{h}\|_2^2$. Setting $\boldsymbol{h} := \varepsilon \cdot \nabla F(\boldsymbol{x})$ one can get

$$1 \overset{\text{Lipschitz}}{\geq} \frac{|F(\boldsymbol{x}) - F(\boldsymbol{x}+\varepsilon\nabla F(\boldsymbol{x}))|}{\|\varepsilon\nabla F(\boldsymbol{x})\|_2} \geq \frac{\langle \varepsilon\nabla F(\boldsymbol{x}), \nabla F(\boldsymbol{x})\rangle - C\|\varepsilon\nabla F(\boldsymbol{x})\|_2^2}{\|\varepsilon\nabla F(\boldsymbol{x})\|_2} = (1-C\varepsilon)\cdot\|\nabla F(\boldsymbol{x})\|_2$$

Sending $\varepsilon \to 0$ implies that $\|\nabla F(\boldsymbol{x})\|_2 \leq 1$. □

Now we come to the lemma that is also called the *dublication trick*. We will bound the exponential moment of the *difference* of the function value at two independent Gaussians.

Lemma 2.11. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be differentiable and Lipschitz. Then*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_0,\boldsymbol{x}_1 \sim N^n(0,1)}\left[\exp\left(\lambda F(\boldsymbol{x}_1) - \lambda F(\boldsymbol{x}_0)\right)\right] \leq e^{\frac{\pi^2}{8}\lambda^2}$$

Proof. For $0 \leq t \leq 1$, let us define interpolate between the two samples by defining

$$\boldsymbol{x}_t := \cos\left(t \cdot \frac{\pi}{2}\right)\cdot \boldsymbol{x}_0 + \sin\left(t\cdot\frac{\pi}{2}\right)\cdot\boldsymbol{x}_1.$$

Note that for every t, the vector $\boldsymbol{x}_t \sim N^n(0,1)$ is a standard Gaussian. Similarly for $0 \leq t \leq 1$, the vector

$$\boldsymbol{y}_t := -\sin\left(t\cdot\frac{\pi}{2}\right)\boldsymbol{x}_0 + \cos\left(t\cdot\frac{\pi}{2}\right)\cdot\boldsymbol{x}_1$$

is distributed as $\boldsymbol{y}_t \sim N^n(0,1)$. Note that for every t, the pair $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ is *independent* as $\cos(\frac{\pi}{2}\cdot t)\cdot(-\sin(t\cdot\frac{\pi}{2}))+\sin(t\cdot\frac{\pi}{2})\cdot\cos(t\cdot\frac{\pi}{2}) = 0$. Next, consider the derivative of the interpolation:

$$\frac{d\boldsymbol{x}_t}{dt} = -\frac{\pi}{2}\cdot\sin\left(t\cdot\frac{\pi}{2}\right)\boldsymbol{x}_0 + \frac{\pi}{2}\cdot\cos\left(t\cdot\frac{\pi}{2}\right)\boldsymbol{x}_1 = \frac{\pi}{2}\cdot\boldsymbol{y}_t.$$

The basic idea behind the proof of the main claim is to track the expectation when one interpolates between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. Using the *Fundamental Theorem of Calculus* we get:

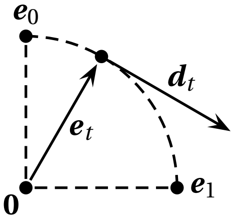
$$F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) = \int_0^1 \langle \nabla F(\boldsymbol{x}_t), \frac{d\boldsymbol{x}_t}{dt}\rangle\, dt = \frac{\pi}{2}\int_0^1 \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{y}_t\rangle\, dt \qquad (*)$$

Now we can express the expectation as

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(F(\boldsymbol{x}_1)-F(\boldsymbol{x}_0))] &\overset{(*)}{=} \mathbb{E}\Big[\exp\Big(\lambda\frac{\pi}{2}\int_0^1 \langle\nabla F(\boldsymbol{x}_t),\boldsymbol{y}_t\rangle\,dt\Big)\Big] \\
&\overset{\text{Jensen+linearity}}{\leq} \int_0^1 \mathbb{E}\Big[\exp\Big(\lambda\frac{\pi}{2}\langle\nabla F(\boldsymbol{x}_t),\boldsymbol{y}_t\rangle\Big)\Big]dt \\
&\overset{\text{Lem. 2.9}}{\leq} \int_0^1 \exp\Big(\frac{1}{2}\cdot\Big(\lambda\frac{\pi}{2}\Big)^2\Big)dt \leq \exp(\frac{\pi^2}{8}\lambda^2).
\end{aligned}$$

Here we use that $(\boldsymbol{x}_t,\boldsymbol{y}_t)$ are independent and $\nabla F(\boldsymbol{x}_t)$ is a vector of length $\|\nabla F(\boldsymbol{x}_t)\|_2 \leq 1$. □

We want to give a visualization for the proof at least for dimension $n=1$. We generate two independent Gaussians $\boldsymbol{x}_0,\boldsymbol{x}_1 \sim N(0,1)$ by picking two orthogonal unit vectors $\boldsymbol{e}_0,\boldsymbol{e}_1 \in \mathbb{R}^2$, drawing a Gaussian $\boldsymbol{g} \sim N^2(0,1)$ and setting $\boldsymbol{x}_t := \langle\boldsymbol{g},\boldsymbol{e}_t\rangle$ for $t\in\{0,1\}$. We also need to be able to interpolate between both random variables. Hence for $0\leq t\leq 1$ we define $\boldsymbol{e}_t := \cos(t\cdot\frac{\pi}{2})\cdot\boldsymbol{e}_1+\sin(t\cdot\frac{\pi}{2})\cdot\boldsymbol{e}_0$. Then $\|\boldsymbol{e}_t\|_2=1$ for all t. We also define another unit vector $\boldsymbol{d}_t := -\sin(t\cdot\frac{\pi}{2})\cdot\boldsymbol{e}_0+\cos(t\cdot\frac{\pi}{2})\cdot\boldsymbol{e}_1$. Note that again $\|\boldsymbol{d}_t\|_2=1$ and $\boldsymbol{e}_t\perp\boldsymbol{d}_t$ for all t.



Then letting $\boldsymbol{x}_t := \langle\boldsymbol{g},\boldsymbol{e}_t\rangle$ and $\boldsymbol{y}_t := \langle\boldsymbol{g},\boldsymbol{d}_t\rangle$ gives two independent Gaussians. This is used for evaluating $F(\boldsymbol{x}_1)-F(\boldsymbol{x}_0)=\frac{\pi}{2}\int_0^1\langle\nabla F(\boldsymbol{x}_t),\boldsymbol{y}_t\rangle\,dt$ as the *position* $\boldsymbol{x}_t$ is independent from the *direction* $\boldsymbol{y}_t$.

Proof of main Theorem. We will prove the bound with weaker constants to keep things simple. With out loss of generality we may assume that $\mathbb{E}_{\boldsymbol{x}\sim N^n(0,1)}[F(\boldsymbol{x})]=0$. On the one hand, we know that

$$\mathbb{E}[\exp(\lambda F(\boldsymbol{x}))] \geq e^{\lambda\cdot 10\lambda}\cdot\Pr[F(\boldsymbol{x})>10\lambda]$$

On the other hand if we draw $\boldsymbol{x},\boldsymbol{y}\sim N^n(0,1)$ independently, then

$$\begin{aligned}
10e^{8\lambda^2/2} &\overset{\text{previous Lemma}}{\geq} \mathbb{E}[\exp(\lambda(F(\boldsymbol{x})-F(\boldsymbol{y})))] \overset{\text{indep.}}{=} \mathbb{E}[\exp(\lambda F(\boldsymbol{x}))]\cdot\mathbb{E}[\exp(-\lambda F(\boldsymbol{y}))] \\
&\overset{\text{Jensen ineq}}{\geq} \mathbb{E}[\exp[\lambda F(\boldsymbol{x})]\cdot\underbrace{\exp(\underbrace{\mathbb{E}[-\lambda F(\boldsymbol{y})]}_{=0})}_{=1} = \mathbb{E}[\exp(\lambda F(\boldsymbol{x}))]
\end{aligned}$$

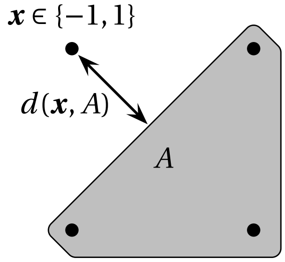
Putting both together, we get

$$\Pr[F(\boldsymbol{x}) > 10\lambda] \leq \frac{\mathbb{E}[\exp(\lambda F(\boldsymbol{x}))]}{e^{10\lambda^2}} \leq \frac{10e^{8\lambda^2}}{e^{10\lambda^2}} \leq 10e^{-2\lambda^2}$$

□

2.4 Talagrand inequality

The *Talagrand inequality* is a remarkably strong concentration inequality for *product measures*. For example if A is a convex set containing half of hypercube points $\{-1,1\}^n$, then Talagrand's inequality tells us that 99.99% of points have Euclidean distance $O(1)$ to A. For this chapter, we follow an excellent blog post of Tao[2].



In the following, we abbreviate $d(\boldsymbol{x},A) := \min\{\|\boldsymbol{x}-\boldsymbol{y}\|_2 \mid \boldsymbol{y} \in A\}$ as the *distance* of $\boldsymbol{x}$ to the set A. Recall that a *median* of a real-valued random variable Y is a value M with $\Pr[Y \leq M] \geq \frac{1}{2}$ and $\Pr[Y \geq M] \geq \frac{1}{2}$. Also recall that there can be an interval of median's for a random variable.

Theorem 2.12. *Let $A \subseteq \mathbb{R}^n$ be a convex set with $A \cap \{\pm1\}^n \neq \emptyset$ and let $\boldsymbol{X} \sim \{\pm1\}^n$ be a uniform vertex of the hypercube. For M being a median of $d(\boldsymbol{X},A)$ and $t \geq 0$ we have*

$$\Pr[d(\boldsymbol{X},A) > M+t] \leq 4\exp\left(-\frac{t^2}{16}\right)$$

Before we come to the proof, we want to recall Hölder's inequality that can be phrased as follows:

Lemma 2.13 (Hölder's Inequality)**.** *Let $X,Y \in \mathbb{R}_{\geq 0}$ be jointly distributed non-negative random variables. For $0 \leq \lambda \leq 1$ one has*

$$\mathbb{E}[X^{1-\lambda} \cdot Y^{\lambda}] \leq \mathbb{E}[X]^{1-\lambda} \cdot \mathbb{E}[Y]^{\lambda}.$$

[2] See https://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/

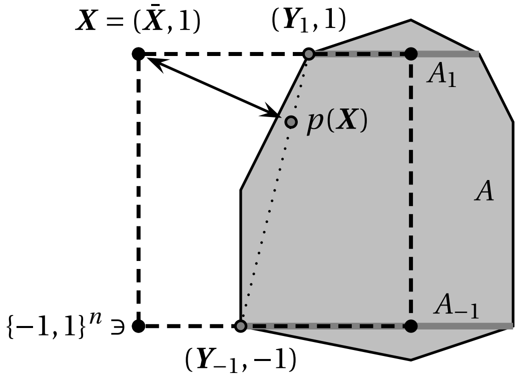
To get some intuition, we can rephrase the inequality. W.l.o.g. the distributions X and Y just pick a uniform coordinate in $\boldsymbol{a} \in \mathbb{R}^n_{\geq 0}$ and $\boldsymbol{b} \in \mathbb{R}^n_{\geq 0}$. Then the inequality says that $\sum_{i=1}^n a_i^{1-\lambda} b_i^{\lambda} \leq (\sum_{i=1}^n a_i)^{1-\lambda} (\sum_{i=1}^n b_i)^{\lambda}$. In particular for $\lambda = \frac{1}{2}$, this is just the *Cauchy-Schwartz inequality*.

Similar to previous concentration inequalities, it suffices to give a bound on the exponential moment[3].

Lemma 2.14. *Let $A \subseteq \mathbb{R}^n$ be convex with $A \cap \{\pm 1\}^n \neq \emptyset$ and abbreviate $\mu_n(A) := \Pr[\boldsymbol{X} \in A]$ where $\boldsymbol{X} \sim \{\pm 1\}^n$ is drawn uniformly. Then for a universal constant $c > 0$,*

$$\mathbb{E}\left[\exp\left(c \cdot d(\boldsymbol{X}, A)^2\right)\right] \leq \frac{1}{\mu_n(A)}.$$

We prove the claim via induction over n. We write $\boldsymbol{X} = (\bar{\boldsymbol{X}}, X_n)$, where $\bar{\boldsymbol{X}} \in \{\pm 1\}^{n-1}$ are the first $n-1$ coordinates. For $t \in \{-1, 1\}$ we consider the two convex *slices* $A_t := \{\bar{\boldsymbol{x}} \in \mathbb{R}^{n-1} \mid (\bar{\boldsymbol{x}}, t) \in A\}$. Let $\boldsymbol{Y}_t \in A_t$ be the closest point to $\bar{\boldsymbol{X}}$ in the slice A_t. The trick is that we can bound the distance of $\boldsymbol{X}$ to A by the distance to any convex combination of the points $(\boldsymbol{Y}_1, 1)$ and $(\boldsymbol{Y}_{-1}, -1)$. Let us abbreviate the point $p(\boldsymbol{X}) := (1-\lambda) \cdot (\boldsymbol{Y}_{X_n}, X_n) + \lambda \cdot (\boldsymbol{Y}_{-X_n}, -X_n)$ which by convexity lies in A. Here $0 \leq \lambda \leq 1$ is a parameter that we determine later. Crucially it is allowed that λ depends on the outcome of X_n.



visualization for $X_n = 1$

[3] For the sake of completeness here the argument how to complete Theorem 2.12. First of all one can make the Lemma work with $c = \frac{1}{16}$. Let $A' := \{\boldsymbol{x} \in \mathbb{R}^n \mid d(\boldsymbol{x}, A) \leq M\}$ which is a convex set as well. By definition of the median $\mu_n(A') = \frac{1}{2}$. Then $\Pr[d(\boldsymbol{X}, A) > M + t] = \Pr[d(\boldsymbol{X}, A') > t] = \Pr[\exp(\frac{1}{16} d(\boldsymbol{X}, A)^2) > \exp(\frac{1}{16} t^2)] \leq \frac{\mathbb{E}[\frac{1}{16} d(\boldsymbol{X}, A)^2]}{\exp(\frac{1}{16} t^2)}) \leq 2\exp(-\frac{1}{16} t^2)$ by Markov's inequality

First we get a useful bound on the distance

$$\begin{aligned} d(\boldsymbol{X},A)^2 &\overset{p(\boldsymbol{X})\in A}{\leq} \left\| p(\boldsymbol{X}) - \boldsymbol{X} \right\|_2^2 \\ &= \left\| (1-\lambda)\begin{pmatrix} \boldsymbol{Y}_{X_n} \\ X_n \end{pmatrix} + \lambda \begin{pmatrix} \boldsymbol{Y}_{-X_n} \\ -X_n \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{X}} \\ X_n \end{pmatrix} \right\|_2^2 \\ &\overset{\text{Phytagoras}}{=} \left\| (1-\lambda)(\boldsymbol{Y}_{X_n} - \bar{\boldsymbol{X}}) + \lambda(\boldsymbol{Y}_{-X_n} - \bar{\boldsymbol{X}}) \right\|_2^2 + \underbrace{(X_n \cdot ((1-\lambda)-\lambda-1))^2}_{=4\lambda^2} \\ &\overset{\|\cdot\|_2^2 \text{ convex}}{\leq} (1-\lambda)\cdot \|\boldsymbol{Y}_{X_n} - \bar{\boldsymbol{X}}\|_2^2 + \lambda \cdot \|\boldsymbol{Y}_{-X_n} - \bar{\boldsymbol{X}}\|_2^2 + 4\lambda^2 \end{aligned}$$

Note the asymmetry as A_{X_n} is the "same side" as $\boldsymbol{X}$ and A_{-X_n} is the "opposite side". Now we apply $\mathbb{E}[\exp(c\cdot ..)]$ to both sides of the equation and get

$$\begin{aligned} & \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\exp(-cd(\boldsymbol{X},A)^2)\right] \\ \leq\; & \mathop{\mathbb{E}}_{\boldsymbol{X}}\left[\exp\left(c\big((1-\lambda)d(\bar{\boldsymbol{X}},A_{X_n})^2 + \lambda\cdot d(\bar{\boldsymbol{X}},A_{-X_n})^2 + 4\lambda^2\big)\right)\right] \\ =\; & \mathop{\mathbb{E}}_{X_n}\left[e^{4c\lambda^2} \mathop{\mathbb{E}}_{\bar{\boldsymbol{X}}}\left[\exp\left(c\cdot d(\bar{\boldsymbol{X}},A_{X_n})^2\right)^{1-\lambda} \cdot \exp\left(c\cdot d(\bar{\boldsymbol{X}},A_{-X_n})^2\right)^{\lambda}\right]\right] \\ \overset{\text{Hölder}}{\leq}\; & \mathop{\mathbb{E}}_{X_n}\left[e^{4c\lambda^2} \mathop{\mathbb{E}}_{\bar{\boldsymbol{X}}}\left[\exp(c\cdot d(\bar{\boldsymbol{X}},A_{X_n})^2)\right]^{1-\lambda} \mathop{\mathbb{E}}_{\bar{\boldsymbol{X}}}\left[\exp(c\cdot d(\bar{\boldsymbol{X}},A_{-X_n})^2)\right]^{\lambda}\right] \\ \overset{\text{induction}}{\leq}\; & \mathop{\mathbb{E}}_{X_n}\left[e^{4c\lambda^2} \frac{1}{\mu_{n-1}(A_{X_n})^{1-\lambda}} \cdot \frac{1}{\mu_{n-1}(A_{-X_n})^{\lambda}}\right] \\ =\; & \underbrace{\mathop{\mathbb{E}}_{X_n}\left[e^{4c\lambda^2} \frac{1}{(1+X_nq)^{1-\lambda}\cdot(1-X_nq)^{\lambda}}\right]}_{=:(*)} \cdot \frac{1}{\mu_n(A)} \end{aligned}$$

where in the last step we write $\mu_{n-1}(A_1) = (1+q)\mu_n(A)$ and $\mu_{n-1}(A_{-1}) = (1-q)\cdot\mu_n(A)$. Let us assume $0 \leq q \leq 1$ for the sake of symmetry. Then we can continue bounding $(*)$ making in particular use of the fact that λ is allowed to depend on X_n. We will need to distinguish two cases:

- *Case* $q \geq 4c$. In this case $A_1 \cap \{\pm 1\}^{n-1}$ has a good fraction more points than $A_{-1} \cap \{\pm 1\}^{n-1}$ and a good choice for λ will be to always be on the side of A_1. We can then estimate

$$(*) \overset{\substack{X_n=1\Rightarrow\lambda=0,\\ X_n=-1\Rightarrow\lambda=1}}{=} \frac{1}{2}\left(\frac{1}{1+q} + \frac{e^{4c}}{1+q}\right) \overset{q\geq 4c}{\leq} \frac{1+e^{4c}}{2\cdot(1+4c)} \overset{0\leq c\leq 0.3}{\leq} 1$$

  as can be easily checked.

- *Case* $0 \leq q < 4c$. In this case we can use $\frac{1}{1+x} \leq \exp(-x+x^2)$ for $|x| \leq 1/4$ to simplify to

$$\begin{aligned}
(*) \quad &\leq \quad \mathop{\mathbb{E}}_{X_n}\left[e^{4c\lambda^2}(e^{-X_nq+q^2})^{1-\lambda}\cdot(e^{X_nq+q^2})^{\lambda}\right] \\
&= \quad \mathop{\mathbb{E}}_{X_n}\left[\exp(4c\lambda^2+\lambda\cdot 2X_nq - X_nq+q^2)\right] \\
&\overset{\substack{X_n=1\Rightarrow\lambda=0,\\ X_n=-1\Rightarrow\lambda=\frac{q}{4c}}}{=} \quad \frac{1}{2}\Big(\exp(-q+q^2)+\exp\Big(\underbrace{-\frac{q^2}{4c}+q^2}_{\leq -3q^2 \text{ as } c\leq\frac{1}{16}}+q\Big)\Big) \overset{0\leq q\leq 1}{<} 1
\end{aligned}$$

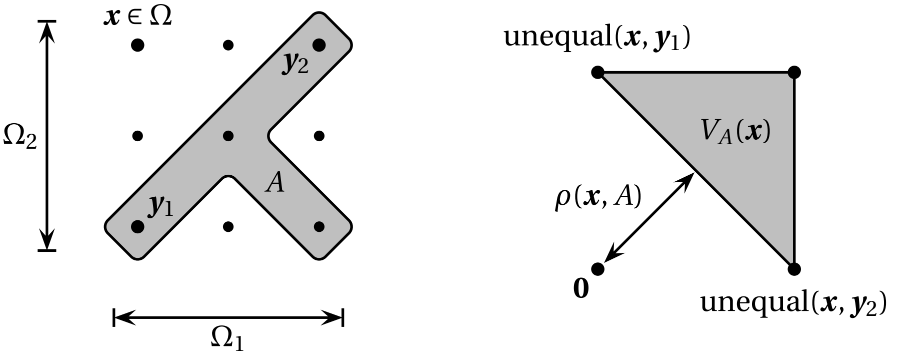
The interpretation of this parameter choice is that if $X_n = 1$, then $\boldsymbol{X}$ lies on the side of the bigger set A_1 we pick $p(\boldsymbol{X})$ should be in A_1; only if $X_n = -1$, then $p(\boldsymbol{X})$ will be a true convex combination.

2.4.1 The general form of Talagrand's inequality

In fact, Talagrand proved his inequality in a much more general form that we want to outline at least. Let $\Omega_1, \ldots, \Omega_n$ be some measurable sets and let $\mathcal{D}$ be a *product measure* on $\Omega := \Omega_1 \times \ldots \times \Omega_n$. Fix a set $A \subseteq \Omega$ (that does not need to be convex). For $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ we define $\text{unequal}(\boldsymbol{x}, \boldsymbol{y}) := \{i \in [n] \mid x_i \neq y_i\}$ as the indices where $\boldsymbol{x}$ and $\boldsymbol{y}$ differ. Recall that for a vector $\boldsymbol{s} \in \mathbb{R}^n$, we have $\text{supp}(\boldsymbol{s}) := \{i \in [n] \mid s_i \neq 0\}$. For each $\boldsymbol{x} \in \Omega$ we define

$$V_A(\boldsymbol{x}) := \text{conv}\{\boldsymbol{s} \in \{0,1\}^n \mid \exists \boldsymbol{y} \in A : \text{unequal}(\boldsymbol{x}, \boldsymbol{y}) \subseteq \text{supp}(\boldsymbol{s})\}$$

Then we define a distance $\rho(\boldsymbol{x}, A) := \min\{\|\boldsymbol{s}\|_2 \mid \boldsymbol{s} \in V_A(\boldsymbol{x})\}$. Intuitively, a small distance $\rho(\boldsymbol{x}, A)$ means that there is a convex combination of *paths* from $\boldsymbol{x}$ to A where on each path one has to change only few coordinates.



$\boldsymbol{x} \in \Omega$
$\boldsymbol{y}_2$
Ω_2
A
$\boldsymbol{y}_1$
Ω_1
$\text{unequal}(\boldsymbol{x}, \boldsymbol{y}_1)$
$V_A(\boldsymbol{x})$
$\rho(\boldsymbol{x}, A)$
$\boldsymbol{0}$
$\text{unequal}(\boldsymbol{x}, \boldsymbol{y}_2)$

Example with $A \subseteq \Omega = \Omega_1 \times \Omega_2$ where $|\Omega_1| = |\Omega_2| = 3$.

Then the concentration inequality is as follows:

Theorem 2.15 (General Talagrand Inequality)**.** *Let $\mathcal{D}$ be a product measure on $\Omega = \Omega_1 \times \ldots \times \Omega_n$. Then for $A \subseteq \Omega$ and $t \geq 0$ one has*

$$\Pr_{\boldsymbol{x}\sim\mathcal{D}}[\rho(\boldsymbol{x}, A) > t] \leq \frac{4\exp(-t^2/16)}{\Pr_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x} \in A]}$$

For a proof of this more general result we refer to [AS16]. A different way of phrasing Talagrand's result is the following:

Corollary 2.16. *Let μ be a product measure on $\Omega = \Omega_1 \times \ldots \times \Omega_n$ and fix a set $A \subseteq \Omega$. For $t \geq 0$, define*

$$A_t := \Big\{\boldsymbol{x} \in \Omega \mid \exists \textit{distribution } \mathcal{D}(\boldsymbol{x}) \textit{ on } A \textit{ so that } \sqrt{\sum_{i=1}^{n} \Pr_{\boldsymbol{y}\sim\mathcal{D}(\boldsymbol{x})}[y_i \neq x_i]^2} \leq t\Big\}$$

Then $\mu(A_t) \geq 1 - \frac{4\exp(-t^2/16)}{\mu(A)}$.

In order to demonstrate the power of Talagrand's general inequality, we show one application. We want to emphasize that the distribution μ does not have to be identical for every coordinate. Also there is nothing special about the interval $[-1,1]$ — any interval works, but the length of the interval goes into the bound.

Lemma 2.17. *Let μ be any product measure on $[-1,1]^n$ and let $f : [-1,1]^n \to \mathbb{R}$ be convex and 1-Lipschitz. Let median(f) be a value with $\Pr_{\boldsymbol{x}\sim\mu}[f(\boldsymbol{x}) \leq \textit{median}(f)] \geq 1/2$ and $\Pr_{\boldsymbol{x}\sim\mu}[f(\boldsymbol{x}) \geq \textit{median}(f)] \geq 1/2$. Then*

$$\Pr_{\boldsymbol{x}\sim\mu}[f(\boldsymbol{x}) > \textit{median}(f) + 2t] \leq 8\exp(-t^2/16)$$

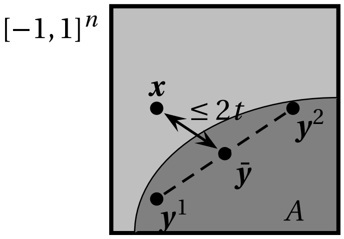
Proof. Let $A := \{\boldsymbol{x} \in [-1,1]^n \mid f(\boldsymbol{x}) \leq \text{median}(f)\}$ so that $\mu(A) \geq 1/2$ and define A_t as before in Cor 2.16. Then by Cor 2.16 we know that $\mu(A_t) \geq 1 - 8\exp(-t^2/16)$. The following claim will then finish the proof:

Claim. Every $\boldsymbol{x} \in A_t$ has $f(\boldsymbol{x}) \leq \text{median}(f) + 2t$.

Proof of claim. By definition of A_t, there are $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^k \in A$ and convex coefficients $\lambda_1, \ldots, \lambda_k \geq 0$ with $\sum_{j=1}^k \lambda_j = 1$ so that the condition from Cor 2.16 is satisfied. By a slight abuse of notation, let us write $\boldsymbol{y} \sim \lambda$ if $\boldsymbol{y} = \boldsymbol{y}^j$ with probability λ_j. Next, consider $p_i := \Pr_{\boldsymbol{y}\sim\lambda}[x_i \neq y_i]$ be the probability that the ith coordinate of $\boldsymbol{y} \sim \lambda$ differs from $\boldsymbol{x}$. Then the condition of Cor 2.16 is that

$$\|\boldsymbol{p}\|_2 = \sqrt{\sum_{i=1}^{n} \Big(\sum_{j=1}^{k} \lambda_j \cdot \mathbb{1}_{y_i^j \neq x_i}\Big)^2} \leq t$$

Now, consider the average $\bar{\boldsymbol{y}} := \sum_{j=1}^{k} \lambda_j \boldsymbol{y}^j = \mathbb{E}_{\boldsymbol{y}\sim\lambda}[\boldsymbol{y}]$ of the obtained points.



By convexity

$$f(\bar{\boldsymbol{y}}) \leq \sum_{j=1}^{k} \lambda_j f(\boldsymbol{y}^j) \leq \text{median}(f)$$

Observe that $|x_i - \bar{y}_i| \leq 2p_i$ since all the points are in $[-1,1]^n$, so

$$\|\boldsymbol{x} - \bar{\boldsymbol{y}}\|_2 = \Big(\sum_{i=1}^{n} |x_i - \bar{y}_i|^2\Big)^{1/2} \leq 2\Big(\sum_{i=1}^{n} p_i^2\Big)^{1/2} \leq 2t$$

Then as the function f is 1-Lipschitz we get $f(\boldsymbol{x}) \leq f(\bar{\boldsymbol{y}}) + 2t \leq \text{median}(f) + 2t$ which gives the claim. □

2.5 Exercises

Exercise 2.1. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with $|A_{ij}| \leq 1$ for all $i, j = 1, \ldots, n$. Give a deterministic polynomial time algorithm to find an $\boldsymbol{x} \in \{-1,1\}^n$ so that $\boldsymbol{A}_i\boldsymbol{x} \leq O(\sqrt{n}\ln(n))$ for all rows i.
Hint: Consider the potential function $\Phi_k(\boldsymbol{x}) := \sum_{i=1}^{n} \exp(\lambda \sum_{j=1}^{k} A_{ij}x_j - 2\lambda^2 k)$ where $k \in \{0,\ldots,n\}$. Show that there is a deterministic polynomial time algorithm to find an $\boldsymbol{x} \in \{-1,1\}^n$ with $\Phi_n(\boldsymbol{x}) \leq n$.

Exercise 2.2 (Balls into bins)**.** Suppose we have n balls and n bins. We throw the balls so that each ball ends up in a uniform random bin. Show that with high probability (say $\geq 1-\frac{1}{n}$) the maximum number of balls in any bin does not exceed $O(\log(n)/\log\log(n))$. Show also that with high probability there is a bin with at least $\Omega(\log(n)/\log\log n)$ many balls.

Exercise 2.3. Consider a random graph $G = ([n], E)$ which contains each edge with probability 1/2. Let X be the random variable that gives the number of triangles in G. Prove that $\Pr[|X - \mathbb{E}[X]| > \lambda \cdot n^2] \leq 2\exp(-\Theta(\lambda^2))$ for all $\lambda \geq 0$.

Exercise 2.4. Derive Theorem 2.12 from Theorem 2.15 (possibly with different constants).

Exercise 2.5. Let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_m$ be unit vectors. Draw a random Gaussian $\boldsymbol{x}$ and consider the random variable $Y := \max\{\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle \mid i = 1,\dots,m\}$. Show that $\Pr[Y > \text{median}(Y) + t] \leq \exp(-\Omega(t^2))$.

Exercise 2.6. Let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_m$ be unit vectors. Draw $\boldsymbol{x} \sim [-1,1]^n$ uniformly at random and consider the random variable $Y := \max\{\langle \boldsymbol{u}_i, \boldsymbol{x}\rangle \mid i = 1,\dots,m\}$. Show that $\Pr[Y > \text{median}(Y) + 2t] \leq 4\exp(-t^2/16)$.

Exercise 2.7. For a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we denote define a function $f : [-1,1]^n \to \mathbb{R}$ with $f(\boldsymbol{M}) := \max\{\langle \boldsymbol{M}, \boldsymbol{x}\boldsymbol{x}^T\rangle : \boldsymbol{x} \in \mathbb{R}^n \text{ with } \|\boldsymbol{x}\|_2 = 1\}$. Here, for two matrices $\boldsymbol{M}, \boldsymbol{N}$ we let $\langle \boldsymbol{M}, \boldsymbol{N}\rangle := \sum_{i=1}^n \sum_{j=1}^n M_{ij}N_{ij}$ be the Frobenius inner product. Let $\mathcal{D}$ be a product distribution that picks each entry M_{ij} independently from some distribution over $[-1,1]$. Let m be a median of $f(\boldsymbol{M})$ if $\boldsymbol{M}$ is drawn from $\mathcal{D}$. Prove that

$$\Pr[f(\boldsymbol{M}) > m + 2t] \leq 8\exp(-t^2/16)$$

for all $t \geq 0$.

Hint. Show that the function f is 1-Lipschitz if you use the Frobenius norm $\|\boldsymbol{M}\|_2 := (\sum_{i=1}^n \sum_{j=1}^n M_{ij}^2)^{1/2}$. Is the function f convex? Then use a result from the lecture.

Exercise 2.8. Prove that for infinitely many $n \in \mathbb{N}$ there is a random variable $X := X_1 + \ldots + X_n$ with $\mathbb{E}[X_j] = 0$ and $|X_j| \leq 1$ for all $j \in \{1,\dots,n\}$ and $\text{Cov}[X_j, X_{j'}] = 0$ for $j \neq j'$ while $\Pr[X = n] \geq \Omega(\frac{1}{n})$.

Hint. You may use without proof that there exists a matrix $\boldsymbol{A} \in \{-1,1\}^{2n\times n}$ so that (i) $\sum_{i=1}^{2n} \boldsymbol{A}_i = \boldsymbol{0}$ and (ii) $\langle \boldsymbol{A}^j, \boldsymbol{A}^{j'}\rangle = 0$ for $j \neq j'$ and (iii) $\boldsymbol{A}_1$ is the all-ones-vector. Here $\boldsymbol{A}_i$ is the ith row and $\boldsymbol{A}^j$ is the jth column. Note that such a matrix can be obtained by taking a Hadamard matrix $\boldsymbol{H}$ and letting $\boldsymbol{A} := \begin{pmatrix} \boldsymbol{H} \\ -\boldsymbol{H} \end{pmatrix}$.

Remark. The exercise shows that Chebychev's inequality is tight and one cannot derive better concentration only based on 1st and 2nd moment.

Exercise 2.9. Let $f : \{-1,1\}^n \to \mathbb{R}$ be a function on vertices of the hypercube and suppose that

$$|f(\boldsymbol{y}) - f(y_1,\dots,y_{i-1},-y_i,y_{i+1},\dots,y_n)| \leq a_i \quad \forall \boldsymbol{y} \in \{-1,1\}^n,$$

meaning that changing the ith coordinate changes the function value by at most a_i. Let $Y \sim \{-1,1\}^n$ be a uniform random vertex. Prove that for $\lambda \geq 0$ one has

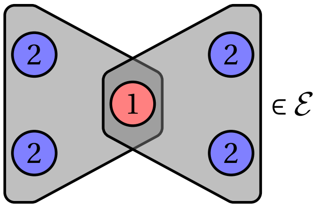
$$\Pr_Y[|f(Y) - \mathbb{E}[f(Y)]| > \lambda\|\boldsymbol{a}\|_2] \leq 2\exp(-\lambda^2/4).$$

Hint. Use may use the following variant of Azuma's inequality without a proof: Let $X_0, \ldots, X_n$ be a Martingale with $|X_t - X_{t-1}| \leq a_t$ for $t = 1, \ldots, n$, then for any $\lambda \geq 0$ one has $\Pr[|X_n - X_0| > \lambda \|\boldsymbol{a}\|_2] \leq 2\exp(-\lambda^2/4)$.

Chapter 3

The Lovász Local Lemma

We want to motivate the next chapter with an application. Suppose we have a *hypergraph* $H = (V, \mathcal{E})$ with vertices V and hyperedges $e \subseteq V$ and suppose that that H is *k-uniform*, which means that $|e| = k$ for all $e \in \mathcal{E}$. A *coloring* is a map $\chi : V \to \{\text{red}, \text{blue}\}$ that gives each vertex a color. Using terminology that goes back to Erdős we say that the hypergraph has *Property B*, if there is a coloring that leaves no edge *monochromatic* (meaning that every edge sees both colors).



Example of a 3-regular hypergraph and a 2-coloring satisfying Property B

It is not hard to see that for large enough k there is always a proper coloring:

Lemma 3.1. *A k-uniform hypergraph with $k \geq \log_2(4|\mathcal{E}|)$ has property B.*

Proof. Take a uniform random coloring χ. Then for each edge $e \in \mathcal{E}$ one has

$$\Pr_{\chi}[e \text{ is monochromatic}] = 2 \cdot 2^{-k} \leq \frac{1}{2|\mathcal{E}|}$$

Then the union bound over all hyperedges gives the claim. □

On the other hand, this bound does not leave too much room for improvement. For example if the hypergraph $H = (V, \mathcal{E})$ has $|V| = 2k$ nodes and contains *all* subsets of k vertices as edges, then $|\mathcal{E}| = \binom{2k}{k} = 2^{\Theta(k)}$. However, any coloring will leave some edge monochromatic.

This discussion brings us to an intermediate case that is less clear. For a vertex v, we define the *degree* $\deg_H(v) := |\{e \in \mathcal{E} \mid v \in \mathcal{E}\}|$ as the number of edges containing v. Suppose that we again have a k-regular hypergraph but instead of a bound on the size $|\mathcal{E}|$ we have a bound on the maximum degree. Let us say for the sake of concreteness that $\deg_H(v) \leq 2^{k/2}$. Is it then possible to find a coloring that leaves no edge monochromatic? After a bit of trying one can see that all the conventional techniques to approach this problem are failing. And indeed it required a beautiful mathematical theorem to resolve the problem.

3.1 The original Local Lemma

Now we show the *Lovász Local Lemma* (which actually appeared in the joint paper by Erdős and Lovász [EL75]). The presentation we give here is inspired from Mitzenmacher and Upfal [MU17]. We also recommend Chapter 5 in [AS16]. Let $G_1, \dots, G_n$ be events ("good" events that we like to be true). A graph $H = ([n], E)$ is called a *dependency graph* if

$$\Pr[G_i] = \Pr\Big[G_i \mid \Big(\bigcap_{j \in I_0} \bar{G}_j\Big) \cap \Big(\bigcap_{j \in I_1} G_j\Big)\Big] \quad \forall i \in [n]\ \forall I_0 \dot{\cup} I_1 \subseteq [n](\backslash N(i) \cup \{i\})$$

Phrased differently, each event G_i needs to be independent from all it's non-neighbors[1]. This is even a bit more strict than what is needed, but intuitive. However, it is important to remember that pairwise independence is not enough.

Theorem 3.2 (Symmetric version of the Lovász Local Lemma)**.** *Let $G_1, \dots, G_n$ be a set of "good" events with $\Pr[\bar{G}_i] \leq p$. Suppose that each G_i is independent of all but at most d events and $d \cdot p \leq \frac{1}{4}$. Then $\Pr[\bigcap_{i=1}^n G_i] > 0$.*

Proof. For a subset $S \subseteq [n]$ of events, we denote $G(S) := \bigcap_{i \in S} G(i)$ as the case that all those good events happen at the same time. We need to prove that given our assumptions, $\Pr[G([n])] > 0$. The main ingredient is to prove that conditioning on any subset of good events does not more than double the probability of any

[1] **Remark 1.** It would have been more intuitive to require an edge $\{i, j\}$ in the dependency graph if the events (G_i, G_j) are dependent. But that is not sufficient. For example consider the complete graph $K_n = ([n], E)$ and pick a uniform coloring $\chi : [n] \to \{-1, 1\}$. For $e = \{u, v\} \in E$, consider the event $G_e :=$ "$\chi(u) \neq \chi(v)''$. Then one can check that for any pair of distinct edges $e, e' \in E$ one has $\Pr[G_e] = \Pr[G_e \mid G_{e'}] = \frac{1}{2}$, hence any *pair* of events is independent. Note that $\Pr[\bigcap_{e \in E} G_e] = 0$ for this construction.

Remark 2. It is not true that there is always a unique minimal dependency graph. Consider the example from Remark 1 for $n = 3$ and say that $E = \{e_1, e_2, e_3\}$ are the all edges in K_3. Then any two edges $F \subseteq \{e_1, e_2, e_3\}$ form a valid dependency graph — but one edge alone is not sufficient.

bad event.

Claim. *For any* $S \subseteq [n]$ *with* $\Pr[G(S)] > 0$ *and* $i \in [n] \setminus S$ *one has* $\Pr[\bar{G}_i \mid G(S)] \leq 2p$.

Proof of claim. We show the claim by induction on $|S|$. If $|S| = 0$, then $\Pr[\bar{G}_i \mid G(S)] = \Pr[\bar{G}_i] \leq p$ by assumption. Now we split S into the events that are independent of i and the ones that are dependent. More precisely, we write $S = S_1 \dot{\cup} S_2$ so that $S_1 := \{j \mid \{i,j\} \in E\}$ are the neighbors and $S_2 := \{j \in S \mid \{i,j\} \notin E\}$ are the non-neighbors in the dependency graph. Then

$$\begin{aligned}
\Pr[\bar{G}_i \mid G(S)] &\overset{S=S_1\dot{\cup}S_2}{=} \Pr[\bar{G}_i \mid G(S_1) \cap G(S_2)] \overset{\text{cond. prob.}}{=} \frac{\Pr[\bar{G}_i \cap G(S_1) \cap G(S_2)]}{\Pr[G(S_1) \cap G(S_2)]} \\
&\overset{\text{cond. prob.}}{=} \frac{\Pr[\bar{G}_i \cap G(S_1) \cap G(S_2) \mid G(S_2)] \cdot \Pr[G(S_2)]}{\Pr[G(S_1) \cap G(S_2) \mid G(S_2)] \cdot \Pr[G(S_2)]} \\
&\overset{\Pr[A\cap B|B]=\Pr[A|B]}{=} \frac{\Pr[\bar{G}_i \cap G(S_1) \mid G(S_2)]}{\Pr[G(S_1) \mid G(S_2)]} \\
&\overset{\Pr[A\cap B]\leq\Pr[A]}{\leq} \frac{\overbrace{\Pr[\bar{G}_i \mid G(S_2)]}^{\leq p \text{ by indep.}}}{\underbrace{\Pr[G(S_1) \mid G(S_2)]}_{\geq 1/2 \text{ by } (*)}} \leq 2p
\end{aligned}$$

Note that we have implicitly used that $\Pr[G(S_2)] > 0$ so that the cancellation is well defined. It remains to argue why $(*)$ is true. If $S_1 = \emptyset$, then $\Pr[G(S_1) \mid G(S_2)] = 1$. So suppose that $|S_1| > 0$ and hence $|S_2| < |S|$. Then we are allowed to apply induction to get that

$$\Pr[G(S_1) \mid G(S_2)] = 1 - \Pr\Big[\bigcup_{j \in S_1} \bar{G}_j \mid G(S_2)\Big] \overset{\text{union bound}}{\geq} 1 - \sum_{j \in S_1} \underbrace{\Pr[\bar{G}_j \mid G(S_2)]}_{\leq 2p \text{ by induction}} \geq 1 - 2p \cdot \underbrace{|S_1|}_{\leq d} \geq \frac{1}{2} \quad \square$$

Using the proven claim, we can quickly conclude that

$$\Pr[G([n])] = \prod_{i=1}^{n} \underbrace{\Pr[G(i) \mid G(\{1,\ldots,i-1\})]}_{\geq 1-2p \text{ by claim}} \geq (1-2p)^n > 0$$

□

Note that the condition $d \cdot p \leq \frac{1}{4}$ can be sharpened to $p \cdot (d+1) \leq \frac{1}{e}$, see e.g. [AS16]. Shearer proved that this bound is tight in the sense that for any $\varepsilon > 0$, there is a probability space with $p \cdot (d+1) \cdot (e+\varepsilon) \leq 1$ where $\Pr[G([n])] = 0$.

A more general version. There is also a more general version of the Lovász Local Lemma in which the probabilities of the events do not have to be bounded by a uniform parameter p. We state the claim for the sake of completeness. The proof is essentially identical to the one of the symmetric case, only the notation needs to be adapted. Again, see e.g. [AS16] for details.

Theorem 3.3 (General Version of the Lovász Local Lemma)**.** *Let $G_1,\dots,G_n$ be a set of "good" events and let $G = ([n],E)$ be a dependence graph. Suppose that there are parameters $x_1,\dots,x_n \in]0,1[$ with*

$$\Pr[\bar{G}(i)] \leq x_i \prod_{j:\{i,j\}\in E} (1-x_j)$$

Then $\Pr[\bigcap_{i=1}^n G(i)] \geq \prod_{i=1}^n (1-x_i) > 0$.

For example, if one sets $x_i := 2\cdot\Pr[\bar{G}(i)]$, then the condition in the theorem is satisfied if $\sum_{j\in N(i)} \Pr[\bar{G}_j] \leq \frac{1}{4}$. In other words, the expected number of bad events happening in a neighborhood should be at most some small enough constant.

Application to hypergraphs. Finally we can resolve our question on hypergraph colorings:

Theorem 3.4. *A k-uniform hypergraph $H = (V,\mathcal{E})$ with $\deg_H(v) \leq \frac{2^k}{8k}$ for all $v \in V$ has property B.*

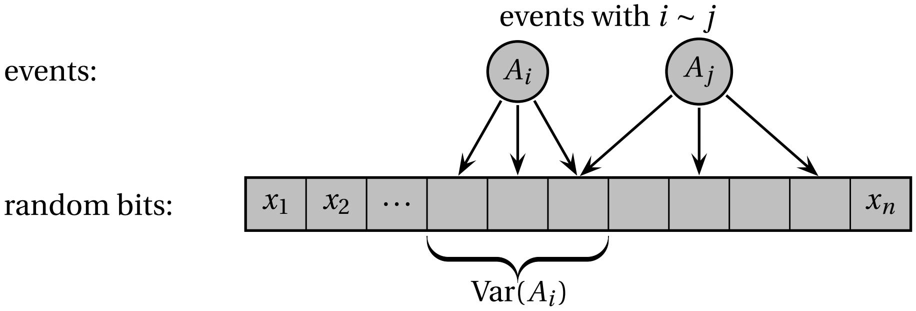
Proof. We consider the random experiment where we pick a uniform random coloring $\chi : V \to \{\text{red},\text{blue}\}$. We have the good events $G_e :=$ "e is not monochromatic$''$. Then the bad events have probability $\bar{G}_e = 2\cdot 2^{-k} =: p$. For disjoint hyperedges $e,e' \in \mathcal{E}$, the events G_e and $G_{e'}$ are independent. In particular, we can define a dependency graph $D = (\mathcal{E},E)$ with $\{e,e'\} \in E :\Leftrightarrow e\cap e' \neq \emptyset$. Observe that every fixed edge e overlaps with at most $\sum_{v\in e}(\deg_H(v)-1) \leq k\cdot\frac{2^k}{8k} = \frac{1}{8}\cdot 2^k =: d$ other edges. Then $d\cdot p = 2\cdot 2^{-k}\cdot\frac{1}{8}\cdot 2^k = \frac{1}{4}$. Hence by the Symmetric LLL the probability that all edges are not monochromatic is positive. □

3.2 An algorithmic proof

Suppose that we actually wanted to find the coloring from Theorem 3.4 in polynomial time. The success probability that for example the symmetric LLL guarantees is only $(1-2p)^n$, which is exponentially small. Hence we cannot simply re-sample uniform random colorings and hope to quickly find a satifying one.

Also the proof does not appear to be of any help other than showing a positive probability. In this section we will describe an algorithm due to Moser and Tardos [MT10] that suffices to make 99% of applications of the LLL algorithmic. The algorithm will also serve as an alternative proof for the LLL. We will not follow the original presentation of [MT10] using the notation of *Witness Trees*, but the view using *Entropy Compression*[2]. We will also be loose in the constants of the LLL condition for the sake of a cleaner exposition.

The variable model. First one has to discuss how one can model a probability space in a way that it can be handled algorithmically. We will model the "randomness source" by drawing a uniform random bit string[3] $\boldsymbol{x} \sim \{0,1\}^n$. We have *m bad events* $A_1,\dots,A_m$ that depend on the bit string $\boldsymbol{x}$. We write $A_i(\boldsymbol{x}) = 1$ if the bad event is true under string $\boldsymbol{x}$ (and $A_i(\boldsymbol{x}) = 0$ otherwise). In general the events will not depend on all random bits. Let $\text{Var}(A_i) \subseteq [n]$ be the indices of random bits that A_i depends on. The goal will be to find a string $\boldsymbol{x}$ so that $A_i(\boldsymbol{x}) = 0$ for all $i = 1,\dots,m$. Naturally we will need to assume that the probabilities $\Pr[A_i] = \Pr_{\boldsymbol{x}\sim\{0,1\}^n}[A_i(\boldsymbol{x}) = 1]$ of the bad events A_i are rather small. To model the independence relation we write $i \sim j$ if the variables of those events overlap, that means if $\text{Var}(A_i) \cap \text{Var}(A_j) \neq \emptyset$.



The algorithm. The main result will be the following:

Theorem 3.5 (Moser-Tardos Algorithm)**.** *Let $C > 0$ be a large enough constant. If $\sum_{i\in[m]:i\sim j} \Pr[A_i] \leq 2^{-C}$ for all $j \in [m]$, then there is a randomized polynomial time algorithm to find an $\boldsymbol{x} \in \{0,1\}^n$ with $A_1(\boldsymbol{x}) = \ldots = A_m(\boldsymbol{x}) = 0$.*

The algorithm is actually as simple and intuitive as one could hope for. One starts with any fixed string $\boldsymbol{x}$. Then most likely some bad events A_i will be true.

[2] See the blog posts by `http://terrytao.wordpress.com/2009/08/05/mosers-entropy-compression-argument` and `http://terrytao.wordpress.com/2009/08/05/mosers-entropy-compression-argument`

[3] The argument can be easily generalized to product distributions. But we use bits to have a simpler notation and get a better intuition.

Then one resamples the variables belonging one bad event. If an overlapping bad event A_j is true (either because it was true from the beginning or flipping the bits in $\mathrm{Var}(A_i)$ made it true) then we resample also A_j and recurse. For an easier statement and analysis we assume w.l.o.g. that $|\mathrm{Var}(A_i)| = k$ for each i. If fewer bits were needed one can imagine to add some dummy random bits that are "private" for the event. The formal statement of the algorithm is as follows:

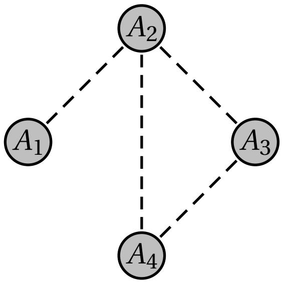
Moser-Tardos Algorithm

```
(1) Set $\boldsymbol{x} := (0,\dots,0)$
(2) WHILE $\exists i : A_i(\boldsymbol{x}) = 1$ DO
        (3) OUTPUT: ROOT+$i$
        (4) FIX($A_i$)
(5) RETURN $\boldsymbol{x}$.
```

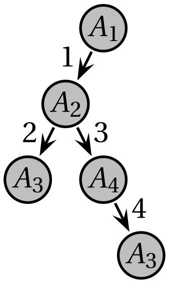
Subroutine FIX(A_i) :

```
(6) Resample $(x_j)_{j\in\mathrm{Var}(A_i)} \sim \{0,1\}^k$
(7) after $T$ iterations: OUTPUT $x_1,\dots,x_n$ and FAIL!
(8) WHILE $\exists A_j : |\mathrm{Var}(A_i)\cap\mathrm{Var}(A_j)| > 0$ s.t. $A_j(\boldsymbol{x}) = 1$ DO
        (9) OUTPUT: RECURSE ON Huffman$_{A_i}(A_j)$ DUE TO Compress$_{A_j}((x_\ell)_{\ell\in\mathrm{Var}(A_j)})$
        (10) FIX($A_j$)
(11) OUTPUT: BACKTRACK
```

Note that line (3) will be executed at most m times as FIX(A_i) only returns once all touched bad events are false. Also note that we number the iterations as $t = 1,\dots,T$ and T is the total number of FIX calls (either recursive or via (3)). Note that one call of FIX(A_i) will result in a recursion tree that might look like this:



dependency graph



possible recursion tree of FIX(A_1) (edge labels give order)

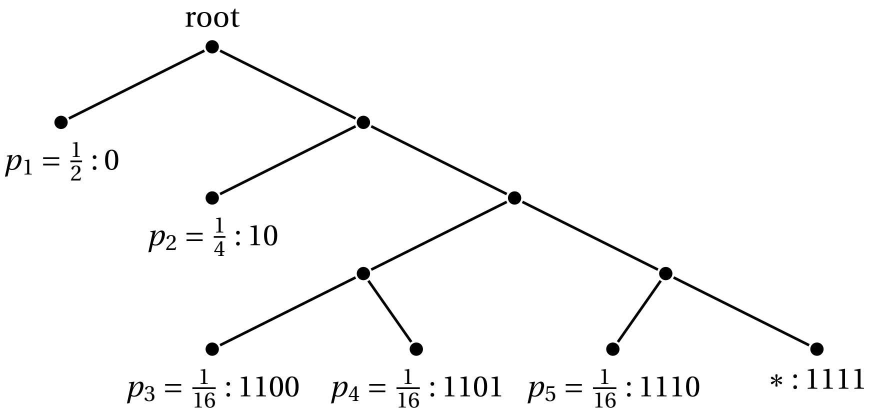
In particular it can happen that the same event is resampled multiple times.

Next, note that the algorithm contains several "OUTPUT" lines that are not doing anything for the functionality. Instead they will be useful for the analysis and the proof that the algorithm terminates with good enough probability. Maybe one can imagine that the algorithm outputs some information into a "debugging log file". Also the algorithm outputs bitstrings $\text{Huffman}_{A_i}(A_j)$ and $\text{Compress}_{A_j}((x_\ell)_{\ell\in\text{Var}(A_j)})$ which we need to define as well. We recall the following classical information-theoretic insight:

Lemma 3.6 (Huffmann Tree)**.** *For any values $p_1,\dots,p_N \geq 0$ with $\sum_{i=1}^N p_i \leq 2^{-C}$ and $C \in \mathbb{Z}_{\geq 0}$, there is a binary tree with leafs $1,\dots,N$ where leaf i has at most distance $\log_2(\frac{1}{p_i}) - C + 1$ to the root.*

Proof sketch. Round p_i down to the nearest power of 2, multiply by 2^C and denote it by p_i' so that now $\sum_{i=1}^N p_i' \leq 1$. Then by induction we can easily construct a tree so that a node i with $p_i' = (1/2)^\ell$ has distance exactly ℓ to the root (and we may have some unused leafs). □

The obtained tree for $C = 0$ could look like this:


root
$p_1 = \frac{1}{2} : 0$
$p_2 = \frac{1}{4} : 10$
$p_3 = \frac{1}{16} : 1100$
$p_4 = \frac{1}{16} : 1101$
$p_5 = \frac{1}{16} : 1110$
$* : 1111$

Huffman encoding for $C = 0$ and $(p_1, p_2, p_3, p_4, p_5) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16})$

This implies that for each fixed event A_i we can encode the indices j with $j \sim i$ with a bitstring $\text{Huffman}_{A_i}(A_j)$ that has length $\log_2(\frac{1}{\Pr[A_j]}) - C + 1$. In particular, this means that indices of likely events will be encoded with short strings, while we can use longer strings for unlikely events. As an unrelated side remark, we would like to mention that the Huffman Tree also proves that a sequence of independent samples of a random variable X can be encoded with an expected number of $H(X) + O(1)$ bits per sample.

To finalize the description of the algorithm consider again step (9) where a bad event A_j is true for the current assignment $\boldsymbol{x}$ and abbreviate $\boldsymbol{y} := (x_\ell)_{\ell\in\text{Var}(A_j)} \in$

$\{0,1\}^k$. The intuition behind $\text{Compress}_{A_j}(\boldsymbol{y})$ is that we want to output the k bits in a more space-efficient way. Note that at this point we know that $A_j(\boldsymbol{y}) = 1$ and hence there are at most $2^k \cdot \Pr[A_j]$ many choices for $\boldsymbol{y}$. Hence $\text{Compress}_{A_j}(\boldsymbol{y})$ can be chosen as a bitstring of length $k - \log(\frac{1}{\Pr[A_j]}) + 1$ that uniquely identifies $\boldsymbol{y}$.

The analysis of the Moser-Tardos Algorithm. We already did most of the work to justify why have chosen the log output in step (9) in this particular way. Note that in the following $\Theta(1)$ will denote a constant that has no dependence on C.

Lemma 3.7. *In every step of (9) we output at most* $k + \Theta(1) - C$ *bits.*

Proof. Whatever index j with $j \sim i$ is picked, the number of bits is

$$\underbrace{\left(\log_2\left(\frac{1}{\Pr[A_j]}\right) - C + 1\right)}_{\text{Bits for Huffman}_{A_i}(A_j)} + \underbrace{\left(k - \log_2\left(\frac{1}{\Pr[A_j]}\right) + 1\right)}_{\text{Bits for Compress}} + \underbrace{O(1)}_{\text{overhead}} \leq k - C + O(1)$$

□

Now, let $\boldsymbol{x}^{(t)}$ be the vector $\boldsymbol{x}$ *after* the t-th iteration. First we argue that the states of the algorithm have high entropy:

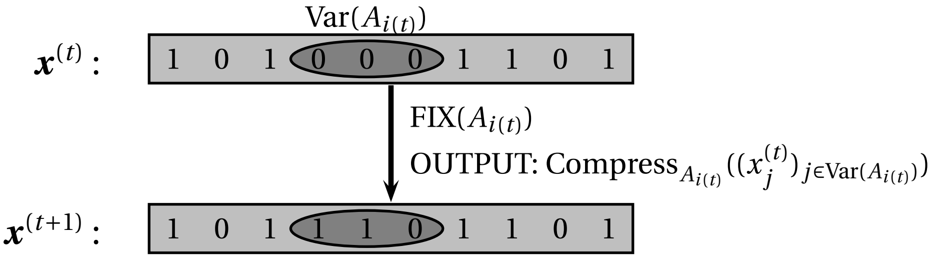
Lemma 3.8. *If the algorithm fails with probability* $1-\varepsilon$*, then* $H(\boldsymbol{x}^{(0)},\ldots,\boldsymbol{x}^{(T)}) \geq T \cdot (1-\varepsilon) \cdot k$.

Proof. If the algorithm fails then we will have sampled $T \cdot k$ independent random bits. □

Lemma 3.9. *One has* $H(\boldsymbol{x}^{(0)},\ldots,\boldsymbol{x}^{(T)}) \leq n + O(m \log_2(m)) + T \cdot (k - C + O(1))$.

Proof. We can split the proof into two claims.
Claim I. *The length of the output log is at most* $n + O(m\log_2(m)) + T \cdot (k - C + O(1))$ *bit.*
Proof of claim. Step (3) is visited at most m times using $O(\log_2(m))$ bits each time to encode the index $i \in [m]$. In each recursive call of FIX(A_j) we spend $k - C + O(1)$ bits. Additionally, to indicate BACKTRACKING we may spend $O(1)$ bits, but the total number of backtrackings is also bounded by T. □
Claim II. *From the output log we can reconstruct* $\boldsymbol{x}^{(0)},\ldots,\boldsymbol{x}^{(T)}$.
Proof of claim. Suppose that the algorithm calls FIX($A_{i(1)}$),...,FIX($A_{i(T)}$) in exactly this order. Clearly from the output log we can reconstruct the indices $i(1),\ldots,i(T)$. Also note that the algorithm outputs the whole assignment $\boldsymbol{x}^{(T)}$ at the end. We

will then reconstruct the sequence $\boldsymbol{x}^{(T)}, \boldsymbol{x}^{(T-1)}, \dots, \boldsymbol{x}^{(0)}$ in this order. Suppose that at some point we know $\boldsymbol{x}^{(t+1)}$. From the output we know the indices of the events $A_{i(t+1)}$ and $A_{i(t)}$ and also the string $\text{Compress}_{A_{i(t)}}(\boldsymbol{x}^{(t)})$ which tells us the bits in $\text{Var}[A_{i(t)}]$ before the resampling took place.



Then we can also reconstruct $\boldsymbol{x}^{(t)}$. □

The overall claim follows from the combination of Claim I and Claim II. □

Lemma 3.10. *The probability that the algorithm terminates with success within* $O(n + m\log(m))$ *iterations is at least* $1 - \frac{1}{2k}$.

Proof. Choosing $C > 0$ as a big enough constant and $\varepsilon := \frac{1}{2k}$ we have

$$T \cdot \left(k - \frac{1}{2}\right) \leq H(\boldsymbol{x}^{(0)}, \dots, \boldsymbol{x}^{(T)}) \leq n + O(m \log_2(m)) + T \cdot (k-1)$$

This gives a contradiction if we pick $T := \Theta(n + m\log(m))$ with a big enough constant. □

3.3 Open problems

A deterministic approach. The algorithm by Moser and Tardos is randomized and it is a natural question whether there is a deterministic polynomial time algorithm for the LLL. This was somewhat answered by Chandrasekaran, Goyal and Haeupler [CGH10] where the authors consider the *witness tree* analysis (that we have not presented here) which can be seen as the recursion tree of the algorithm. The original work of Moser and Tardos shows that the expected size of this recursion tree is polynomial. Then [CGH10] show that the expected size can be computed and one can deterministically make choices so that the remaining expected size goes down. While this technically answers the question, it does not provide additional insides into the LLL. Alternatively, one might hope for an explicit, intuitive potential function (differently from the expected witness tree size) that can be easily computed and more naturally provides a deterministic algorithm.

The Kadison-Singer problem. The Kadison-Singer Conjecture from 1959 was a problem that somewhat independently appeared in an unusually large number of subfields of mathematics. Intuitively the question was whether a set of short vectors in isotropic position could be partitioned into two spectrally similar parts. It was very recently resolved by Marcus, Spielman and Srivastava [MSS15] who proved the following:

Theorem 3.11 (Marcus, Spielman and Srivastava 2013)**.** *Given vectors* $\boldsymbol{v}_1,\dots,\boldsymbol{v}_m \in \mathbb{R}^n$ *with* $\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^T = \boldsymbol{I}_n$ *and* $\|\boldsymbol{v}_i\|_2 \leq \varepsilon$*, there exists a partition* $[n] = I_1\dot{\cup}I_2$ *so that*

$$\Big(\frac{1}{2}-\Theta(\varepsilon)\Big)\cdot\boldsymbol{I}_n \preceq \sum_{i\in I_j}\boldsymbol{v}_i\boldsymbol{v}_i^T \preceq \Big(\frac{1}{2}+\Theta(\varepsilon)\Big)\cdot\boldsymbol{I}_n \quad \forall j\in\{1,2\}.$$

The proof of this spectacular result uses *interlacing polynomials* and the result is non-constructive in the sense that it is currently unknown whether the partition $[n] = I_1\dot{\cup}I_2$ can be computed in polynomial time. In fact, it is a challenging open problem to find such a polynomial time algorithm. However, this can also be seen as an open problem connected to the Lovász Local Lemma. First note that the theorem of [MSS15] can be rephrased, as the fact that there is a *coloring* $\boldsymbol{x} : [m] \to \{-1,1\}$ so that

$$\sum_{i=1}^m x_i\,\langle\boldsymbol{v}_i,\boldsymbol{y}\rangle^2 = \langle\sum_{i=1}^m x_i\boldsymbol{v}_i\boldsymbol{v}_i^T,\boldsymbol{y}\boldsymbol{y}^T\rangle \leq O(\varepsilon) \quad \forall\|\boldsymbol{y}\|_2 = 1$$

This condition feels like finding a coloring that is good for infinitely many directions or "sets". For the sake of argument imagine to just take the coloring $\boldsymbol{x} \sim \{-1,1\}^m$ uniformly at random. Then the variance in some direction $\boldsymbol{y}$ with $\|\boldsymbol{y}\|_2 = 1$ is

$$\mathrm{Var}\Big[\sum_{i=1}^m x_i\,\langle\boldsymbol{v}_j,\boldsymbol{y}\rangle^2\Big] = \sum_{i=1}^m\langle\boldsymbol{v}_i,\boldsymbol{y}\rangle^4 \leq O(\varepsilon^2)\sum_{i=1}^m\langle\boldsymbol{v}_i,\boldsymbol{y}\rangle^2 = O(\varepsilon^2)$$

That means to achieve the MSS Theorem one can only afford a deviation in each direction of $O(1)$ times the standard deviation. If we are aiming for a weaker bound and replace the bound of $\frac{1}{2}\pm\Theta(\varepsilon)$ with say $\frac{1}{2}\pm\frac{1}{4}$, then we can afford a deviation of $\Theta(\frac{1}{\varepsilon})$ times the standard deviation. While one has infinitely many vectors, the events in most directions are little correlated. It would be beautiful if one could reprove the Kadison-Singer question (even in a weaker form) using a "spectral variant" of the Lovász Local Lemma. This might also give a polynomial time algorithm.

3.4 Exercises

Exercise 3.1. Let $n \geq 2k$ and $D \in \mathbb{N}$. Prove that there exists a k-uniform hypergraph $H = (V, \mathcal{E})$ on n vertices with degree at most D so that the fraction of colorings $\chi : V \to \{\text{red}, \text{blue}\}$ that have property B is at most $\exp(-D \cdot 2^{-\Theta(k)} n))$.
Hint: A random k-uniform hypergraph will do the job.

Exercise 3.2. Let $H = (V, \mathcal{E})$ be a t-uniform hypergraph where $\deg_H(v) \leq t$ for all $v \in V$. Prove that there exists a coloring $\chi : V \to \{\text{red}, \text{blue}\}$ so that for every edge $e \in \mathcal{E}$ one has

$$|(\#i \in e : \chi(i) = \text{red}) - (\#i \in e : \chi(i) = \text{blue})| \leq O(\sqrt{t \log(t)}).$$

Exercise 3.3. Fix a $\alpha > 0$. For a graph $G = (V, E)$ with $|V| = n$ one defines the *edge expansion* as

$$\Phi(E) := \min_{S \subseteq V : |S| \leq n/2} \left\{ \frac{|\delta(S)|}{|S|} \right\}$$

Assume that G is r-regular and $\Phi(E) \geq \alpha \cdot r$ (which means that the graph G is an expander graph). Prove that if r is big enough — dependent on α, but not on n — then there is a partition $E = E_1 \dot\cup E_2$ so that $\frac{1}{3} r \leq \deg_{E_i}(v) \leq \frac{2}{3} r$ and $\Phi(E_i) \geq \frac{\alpha}{3} r$ for both $i \in \{1, 2\}$.
Hint: You will need the general LLL. Consider the family $\mathcal{F}_{v,s} := \{S \subseteq V \mid v \in S \text{ and } |S| = s \text{ and } S \text{ is connected}\}$. Follow the following strategy: (i) show that it suffices to lowerbound $|\delta_{E_i}(S)|/|S|$ for S so that the induced subgraph $G[S]$ is *connected*; (ii) prove that $|\mathcal{F}_{v,s}| \leq (\Theta(r))^{s-1}$; (iii) what is the probability for a fixed $S \in \mathcal{F}_{v,s}$ that $|\delta_{E_i}(S)|/|S| < \frac{\alpha}{3} r$? If we have a bound $p_{|S|}$ on the probability of the event $\frac{1}{3} \alpha r |S| \leq |\delta_{E_1}(S)| \leq \frac{2}{3} \alpha r |S|$ failing, then pick an overestimate of say $x_S := p_S^{1/10}$.

Exercise 3.4. We consider the following routing problem: we are given an undirected graph $G = (V, E)$ with pairs $(s_1, t_1), \ldots, (s_k, t_k) \in V \times V$ and a parameter $D \in \mathbb{N}$. The goal is to find to find s_i-t_i paths P_i of length at most D so that the *congestion* $\max_{e \in E} |\{i \in [k] \mid e \in P_i\}|$ is minimized. Assume that in the optimum solution the congestion is 1. One can use linear-programming techniques so that for each pair i we find a *collection* of s_i-t_i pairs $\{P_i^1, \ldots, P_i^N\}$ with small congestion. In particular the paths will satisfy the length bound of $|P_i^j| \leq D$ for all i and j and the average congestion is $\frac{|\{j : e \in P_i^j\}|}{N} \leq 1$ for all i (you do not have to prove these properties). Let us also assume for the sake of simplicity that N is a power of 2. Prove that under these assumptions there are s_i-t_i paths so that each edge has congestion at most $O(\log D)$.
Hint. Apply the LLL iteratively to sparsify the collection of paths.

Exercise 3.5. A function $c : \mathbb{R} \to \{1, 2, \ldots, k\}$ is called a *k-coloring* of the real numbers. We say that a subset $T \subseteq \mathbb{R}$ is *multicolored* if $c(T) = \{1, \ldots, k\}$, meaning that T contains elements of all colors.

Prove that for parameters $k, m \in \mathbb{N}$ with $m \geq 10k \ln(10k)$ the following holds: for any finite set $S \subseteq \mathbb{R}$ with $m = |S|$ elements and any finite $X \subseteq \mathbb{R}$, there is a k-coloring $c : \mathbb{R} \to \{1, \ldots, k\}$ so that the translates[4] $x + S$ are multicolored for all $x \in X$.

[4] Here $x + S$ is defined as $\{x + s \mid s \in S\}$.

Chapter 4

Point line incidences and the Crossing Number Theorem

The probabilistic method is also a powerful tool in answering geometric questions. The motivating question for this chapter is the following: suppose we have a set of *lines* L and *points* $P \subseteq \mathbb{R}^2$ in the Euclidean plane. Let $\mathcal{I}(P,L)$ be the number of *incidences* between those lines, meaning it is the number of pairs $(\ell, p) \in L \times P$ so that p lies on ℓ.

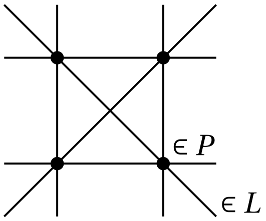


Fig: Example with $|P| = 4$ points, $|L| = 6$ lines, $\mathcal{I}(P,L) = 4 \cdot 3 = 12$ incidences

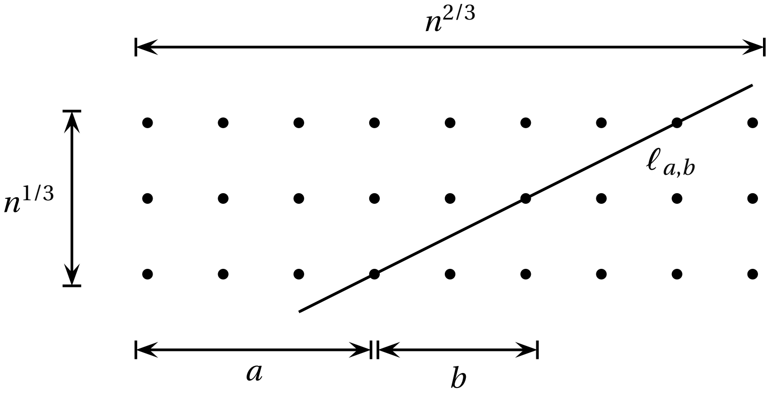
Then the question is: *how large can the number of incidences $\mathcal{I}(P,L)$ be, dependent on the number $|P|$ of points and the number $|L|$ of lines.* And for simpler notation we will be focusing on the symmetric case with $|L| = n = |P|$. For this chapter we will mostly follow the brilliant exposition of Matousek [Mat02].

4.0.1 A lower bound

To get some intuition on what the right answer might be, we want to start with a lower bound construction.

Lemma 4.1. *There are n points and n lines so that $\mathcal{I}(P,L) \geq \Omega(n^{4/3})$.*

Proof. One can quickly come up with the educated guess that the points should be arranged in some form of a *grid* in order to maximize the incidences. So let P be an $n^{2/3} \times n^{1/3}$ size grid. Let $\ell_{a,b}$ be the line going through point $(a,1)$ and having slope $\frac{1}{b}$. We pick the lines $L := \{\ell_{a,b} \mid a \in \{1,\dots,n^{2/3}\}$ and $b \in \{1,\dots,n^{1/3}\}\}$ as our lines.



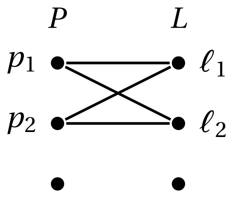
Then we claim that there are $\Omega(n^{4/3})$ many incidences. The best way to see this is by observing that for $a \le \frac{1}{2}n^{2/3}$ and $b \le \frac{1}{2}n^{1/3}$ the line $\ell_{a,b}$ has exactly $n^{1/3}$ incidences. □

4.0.2 An upper bound based on forbidden subgraphs

Many results in particular in geometric combinatorics are obtained by using *forbidden subgraph* arguments. We say that $H = (V_H, E_H)$ is a *subgraph* of $G = (V,E)$ if there is an injective map $f : V_H \to V$ so that $\{u,v\} \in E_H \Rightarrow \{f(u), f(v)\} \in E$. In particular this gives a monotone relation in the sense that if H is a subgraph of G then this remains true if we delete edges from H or add more edges to G. Particularly important classes of subgraphs are K_r — the complete graph on r nodes; and $K_{r,s}$ — the complete bipartite graph with r nodes on the left side and s nodes on the right hand side.

Theorem 4.2. *For any points P and lines L in the plane with $|P| = n = |L|$ one has $\mathcal{I}(P,L) \le O(n^{3/2})$.*

Proof. Define a bipartite graph $G = (P \dot{\cup} L, E)$ with an edge $\{p,\ell\} \in E$ if the point p lies on ℓ. In particular $|E| = \mathcal{I}(P,L)$. Then this graph does *not* contain a $K_{2,2}$ as two lines can only intersect in one point, not more.



forbidden $K_{2,2}$

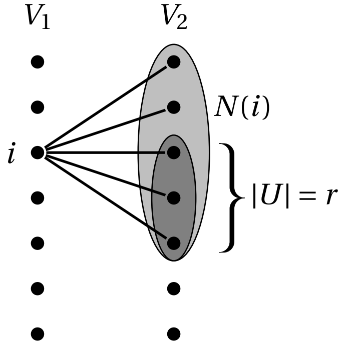
But any graph without a $K_{2,2}$ subgraph has at most $O(n^{3/2})$ many edges as we will see in the next Theorem in more generality. □

Theorem 4.3 (Kövari-Sos-Turan Theorem)**.** *Let $G = (V,E)$ be a graph without an induced $K_{r,r}$ subgraph. Then $|E| \leq O(r)\cdot n^{2-1/r}$.*

Proof. It suffices to prove the bound for a bipartite graph $G = (V_1 \dot\cup V_2, E)$ with $|V_1| = n = |V_2|$. The reason is that we could partition the original graph G into two partitions while keeping at least half of the edges. Let $N(i) \subseteq V_2$ be the set of neighbors for a node $i \in V_1$. Basically the proof consists of the following 1-line estimate that double-counts the number of r-tuples in neighborhoods $N(i)$:

$$\frac{|E|^r}{r^r n^{r-1}} = \frac{n}{r^r}\cdot\Big(\frac{1}{n}\underbrace{\sum_{i\in V_1}|N(i)|}_{=|E|}\Big)^r \overset{(2)}{\leq} \frac{1}{r^r}\sum_{i\in V_1}|N(i)|^r \leq \sum_{i\in V_1}\binom{|N(i)|}{r} \overset{(1)}{\leq} (r-1)\cdot\binom{n}{r} \leq r\cdot n^r$$

The crucial argument for (1) is the following: consider any set $U \subseteq V_2$ of size $|U| = r$. Then there are at most $r-1$ many nodes $i \in V_1$ so that $U \subseteq N(i)$ as otherwise we would have found a $K_{r,r}$.



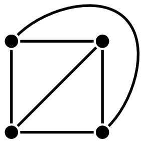
Then the r-tuple U is counted at most $r-1$ times on the left hand side of (1). In (2) we use *Jensen's inequality* with the fact that $f(x) = x^r$ is a convex function. Rearranging the inequality gives $|E| \leq (r\cdot r^r \cdot n^{2r-1})^{1/r} = O(r)\cdot n^{2-1/r}$ as claimed. □

Note that there are indeed graphs without a $K_{2,2}$ subgraph that have $\Theta(n^{3/2})$ many edges. The proof can be slightly extended to obtain that any graph without a $K_{r,s}$ subgraph with $r \leq s$ has at most $O_{r,s}(1) \cdot n^{2-1/r}$ many edges. In other words: more or less only the smaller side of the forbidden subgraph counts.

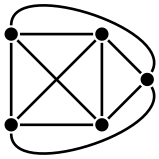
4.1 Crossing numbers

For an undirected graph $G = (V, E)$, a *drawing* consists of an injective map $f : V \to \mathbb{R}^2$, mapping points into the plane and continuous injective maps $g_{uv} : [0,1] \to \mathbb{R}^2$ for every edge $\{u, v\} \in E$ modelling the *arcs*. We require that the maps g_{uv} start and end at $f(u)$ and $f(v)$. Moreover, arcs may not run through vertices other than their own, meaning the interior of the arcs does not contain a node. A *crossing* is a point that is common to at least two arcs. The *crossing number* of a drawing is the number of crossings, but counting a crossing of k arcs with multiplicity $\binom{k}{2}$. In other words, one could always perturb the arcs a little so that the crossing number remains the same but at any point no more than 2 arcs cross.

Obviously a graph can have drawings with many crossings and some with few crossings. The *crossing number* $\text{cr}(G)$ of a graph G is then the minimum number of crossings of any drawing. A drawing is *planar* if the number of crossings is 0 and a graph is planar if it has a planar drawing, i.e. $\text{cr}(G) = 0$.



planar drawing of K_4

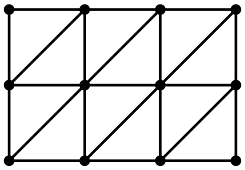


drawing of K_5 with 1 crossing

The question that we will aim to solve now is: *what lower bound on the crossing number can one give, depending only on the number of edges*? A well known condition is the following:

Lemma 4.4. *Any planar graph $G = (V, E)$ has $|E| \leq 3n - 6$.*

To see this, we can denote F as the *faces* of a planar drawing. Then *Euler's formula* says that $|E| - |F| = |V| - 2$. As every face is bounded by at least 3 edges and every edge is incident to only 2 faces we also have $|F| \leq \frac{2}{3}|E|$. Putting this together gives $|E| \leq 3n - 6$. Note that this inequality can be read as the fact that in a planar graph, the *average degree* is slightly below 6. And in fact there are planar graphs whose average degree is tend to 6 as $n \to \infty$, for example the "grid graph with diagonals".



Now we can come to the *Crossing Number Theorem* and it's beautiful probabilistic proof. To get some intuition about the quantity note that it says that a graph with $|E| = 5|V|$ many edges might not have a crossing, but a dense graph with $|E| = \Theta(n^2)$ edges will have $\Theta(n^4)$ many crossings. In other words, for dense graphs, there is no "good" drawing — all drawings are equally bad up to a constant factor.

Theorem 4.5 (Crossing Number Theorem)**.** *Let $G = (V,E)$ be an undirected graph. Then $cr(G) \geq \frac{1}{64} \cdot \frac{|E|^3}{|V|^2} - |V|$.*

Proof. We begin by proving a rather weak lower bound on the crossing number:
Claim. One has $\mathrm{cr}(G) \geq |E| - 3|V|$.
Proof of claim. Fix the drawing that attains the crossing number and set $k := \mathrm{cr}(G)$. Remove one edge after the other while each removed edge was involved in a crossing. That means we can find edges $e_1,\dots,e_k$ so that the "inherited" drawing is crossing free for the edges in $E \setminus \{e_1,\dots,e_k\}$. Then $|E \setminus \{e_1,\dots,e_k\}| \leq 3|V| - 6 \leq 3|V|$ by the previous Lemma. This can be rearranged to $k \geq |E| - 3|V|$. □

The claim gives already the right asymptotic bound if the graph is *sparse*, i.e. $|E| = \Theta(|V|)$. We can now use a probabilistic argument to reduce the general case to the sparse case. Take an arbitrary graph $G = (V,E)$ and let $n := |V|$ and $m := |E|$. We may assume that $m \geq 4n$ otherwise we are done. For a probability $p \in [0,1]$, we pick a subset $V' \subseteq V$ at random that contains each node independently with probability p. Then we obtain a random subgraph $G' = (V',E')$ that inherits an edge $\{u,v\} \in E$ if both end points are sampled into V'. Now, for the crossing number $\mathrm{cr}(G')$ of the random subgraph we obtain that

$$p^4\mathrm{cr}(G) \overset{(*)}{\geq} \mathbb{E}[\mathrm{cr}(G')] \overset{\text{Claim+linearity}}{\geq} \underbrace{\mathbb{E}[|E'|]}_{=p^2m} - 3\underbrace{\mathbb{E}[|V'|]}_{=pn} = p^2m - 3pn$$

In $(*)$ we use the following observation: consider the optimum drawing for G and delete all nodes not in V' and all edges that do not have both end points in V'.

Then a crossing between $\{u_1, u_2\}$ and $\{u_3, u_4\}$ only remains if $u_1, u_2, u_3, u_4 \in V'$. That means a crossing only survives with probability[1] p^4. Hence $\mathbb{E}[\mathrm{cr}(G')] \leq p^4 \cdot \mathrm{cr}(G)$ (and this is only an upper bound because there is no guarantee that the inherited drawing is still optimal for G'). Rearranging and choosing p gives

$$\mathrm{cr}(G) \geq \frac{m}{p^2} - \frac{3n}{p^3} \overset{p:=\frac{4n}{m}}{=} \left(\frac{1}{16} - \frac{3}{64}\right) \cdot \frac{m^3}{n^2}$$

as desired. □

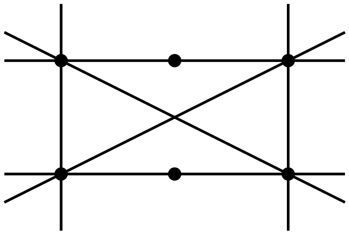
Note that for the choice of $p = \frac{4n}{m}$ we have $\mathbb{E}[|V'|] = pn = \Theta(\frac{n^2}{m})$ and $\mathbb{E}[|E'|] = pm^2 = \Theta(\frac{n^2}{m})$. That means we have indeed reduced to the case of sparse graphs as we have claimed earlier.

4.2 The Szemerédi-Trotter Theorem

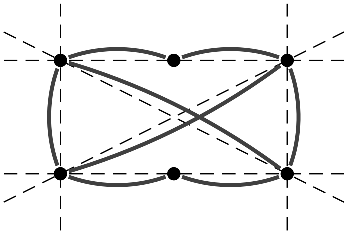
Now we come to a tight upper bound on the number of incidences. The simple proof that we see here is due to Székely.

Theorem 4.6 (Szemerédi-Trotter Theorem)**.** *For any set of points P and lines L in the plane with $|P| = |L| = n$ one has $\mathcal{I}(P, L) \leq O(n^{4/3})$.*

Proof. We define a graph $G = (P, E)$ that has a vertex for each of the points in P. Then we insert an edge $\{p, p'\} \in E$, if p and p' are consecutive on some line $\ell \in L$. Phrased differently, for a line $\ell \in L$ let $p_1, \ldots, p_k \in P$ be the points in their natural order; then $\{p_i, p_{i+1}\} \in E$ for all $i = 1, \ldots, k-1$.



points P and lines L



graph $G = (P, E)$

[1] Note that the bound of p^4 would not hold if there would be a crossing of two edges incident to the *same* node u. But one can argue that the optimum drawing would never have arcs coming from the one node intersect. If there was such an intersection, one could swap the segments between the first crossing and the common node u and then remove the crossing itself, resulting in a drawing with lower crossing number.

Then a line ℓ that contributes k many incidences to $\mathcal{I}(P,L)$, contributes exactly $k-1$ edges to E. Hence $\mathcal{I}(P,L) \leq |E| + n$. So it suffices to bound the number of edges $|E|$. For this sake, we inspect the *crossing number* of the constructed graph. And indeed one has

$$\frac{|E|^3}{64n^2} - n \overset{(*)}{\leq} \mathrm{cr}(G) \overset{(**)}{\leq} n^2$$

Here we use the crossing number theorem in $(*)$. To get the bound on the crossing number in $(**)$, we use the straightline drawing where the arc for an edge $\{p,p'\} \in E$ goes along the line ℓ with $p,p' \in \ell$. Then arcs cross only where lines cross; and there are at most n^2 such crossings of lines.

Rearranging the above inequality then gives that $|E| \leq O(n^{4/3})$ as needed. □

4.3 Exercises

Exercise 4.1 (From Matousek [Mat02])**.** For a set of points P and unit circles C in the plane, let $\mathcal{I}(P,C)$ be the number of incidences. Show that for $|P| = n = |C|$ one has $\mathcal{I}(P,C) \leq O(n^{4/3})$.

Exercise 4.2 (From Matousek [Mat02])**.** Prove that for any n and $Cn \leq m \leq \binom{n}{2}$, there is an n-vertex graph with m edges and crossing number at most $O(m^3/n^2)$, where $C > 0$ is a big enough constant.

Exercise 4.3. Fix a prime number $p \in \mathbb{N}$. Define $A := \{a = (a_1, a_2, 1) \in \mathbb{Z}_p^3\}$ and $B := \{b = (b_1, b_2, 1) \in \mathbb{Z}_p^3\}$. Consider the bipartite graph $G = (A \dot\cup B, E)$ with $E = \{(a,b) \in A \times B \mid \sum_{i=1}^3 a_i b_i \equiv_p 0\}$. Prove that G has $\Theta(n^{3/2})$ many edges where $n := |A| = |B|$ but does not contain a $K_{2,2}$.

Exercise 4.4. The Crossing Number Theorem only applies to simple graphs without parallel edges. We will now extend it to graphs having parallel edges. Let $G = (V,E)$ be a *multi-graph* and k be the *maximum edge multiplicity* (meaning that an edge $\{u,v\}$ may exist up to k many times). Prove that $\mathrm{cr}(G) \geq \Omega(\frac{|E|^3}{k|V|}) - O(k^2|V|)$, where $|E|$ counts each edge with multiplicity.

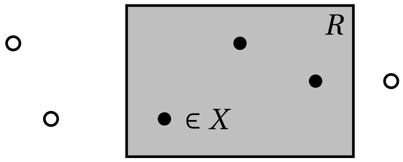
Exercise 4.5. Use the probabilistic method to prove that for any $r, n \in \mathbb{N}$ with $2 \leq r \leq n$, there is a bipartite graph $G = (V_1 \dot\cup V_2, E)$ with $|V_1| = |V_2| = n$ that does not contain a $K_{r,r}$ subgraph and has $|E| \geq C_r \cdot n^{2-2/(r+1)}$ many edges. Here $C_r > 0$ is some constant that may depend on r.
Hint. Use the method of alterations.

Chapter 5

VC dimension and ε-nets

In the following, we will study the complexity of set systems $(X, \mathcal{F})$, where X is the *ground set* with $|X| = n$ elements and $\mathcal{F} \subseteq 2^X$ is the *family of subsets*. Here we will follow in part [Mat02] and in part [AS16]. To make it more concrete, consider a set $X \subseteq \mathbb{R}^2$ of n points in the plane and let

$$\mathcal{F} := \{|R \cap X| : R \text{ is an axis parallel rectangle}\}$$



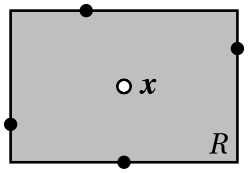
What can we say about the *complexity* of this set system? On the one hand, there are infinitely many rectangles — but the intersection with X can trivially give at most $|\mathcal{F}| \leq 2^n$ many combinations. But it is hard to imagine that all possible subsets of X could be realized as intersection with a rectangle. In fact, later we will see that even $|\mathcal{F}| \leq O(n^4)$.

One can imagine that in many other geometric settings the subsets have a lot of structure — they are "simpler" than arbitrary set systems. This phenomenon is brilliantly captured by the concept of *VC dimension*, named after Vapnik and Chervonenkis. First, we say that a subset $A \subseteq X$ is *shattered* by $\mathcal{F}$, if all $2^{|A|}$ possible subsets of A can be obtained as intersections of the form $A \cap S$ with $S \in \mathcal{F}$. Phrased differently, A can be shattered if $\mathcal{F}_{|A} = 2^A$, where $\mathcal{F}_{|A} = (A, \{A \cap S : S \in \mathcal{F}\})$ is the *set system induced on A*. Then the *VC dimension* of $(X, \mathcal{F})$ is the largest cardinality $|A|$ of a subset that can be shattered.

In planar geometry one often assumes that some set of points $X \subseteq \mathbb{R}^2$ is in *general position* meaning that the points have no constellation that has a zero probability of occurring if the points would have been picked at random from some continuous distribution. For example if X is in general position, no 3 points should be on a line and no 2 points should have the same x-coordinate or the same y-coordinate.

Lemma 5.1. *Let $X \subseteq \mathbb{R}^2$ be points in general position and $|X| \geq 4$ and let $\mathcal{F} = \{|R \cap X| : R$ is an axis parallel rectangle$\}$. Then the VC dimension of $(X, \mathcal{F})$ is at most* 4.

Proof. Suppose for the sake of contradiction that some set $A \subseteq X$ of $|A| = 5$ points can be shattered. Take the unique minimal rectancle R containing A. By the general position assumption, each side of R contains exactly 1 point. So there is one point $x \in A$ in the interior of R. Then there is no set $S \in \mathcal{F}$ with $S \cap A = A \setminus \{x\}$.



□

With some more case distinctions one can also show the claim for points that are not in general position. Note that in general one will not be able to improve this bound. For example if there is a rectangle R that has exactly one point of A on each side, then one can shatter A. To do so, for a subset $A' \subseteq A$ take the rectangle R and simply shorten the sides containing points of $A \setminus A'$. We will see more examples in the exercises.

5.1 The Shatter-function

We want to go back to the question of how many different sets are there in a set system with bounded VC dimension. In our example $(X, \mathcal{F})$ of axis parallel rectangles, we determined that the VC dimension is 4. Also we have seen that any *minimal* rectangle is uniquely determined by at most 4 points. Hence we can argue that at least for axis parallel rectangles we will only have $|\mathcal{F}| \leq O(n^4)$ many different sets. From that consideration one might get hopeful that for a general set system one could prove a $O(n^d)$ bound, where $d = \dim(\mathcal{F})$. And indeed that is true:

Lemma 5.2 (Shatter Function Lemma)**.** *Define* $\Phi_d(n) := \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{d}$*. Then any set system* $(X, \mathcal{F})$ *with* $|X| = n$ *elements and VC dimension* $d = \dim(\mathcal{F})$ *has* $|\mathcal{F}| \leq \Phi_d(n)$*.*

Proof. We will prove that $|\mathcal{F}| \leq \Phi_d(n)$ by induction. Fix an element $x \in X$. We define

$$\mathcal{F}_1 := \mathcal{F}_{|X \setminus \{x\}} \quad \text{and} \quad \mathcal{F}_2 = \{S \mid S \in \mathcal{F} \text{ and } S \dot{\cup} \{x\} \in \mathcal{F}\}$$

Then $\mathcal{F}_1$ is the set system that we obtain by deleting x. Note that $\mathcal{F}_1$ might have fewer sets than $\mathcal{F}$ as any set $S \subseteq X \setminus \{x\}$ with $S \in \mathcal{F}$ and $S \cup \{x\} \in \mathcal{F}$ would be collapsed to just one set. But these sets are being count in $\mathcal{F}_2$. Hence $|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_2|$. By induction we have $|\mathcal{F}_1| \leq \Phi_d(n-1)$. Next, considering $\mathcal{F}_2$ we observe hat $\dim(\mathcal{F}_2) \leq d-1$. The reason is that if $A \subseteq X \setminus \{x\}$ was shattered in $\mathcal{F}_2$, then $A \cup \{x\}$ is shattered by $\mathcal{F}$. Then

$$|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_2| \leq \Phi_d(n-1) + \Phi_{d-1}(n-1) = \Phi_d(n)$$

using the recurrence $\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}$. □

We also want to introduce a general notion that is usually used in this context. We define the *Shatter Function* of the set system $(X, \mathcal{F})$ as

$$\pi_{\mathcal{F}}(m) := \max_{Y \subseteq X : |Y| = m} |\mathcal{F}_{|Y}|$$

In other words, for every $m \in \{1, \ldots, n\}$, $\pi_{\mathcal{F}}(m)$ gives the maximum number of different sets that any induced set system with m elements may have. Then the previous Lemma is equivalent to $\pi_{\mathcal{F}}(m) \leq \Phi_d(m)$ for all $m \in \{0, \ldots, n\}$.

5.2 Epsilon-nets

For a set system $(X, \mathcal{F})$, a *transversal* is a set of elements $N \subseteq X$ that intersect every set, meaning that $|N \cap S| \geq 1$ for all $S \in \mathcal{F}$. One might naively wonder whether set systems with bounded VC dimension have small transversals? But that is not true. Even for axis-parallel rectangles, one can of course have a rectangle $R \cap X = \{x\}$ for every single element $x \in X$ and hence the only transversal is $N = X$.

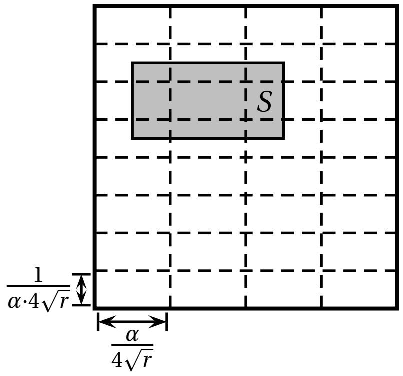
But maybe there are transversals that hit at least the *large* sets. So we call $N \subseteq X$ an *ε-net,* if for all sets $S \in \mathcal{F}$ with $|S| \geq \varepsilon |X|$ one has $|N \cap S| \geq 1$. In words: N has to intersect every set that contains at least an ε-fraction of elements of the universe.

For the sake of argument, suppose we are aiming for a $\frac{1}{r}$-net for a set system $(X,\mathcal{F})$. Then it is an easy exercise to argue that if we sample $O(r \cdot \log|\mathcal{F}|) \leq O(rd\log(n))$ elements uniformly at random, then any $\frac{1}{r}$-large set will be hit at least once. But this leaves us wondering whether a dependence on n is really needed?

Question. *Is it possible to find an ε-net of size $f(\varepsilon, \dim(\mathcal{F}))$?*

We can also allow to give weights to the elements. So, let μ be a *probability measure* on X. Then an ε-net must have the property that $(S \in \mathcal{F}$ with $\mu(S) \geq \varepsilon) \Rightarrow |N \cap S| \geq 1$. In fact, the notion of VC dimension and ε-nets then also make sense if X is an infinite (say compact) set and $\mathcal{F}$ is a family (again compact) subsets of X.

To get some intution we want to study the continuous version of our previous setting. We set $X := [0,1]^2$ and $\mathcal{F} := \{R \subseteq [0,1]^2 \mid R \text{ is axis-parallel rectangle}\}$ and let μ be the uniform measure on $[0,1]^2$. Suppose we are aiming to intersect every rectangle S of size $\mu(S) \geq \frac{1}{r}$ where $r \in 2^{\mathbb{N}}$. Here is one possible line of attack. Take $\alpha \in 2^{\mathbb{Z}}$ with $\frac{1}{4\sqrt{r}} \leq \alpha \leq 4\sqrt{r}$ and partition $[0,1]^2$ into cells of width $\frac{\alpha}{4\sqrt{r}}$ and height $\frac{1}{\alpha \cdot 4\sqrt{r}}$. Every $S \in \mathcal{F}$ with $\mu(S) \geq \frac{1}{r}$ will fully contain one of these cells (for the right α). Here α gives the right ratio of width over height of the target rectangles that we try to hit.



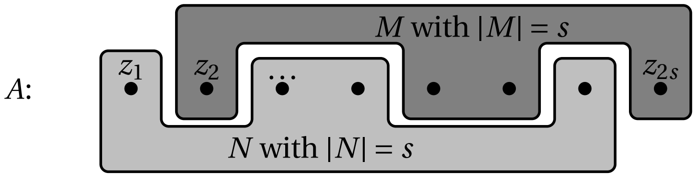
There are only $O(r\log(r)) \leq \text{poly}(r)$ many such cells. If we now sample N as $O(r\log(r))$ many points from X at random, then we miss a cell only with probability $\frac{1}{\text{poly}(r)}$ and hence are likely to hit every of the cells and hence also every $\frac{1}{r}$-large rectangle. Hence there are indeed $\frac{1}{r}$-nets for axis-parallel rectangles of size $O(r\log(r))$.

This reasoning obviously uses a lot of the geometry of the set system. But we will now see the random sampling still works for general set systems of bounded

VC dimension. We present the beautiful probabilistic proof of the ε-net Theorem. As Matousek [Mat02] writes it: *one might be tempted to believe that it works by some magic.*

Theorem 5.3 (ε-Net Theorem)**.** *Let $(X, \mathcal{F})$ be a set system with VC dimension d and let μ be a measure on X. Then a uniform sample N of $O(dr\ln(r))$ many elements is a $\frac{1}{r}$-net with probability at least $1/2$.*

Proof. We can delete small sets from our set family and hence assume that $\mu(S) \geq \frac{1}{r}$ for every $S \in \mathcal{F}$. Let C be a large enough constant and set $s := Cdr\ln(r)$ be our target sample size. However, we do the sampling process in a non-intuitive way. First we sample *twice* as many points as we actually need. More precisely, we sample $z_1, \ldots, z_{2s} \in X$ independently and possibly with repetition according to μ and set $A := \{z_1, \ldots, z_{2s}\}$. Then we pick a uniform sample $N \subseteq A$ of exactly $|N| = s$; the remaining elements are denoted by $M := A \setminus N$.



The expected number of samples for a set S is then $\mathbb{E}[|S \cap N|] \geq \frac{s}{r} = Cd\ln(r)$ (counting elements with multiplicity if they have been sampled several times). So we set a value of $k := \frac{s}{2r}$ which is half of the expectation. Now we define two events that are going to be crucial:

$$\begin{aligned} \mathcal{E}_0 &:= \text{ There is an } S \in \mathcal{F} \text{ with } N \cap S = \emptyset \\ \mathcal{E}_1 &:= \text{ There is an } S \in \mathcal{F} \text{ with } (N \cap S = \emptyset \text{ and } |M \cap S| \geq k) \end{aligned}$$

First we show that the probabilities for both events are close:
Claim I. One has $\frac{1}{2}\Pr[\mathcal{E}_0] \leq \Pr[\mathcal{E}_1] \leq \Pr[\mathcal{E}_0]$.
Proof of claim. The inequality $\Pr[\mathcal{E}_1] \leq \Pr[\mathcal{E}_0]$ is clear because $\mathcal{E}_1$ emerges from $\mathcal{E}_0$ by adding an extra condition. For the 1st inequality it suffices to prove that $\Pr[\mathcal{E}_1 \mid N] \geq \frac{1}{2}\Pr[\mathcal{E}_0 \mid N]$ for every fixed sample N. In fact, it suffices to consider the case where N is not an $\frac{1}{r}$-net as otherwise $\Pr[\mathcal{E}_1 \mid N] = 0 = \Pr[\mathcal{E}_0 \mid N]$. Then fix any set $S^* \in \mathcal{F}$ with $N \cap S^* = \emptyset$. Note that even if N is fixed, M is still a uniform sample of s elements from μ (with repetition). Then $Y := |M \cap S^*|$ is the sum of independent 0/1 random variables (again considering M as a multiset) and $\mathbb{E}[Y] \geq 2k = Cd\ln(r)$. Then using Chernov bound II we get

$$\Pr[\mathcal{E}_1 \mid N] \geq \Pr[|M \cap S^*| \geq k] \geq 1 - \Pr\Big[Y < \frac{1}{2}\mathbb{E}[Y]\Big] \geq 1 - \exp\Big(-\frac{1}{8}\mathbb{E}[Y]\Big)\Big) \geq \frac{1}{2}$$

for C large enough. □

Claim II. For every fixed choice of A one has $\Pr[\mathcal{E}_1 \mid A] < \frac{1}{4}$.

Proof of claim. In principle, the proof consists of the usual combination of concentration plus union bound. The crucial question though is: what sets do we need to take into account. Once A is fixed it suffices to consider the induced set system $(A, \mathcal{F}_{|A})$. This set system still has VC dimension at most d, hence it only has $|\mathcal{F}_{|A}| \leq \Phi_d(2s) \leq (\frac{12s}{d})^d$ many sets[1]. Now for every $S \in \mathcal{F}_{|A}$ one has

$$\begin{aligned}\Pr[N \cap S = \emptyset \text{ and } |M \cap S| \geq k] &\overset{(*)}{\leq} \frac{\binom{2s-k}{s}}{\binom{2s}{s}} \leq \Big(1 - \frac{k}{2s}\Big)^s \leq \exp(-k/2) \\ &= \exp\Big(-\frac{C}{4} d \ln(r)\Big) = r^{-Cd/4}\end{aligned}$$

In $(*)$ we use the following reasoning: the probability $\Pr[N \cap S = \emptyset$ and $|M \cap S| \geq k]$ can only be non-zero if $|A \cap S| \geq k$. And in that case to get $N \cap S = \emptyset$ it needs to happen that each of the s samples for N does not come from the $A \cap S$. Now we can apply the union bound to get.

$$\Pr[\mathcal{E}_1 \mid A] \leq \Phi_d(2s) \cdot r^{-Cd/4} \leq \Big(\frac{6 \cdot 2Cdr\ln(r)}{d}\Big)^d \cdot r^{-Cd/4} \leq (12Cdr^2 \cdot r^{-C/4})^d < \frac{1}{4}$$

if we pick C large enough.

Now we can finish the proof. Combining both proven claims we have $\Pr[\mathcal{E}_0] \leq 2\Pr[\mathcal{E}_1] < 2 \cdot \frac{1}{4} = \frac{1}{2}$. □

We want to briefly mention a notion that is closely related to ε-nets. We say that a set $A \subseteq X$ is an *ε-approximation* for a set system if $|\frac{|A \cap S|}{|A|} - \mu(S)| \leq \varepsilon$. In particular a $\frac{\varepsilon}{2}$-approximation is also an ε-net. One can prove that there is always a $\frac{1}{r}$-approximation of size $O(dr^2\ln(r))$.

5.3 Dual set systems

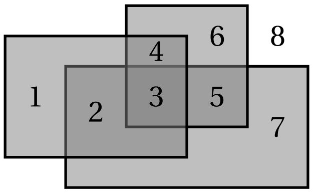
For a set system $(X, \mathcal{F})$, the *incidence matrix* is the matrix $\boldsymbol{A} \in \{0,1\}^{\mathcal{F} \times X}$ with entries

$$A_{S,i} = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases} \quad \forall S \in \mathcal{F} \ \forall i \in X.$$

[1] Here can use the following estimate.
Claim. For $d, m \in \mathbb{N}$ with $m \geq 2d$ one has $\Phi_d(m) \leq (\frac{6m}{d})^d$.
Proof. Simply bound $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \leq (d+1) \cdot \binom{m}{d} \leq (d+1) \cdot (\frac{em}{d})^d \leq 2^d \cdot (\frac{em}{d})^d \leq (\frac{6m}{d})^d$. Here we use that $\binom{m}{i} \leq \binom{m}{i+1}$ for all $i \leq m/2$ as well as the generous bound of $d+1 \leq 2^d$ for $d \geq 1$.

The *dual set system* $(X^*, \mathcal{F}^*)$ is the system that we obtain by reversing the roles of elements and set. Formally, we have elements $X^* := \mathcal{F}$ and $\mathcal{F}^* = \{S_x \mid x \in X\}$ with sets $S_x := \{S \in \mathcal{F} \mid x \in S\}$. The incidence matrix of $(X^*, \mathcal{F}^*)$ is the transpose of the incidence matrix of $(X, \mathcal{F})$. The *dual shatter function* $\tau^*_{\mathcal{F}}(m)$ is the shatter function of the dual set system. In particular $\tau^*_{\mathcal{F}}(m)$ is the maximum number of *equivalence classes* that can be formed by selecting m sets from $\mathcal{F}$.

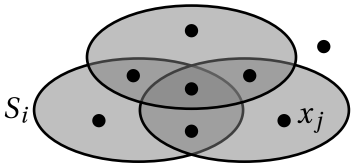
In our example of $(X, \mathcal{F})$ with $X = [0,1]^2$ and $\mathcal{F}$ being all axisparallel rectangles we have $\tau^*_{\mathcal{F}}(3) = 8$ because with 3 rectangles one can indeed get 8 equivalence classes:



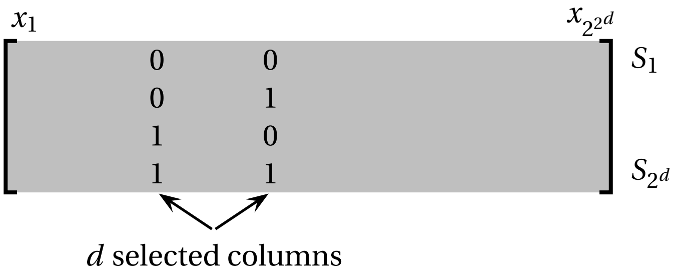
One can prove that the VC dimensions of a dual pair of set systems is related:

Lemma 5.4. *The dual set system satisfies* $\dim(\mathcal{F}^*) \leq 2^{\dim(\mathcal{F})}$.

Proof. Suppose that $\dim(\mathcal{F}^*) \geq 2^d$. Then there are sets $S_1, \ldots, S_{2^d} \in \mathcal{F}$ whose Venn diagram has all the 2^{2^d} many possible equivalence classes. Let $x_1, \ldots, x_{2^{2^d}} \in X$ be points such that we pick one point from each equivalence class.



Consider the $2^d \times 2^{2^d}$ *incidence matrix* of those elements and sets. Then this matrix contains *all* possible 0/1 columns with 2^d entries.



Then in particular it will contain the d many columns that correspond to d shattered elements. Hence $\dim(\mathcal{F}) \geq d$. □

5.4 Discrepancy of set systems

For set system $(X,\mathcal{F})$, we define a *coloring* as a map $\chi : X \to \{-1,1\}$ that "colors" every element with either -1 or $+1$. The *discrepancy* of that coloring is the maximum *in-balance* of any set, i.e.

$$\text{disc}(\mathcal{F},\chi) := \max_{S\in\mathcal{F}} |\chi(S)|$$

where we abbreviate $\chi(S) = \sum_{j\in S}\chi(j)$. Then the discrepancy of the whole set system is defined the discrepancy of the best coloring:

$$\text{disc}(\mathcal{F}) := \min_{\chi:X\to\{-1,1\}} \max_{S\in\mathcal{F}} |\chi(S)|.$$

The following is a folklore result:

Lemma 5.5. *For a set system $(X,\mathcal{F})$ with $|X| = n$ elements and $|\mathcal{F}| = m$ sets one has disc$(\mathcal{F}) \leq O(\sqrt{n\log(m)})$.*

Proof. Pick a uniform random coloring $\chi : X \to \{-1,1\}$. Then for each set $S \in \mathcal{F}$ one has $\Pr[|\chi(S)| > \lambda\sqrt{|S|}] \leq 2\exp(-\lambda^2/2)$ using standard concentration bounds. Then setting $\lambda := \Theta(\sqrt{\log m})$ and applying the union bound gives the claim. □

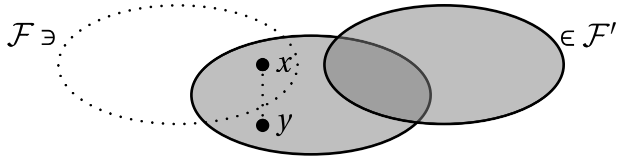
For example if $m = n$, this gives a $O(\sqrt{n\log(n)})$ bound. On the other hand, for example for a random set system one can show that disc$(\mathcal{F}) \geq \Omega(\sqrt{n})$. It takes some work to remove the extra $\sqrt{\log n}$-term in the upper bound, but this is indeed possible as we will see with Spencer's Theorem in a later chapter.

For now we want to bring our attention back to set systems $(X,\mathcal{F})$ with bounded discrepancy. It turns out that for those one can even improve the exponent of $\frac{1}{2}$ to $\frac{1}{2} - \frac{1}{2d}$ where d is the dual VC dimension.

We say that two elements $x, y \in X$ *cross* a set S, if $|\{x, y\} \cap S| = 1$.

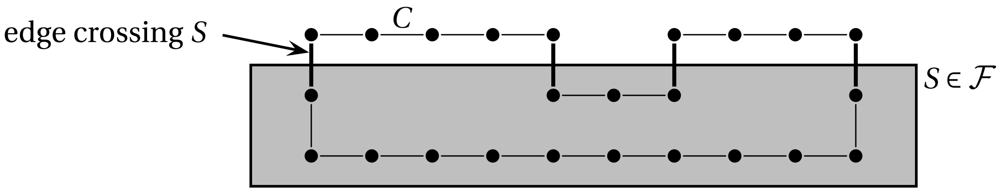
Lemma 5.6. *Let $(X,\mathcal{F})$ be a set system with $|X| = n$, $|\mathcal{F}| = m$ and dual VC dimension d. Then there is a pair $x, y \in X$ that crosses at most $O(m \cdot \frac{\log(n)}{n^{1/d}})$ many sets.*

Proof. Let cross$_{\mathcal{F}}(x, y)$ the number of sets that $\{x, y\}$ crosses in $\mathcal{F}$. Sample each set from $\mathcal{F}$ independently into $\mathcal{F}' \subseteq \mathcal{F}$ with probability $p := \min\{\frac{n^{1/d}}{8m}, 1\}$. The idea of the proof is to use that cross$_{\mathcal{F}'}(x, y)$ is a good enough approximation to cross$_{\mathcal{F}}(x, y)$ (up to some error). First note that $\mathbb{E}[|\mathcal{F}'|] = pm \leq \frac{1}{8}n^{1/d}$ and with probability at least 1/2 one has $|\mathcal{F}'| \leq \frac{1}{4}n^{1/d}$. If we consider the *Venn diagram* formed by $\mathcal{F}'$, it will have at most $\tau_{\mathcal{F}^*}(|\mathcal{F}'|) \leq (2|\mathcal{F}'|)^d < n$ many equivalence classes[2]. In particular there have to be two elements $x, y \in X$ in the same equivalence class. Note that $\text{cross}_{\mathcal{F}'}(x, y) = 0$. If $p = 1$, then $\mathcal{F}' = \mathcal{F}$ and we are done. So, suppose that $p = \frac{1}{8m} n^{1/d}$. We will see that with high probability, any such pair will satisfy the claim.



In fact, consider a pair $x, y \in X$ with $\text{cross}_{\mathcal{F}}(x, y) \geq m\frac{32\ln(n)}{n^{1/d}}$ many sets in the original set system $\mathcal{F}$. Let $Y_S \in \{0, 1\}$ be the indicator random variable telling whether $S \in \mathcal{F}'$. Then we can write let $\text{cross}_{\mathcal{F}'}(x, y) = \sum_{S \in \mathcal{F}: |\{x,y\} \cap S| = 1} Y_S$. Note that $\mathbb{E}[\text{cross}_{\mathcal{F}'}(x, y)] \geq p \cdot m\frac{\log(n)}{n^{1/d}} = 4\ln(n)$ and hence we have $\Pr[\text{cross}_{\mathcal{F}'}(x, y) = 0] \leq \exp(-\mathbb{E}[\text{cross}_{\mathcal{F}'}(x, y)]) \leq \frac{1}{n^4}$. As there are at most n^2 many pairs, the pair (x, y) will have a low crossing number with high probability. □

Lemma 5.7. *Let $(X, \mathcal{F})$ be a set system with $|X| = n$, and dual VC dimension $d \geq 2$. Then there exists a Hamiltonian cycle C on X that crosses every set $S \in \mathcal{F}$ at most $O_d(n^{1-1/d} \cdot \log(n))$ times.*



Proof. First observe that it suffices to find a *spanning tree* T that crosses every set at most $K := \Theta_d(n^{1-1/d} \cdot \log(n))$ times. Then we can use the following standard argument: double the edges of T to obtain a connected subgraph with even degrees. Then there is an Euler tour visiting every edge exactly once. Then shortcut that Euler tour by skipping revisited nodes. We obtain a Hamiltonian cycle that crosses every set at most $2K$ times. Now to the argument how to find the spanning tree.

Let $\mathcal{F} = \{S_1, \ldots, S_m\}$ be the sets in the set system. We give each set i an initial weight of $w_i := 1$. We start the subgraph $T := (X, \{\emptyset\})$, which will be a forest at any time and finish as a spanning tree. Now we call the previous lemma to obtain a pair $\{x, y\}$ crossing at most $O(m \cdot \frac{\log(n)}{n^{1/d}})$ many edges. Actually the lemma also works if the sets are weighted (as dublicating a set does not change the VC

[2] **Claim.** For all $m, d \in \mathbb{N}$ one has $\Phi_d(m) \leq (2m)^d$. **Proof.** $\Phi_d(m) = \sum_{i=0}^{d} \binom{n}{d} \leq \sum_{i=0}^{d} n^i \leq (2n)^d$.

dimension). So we can always get a pair $\{x,y\}$ crossing sets of weight at most $O(\|\boldsymbol{w}\|_1 \cdot \frac{\log(n)}{n^{1/d}})$. We add the edge (x,y) to T and *double* the weight of all the sets that were crossed by the edge. In the next iteration (and in any further iteration) we repeat the argument, just that we do not apply the previous Lemma to all points, but we pick a subset $Y \subseteq X$ that has one point from every connected component of T. The total weight of sets increases by at most a factor of $1 + \frac{O(\log|Y|)}{|Y|^{1/d}}$ per iteration. At the end of this prodedure, we have selected $n-1$ pairs and T is a spanning tree. Let $\text{cross}(S,T) := |\{\{x,y\} \in T \mid |S \cap \{x,y\}| = 1\}|$ be the number of times that edges in T cross S and let $S^* \in \mathcal{F}$ be the set that maximizes $\text{cross}(S^*,T)$. If we analyze the final weights w_S of all the sets, then[3]

$$\begin{aligned}
2^{\text{cross}(S^*,T)} &\overset{(*)}{\leq} \sum_{S\in\mathcal{F}} w_S \overset{(**)}{\leq} |\mathcal{F}| \cdot \prod_{i=1}^{n} \Big(1 + \frac{O(\log n)}{(n-i)^{1/d}}\Big) \\
&\leq |\mathcal{F}| \cdot \exp\Big(O(\log n) \cdot \underbrace{\sum_{i=1}^{n} \frac{1}{(n-i)^{1/d}}}_{\leq O(n^{1-1/d}) \text{ as } d\geq 2}\Big) \\
&\leq \exp\Big(O_d(\log(n)) + O(\log n) \cdot n^{1-1/d}\Big)
\end{aligned}$$

Here we use in $(*)$ that the weight of S^* was doubled $\text{cross}(S^*,T)$ many. We can bound $|\mathcal{F}| \leq n^{\dim(\mathcal{F})} \leq n^{2^d} = 2^{O_d(\log(n))}$, which only results in a lower order term. □

The argument in the proof to maintain weights that increase exponentially with any violation belongs to a very general technique called *multiplicative weight update method* with numerous applications in machine learning and theoretical computer science.

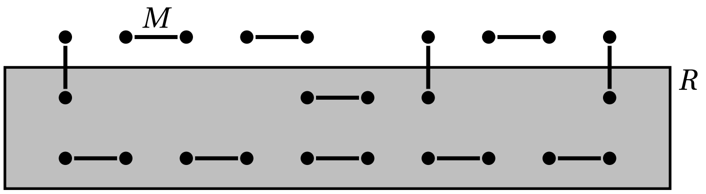
The trick is that the Hamiltonian path allows us to pick the random coloring in a smarter way so that the standard deviation per set is a lot smaller than $\sqrt{n}$.

Theorem 5.8. *Let $(X,\mathcal{F})$ be a set system with $|X| = n$ and dual VC dimension d. Then disc$(\mathcal{F}) \leq O_d(n^{1/2-1/(2d)} \cdot \log(n))$.*

Proof. After possibly dropping an element we may assume that n is even. Let C be the Hamiltonian cycle from the previous Lemma. By picking every 2nd edge of C, we can get a *perfect matching* $M \subseteq C$ crossing any set S at most $k := \Theta(n^{1-1/d}\log(n))$ times. Now we pick again a random coloring $\chi : X \to \{-1,1\}$. But

[3] Here note that if n is a power of 2, then $\sum_{i=1}^{n} \frac{1}{(n-i)^{1/d}} \leq \sum_{k=1}^{\log_2(n)} \frac{n}{2^k} \cdot (\frac{1}{n/2^k})^{1/d} \leq n^{1-1/d} \sum_{k\geq 1} (\frac{2^{1/d}}{2})^k$ and the latter sum converges. The general bound can be obtained by rounding n up to the nearest power of 2.

differently from before, we consider each edge $\{u,v\} \in M$ and with probability $1/2$ we color $\chi(u) = 1, \chi(v) = -1$ and otherwise we color $\chi(u) = -1, \chi(v) = 1$.



If we now consider the discrepancy $\chi(S)$, then every edge completely inside of S or completely outside of S contributes 0 to the discrepancy. Still $\mathbb{E}[\chi(S)] = 0$. The crucial observation is that the standard deviation of $\chi(S)$ is now bounded by $\sqrt{k}$ instead of $\sqrt{|S|}$ which could be up to $\sqrt{n}$. Then $\Pr[|\chi(S)| > 10\sqrt{\log(m)}\sqrt{k}] \leq \frac{1}{2m}$ by the Chernov bound and via the union bound we can conclude that

$$\mathrm{disc}(\mathcal{F}) \leq O\Big(\sqrt{k \log m}\Big) \leq O\Big(\sqrt{d}\sqrt{\log(n)} \cdot n^{1/2-1/(2d)}\sqrt{\log(n)}\Big)$$

which gives the claimed bound. Here we plug in the bound on k and $m \leq \Phi_d(n) \leq O(n^d)$. □

5.5 Exercises

Exercise 5.1. Let $(X,\mathcal{F})$ be a set system with $|X| = n$ elements and VC dimension d. We have seen in the ε-net theorem that there is a $\frac{1}{r}$-net whose size only depends on r and d and we know that a $O_{d,r}(\log n)$ bound is simple. Now we wan to give a intermediate argument without the epsilon-net theorem that gives a $O_{d,r}(\log\log n)$ size $\frac{1}{r}$-net.

(i) Sample $O(rd\log(n))$ elements $Y \subseteq X$ uniformly (with repetition). Show that with high probability any $\frac{1}{2r}$-net for the induced set system $(Y,\mathcal{F}_{|Y})$ is a $\frac{1}{r}$-net for $\mathcal{F}$.

(ii) Show that there is a $O(rd\log(|Y|))$-size $\frac{1}{2r}$-net for $\mathcal{F}_{|Y}$.

Exercise 5.2. Let $(X,\mathcal{F})$ be the set system with $X = [0,1]^2$ and $\mathcal{F} = \{R \subseteq [0,1]^2 \mid R \text{ is axis-parallel rectangle}\}$. Prove without using Lemma 5.4 that the dual shatter function satisfies $\tau_{\mathcal{F}^*}(m) \leq O(m^2)$ and the dual VC dimension is bounded by $\dim(\mathcal{F}^*) \leq O(1)$.

Exercise 5.3. Prove that there exists a set system with n elements and n sets so that the VC dimension is $\Theta(\log n)$. Solve the problem in the spirit of this class!

Exercise 5.4. For $k \in \mathbb{N}$, let $\mathcal{P}$ be the family of full-dimensional convex polytopes in $\mathbb{R}^2$ with at most k vertices. Prove that $\mathcal{P}$ has a VC-dimension that is bounded by a function of k.

Chapter 6

The Regularity Lemma

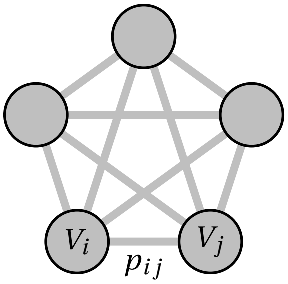
We start again with an example application. Suppose that $G = (V, E)$ is an undirected graph and we want to test a property like "*Does G contain a triangle*"? The only access that we have for the graph is that we can sample a *constant-size* subset U and inspect the induced subgraph $G[U]$. Clearly we are asking for a too strong property — the graph G might only contain a single triangle and that one triangle will likely not be contained in $G[U]$.

But let us say that graphs G and H are *ε-far* if they differ in εn^2 many edges. Maybe there exists a tester that samples constantly many nodes and at least distinguishes the following cases:

- G is triangle-free
- G is ε-far from any triangle-free graph

It still is not clear whether or not this distinction can be made with $O_\varepsilon(1)$-many samples.

Let us first show an algorithm for a particular type of *random* graphs. Suppose we generate $G = (V, E)$ from a distribution $\mathcal{D}$ as follows: we have a partition $V = V_1 \dot\cup \ldots \dot\cup V_k$ for $\frac{1}{\varepsilon} \ll k \leq f(\varepsilon)$ with $|V_1| = \ldots = |V_k|$. For each pair of blocks $i, j \in [k]$ we have a probability p_{ij}. Then every edge (u, v) with $u \in V_i$ and $v \in V_j$ materializes independently with probability p_{ij}.



Could we test triangle-freeness for such random graphs? We can make a crucial observation: if there is a triangle in the graph $H = ([k], E(H))$ with $(i,j) \in E(H) :\Leftrightarrow p_{ij} \geq \varepsilon/2$, then with high probability G will contain $\Omega_\varepsilon(n^3)$ many triangles. On the other hand if there is no triangle in H, then we can destroy all triangles in G by deleting all edges coming from the low density pairs V_i-V_j with $p_{ij} < \varepsilon/2$. Note that the number of such edges will likely not exceed εn^2. In particular if there are $\Omega_\varepsilon(n^3)$ many triangles, it suffices to sample $O_\varepsilon(1)$ many nodes for a positive test.

It appears that this line of arguments only works for random graphs that come from a "density template". The amazingly powerful result that we will see now shows: *Every dense graph has such a density template!*

6.1 The Szemerédi Regularity Lemma

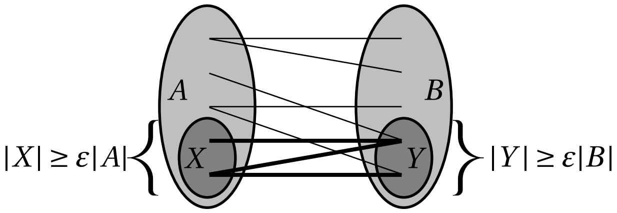
We will now see the statement and proof of the *Regularity Lemma*, proven by Szemerédi with the motivation of resolving questions in combinatorial number theory (we will see one such application in the exercises). We closely follow the excellent exposition in Alon and Spencer [AS16] and refer to the textbook for any further details. First we have to discuss how the definition of the "template" that we promised earlier. Fix an undirected graph $G = (V, E)$ and for disjoint subsets $A, B \subseteq V$, let $e(A,B) := |\{e \in E : |e \cap A| = |e \in B| = 1\}|$ be the number of edges going between A and B. The *density* of the pair is the quantity

$$d(A,B) := \frac{e(A,B)}{|A| \cdot |B|}$$

That means $0 \leq d(A,B) \leq 1$ is the fraction of all possible edges between A and B that exist in G. For a constant $\varepsilon > 0$, we say that the pair (A,B) is ε*-regular* if

$$|d(A,B) - d(X,Y)| \leq \varepsilon \quad \forall X \subseteq A, Y \subseteq B \text{ with } |X| \geq \varepsilon|A| \text{ and } |Y| \geq \varepsilon|B|.$$

In words: for a regular pair, the edge density should be approximately the same for every subsets $X \subseteq A$ and $Y \subseteq B$ of a constant fraction of nodes.



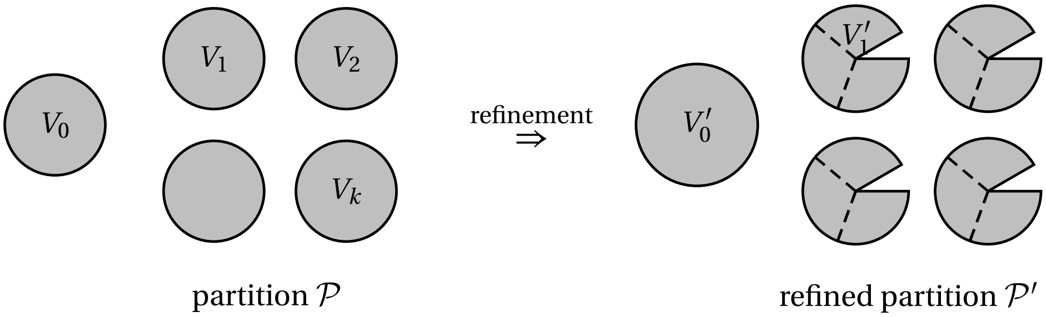
We can now come to the crucial definition:

Definition 1. We call a partition $V = V_0 \dot\cup V_1 \dot\cup \ldots \dot\cup V_k$ an *equipartition* if $|V_1| = \ldots = |V_k|$. Here V_0 is called the *exceptional set*. We call an equipartition *ε-regular*, if all pairs (V_i, V_j) except at most εk^2 are ε-regular and the size of the exceptional set is bounded by $|V_0| \leq \varepsilon |V|$.

Then the main theorem for this chapter is the following:

Theorem 6.1 (Regularity Lemma, Szemerédi 1978)**.** *For every $\varepsilon > 0$, there is a constant $T = T(\varepsilon)$ so that every graph with $|V| \geq T$ vertices has an ε-regular partition $\mathcal{P} = (V_0, \ldots, V_k)$ with $\frac{1}{\varepsilon} \leq k \leq T$.*

The proof idea is simple: we start with any partition $\mathcal{P}$ with $k := \frac{1}{\varepsilon}$ blocks. Then we find a *refinement* $\mathcal{P}'$ that is more regular. A refinement $\mathcal{P}'$ of $\mathcal{P}$ is a partition so that every block of $\mathcal{P}$ is the disjoint union of some blocks of $\mathcal{P}'$. Here we are interpreting V_0 as $|V_0|$ many separate blocks of singleton nodes. In particular that means in order to obtain a refinement, we can always move a small number of nodes into the exceptional set V_0'.



The crucial ingredient is how to measure the regularity so that it can be monotonically improved. For disjoint sets $U, W \subseteq V$ in an n-node graph we define a quantity

$$q(U,W) := \frac{|U| \cdot |W|}{n^2} \cdot d(U,W)^2$$

If $\mathcal{U}$ is a partition of U and $\mathcal{W}$ is a partition of W, then we denote

$$q(\mathcal{U},\mathcal{W}) := \sum_{\substack{U' \in \mathcal{U} \\ W' \in \mathcal{W}}} q(U',W')$$

as the weighted average squared density. It will be useful to introduce a random variable $Z \sim \mathcal{D}(\mathcal{U},\mathcal{W})$ as follows: pick uniform random elements $u \in U$ and $w \in W$. Then set $Z := d(U',W')$ where $u \in U' \in \mathcal{U}$ and $w \in W' \in \mathcal{W}$. In other words, Z gives the average density between blocks of $\mathcal{U}, \mathcal{W}$ where blocks are picked proportional to their number of nodes. The usefulness of this random variable is that

$$q(\mathcal{U},\mathcal{W}) = \frac{|U|\cdot|W|}{n^2}\cdot\mathbb{E}[Z^2] \qquad (*) \tag{6.1}$$

Next, if $\mathcal{P} = (V_0, V_1, \ldots, V_k)$ is a partition with an exceptional set V_0, then we define $q(\mathcal{P}) := \sum_{\text{blocks } U,W \text{ of } \mathcal{P}} q(U,W)$, where the sum is over the $\binom{k+|V_0|}{2}$ many unordered pairs of blocks, counting each singleton in V_0 as one block. The quantity $q(\mathcal{P})$ is called the *index* of the partition. Again, $q(\mathcal{P})$ can be seen as the weighted average of squared densities of its partitions. As the densities are in $[0,1]$ and the sum of the weights is at most $1/2$ one has $0 \leq q(\mathcal{P}) \leq \frac{1}{2}$. In fact, as long as the partition is not regular, we will be able to find refinements that increase $q(\mathcal{P})$. We need a crucial lemma that shows the following: (i) refining can only increase the value $q(\mathcal{P})$; (ii) an irregular pair can be used to get a refinement that strictly increases $q(\mathcal{P})$. The improvement comes from the strict convexity of the function $x \mapsto x^2$,

Lemma 6.2. *The following holds:*

i) Let $U, W \subseteq V$ be disjoint. Let $\mathcal{U}$ and $\mathcal{W}$ be partitions of U and W. Then $q(\mathcal{U},\mathcal{W}) \geq q(U,W)$.

ii) If $\mathcal{P}'$ is a refinement of $\mathcal{P}$, then $q(\mathcal{P}') \geq q(\mathcal{P})$.

iii) Suppose a disjoint pair (U,W) is not ε-regular due to (U_1, W_1) with $U_1 \subseteq U$ and $W_1 \subseteq W$. Then the partition $\mathcal{U} := \{U_1, U \setminus U_1\}$ and $\mathcal{W} := \{W_1, W \setminus W_1\}$ satisfies $q(\mathcal{U},\mathcal{W}) > q(U,W) + \varepsilon^4 \cdot \frac{|U|\cdot|W|}{n^2}$.

Proof. We prove the items separately:

i) We study the random variable $Z \sim \mathcal{D}(\mathcal{U},\mathcal{W})$ that gives the density of a random pair (U', W') of the partitions. Note that $\mathbb{E}[Z] = d(U,W)$ is just the overall edge density. Then

$$\frac{n^2}{|U||W|}q(\mathcal{U},\mathcal{W}) \overset{(6.1)}{=} \mathbb{E}[Z^2] \overset{\text{Jensen}}{\geq} \mathbb{E}[Z]^2 = d(U,W)^2 \overset{\text{Def } q(U,W)}{=} \frac{n^2}{|U|\cdot|W|}q(U,W)$$

ii) Follows from *i*).

iii) Recall that

$$\operatorname{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \frac{n^2}{|U|\cdot|W|}(q(\mathcal{U},\mathcal{W}) - q(U,W)),$$

hence to show the claim we only need to lower bound the variance of Z. But if we pick the irregular pair (U_1, W_1), then we get a density that is different from $\mathbb{E}[Z] = d(U,W)$. Hence

$$\mathrm{Var}[Z] = \mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right] = \underbrace{\Pr_{u\sim U, w\sim W}[u \in U_1, w \in W_1]}_{\geq \varepsilon^2} \cdot \underbrace{(d(U_1,W_1) - d(U,W))^2}_{\geq \varepsilon^2} \geq \varepsilon^4$$

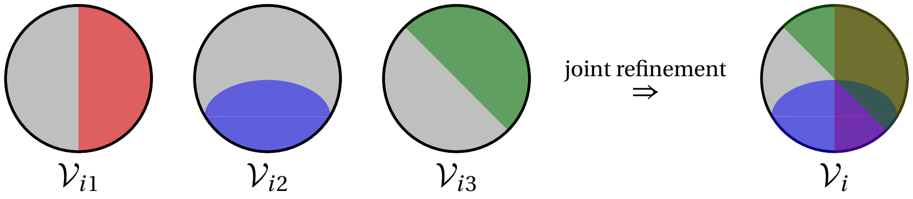
Combining this, we get the claim.

□

Now we can come to the central part of the proof for the Regularity Lemma in which we show that if the current partition is not ε-regular, then we can find a refinement $\mathcal{P}'$ with $q(\mathcal{P}') \geq q(\mathcal{P}) + \frac{1}{2}\varepsilon^5$. Here the size of the exceptional set only increases marginally:

Lemma 6.3. *Let $0 < \varepsilon \leq \frac{1}{4}$. Suppose $\mathcal{P} = \{V_0, \dots, V_k\}$ is an equipartition with $|V_0| \leq \varepsilon n$ that is not ε-regular. Then there exists a refinement $\mathcal{P}' = \{V_0', \dots, V_\ell'\}$ that is an equipartition with $k \leq \ell \leq k4^k$ parts satisfying $|V_0'| \leq |V_0| + \frac{n}{2^k}$ and $q(\mathcal{P}') \geq q(\mathcal{P}) + \frac{1}{2}\varepsilon^5$.*

Proof. Consider a pair (V_i, V_j) with $1 \leq i < j \leq k$. If this is a regular pair, then we set $\mathcal{V}_{ij} := \{V_i\}$ and $\mathcal{V}_{ji} := \{V_j\}$. If the pair is not regular and $U \subseteq V_i$ and $W \subseteq V_j$ is the irregular part, then we set $\mathcal{V}_i := \{U, V_i \setminus U\}$ and $V_{ji} := \{W, V_j \setminus W\}$. Now, let $\mathcal{V}_i$ be the joint refinement of the partitions $\{\mathcal{V}_{ij}\}_{j\in[k]\setminus\{i\}}$ that have 1 or 2 parts. In other words $\mathcal{V}_i$ is the *Venn diagram* consisting of at most 2^{k-1} parts.



Let $\tilde{\mathcal{P}}$ be the partition containing $\mathcal{V}_1, \dots, \mathcal{V}_k$ together with the exceptional set V_0. As $\mathcal{P}$ was not ε-regular, there will be εk^2 many pairs that are irregular and by Lemma 6.2.iii), each one will increase the function q. We can estimate that

$$q(\tilde{\mathcal{P}}) \geq q(\mathcal{P}) + \underbrace{\sum_{(i,j)\text{ irregular}}}_{\varepsilon k^2 \text{ pairs}} \varepsilon^4 \frac{|V_i| \cdot |V_j|}{n^2} \overset{|V_i| \geq \frac{3}{4}\frac{n}{k}}{\geq} q(\mathcal{P}) + \varepsilon^4 \cdot \varepsilon k^2 \cdot \frac{3}{4}\frac{1}{k^2} \geq q(\mathcal{P}) + \frac{1}{2}\varepsilon^5.$$

Here we use $|V_i| = \frac{n-|V_0|}{k} \geq \frac{3}{4}\frac{n}{k}$.

The regularity of $\tilde{\mathcal{P}}$ has increased as desired, but $\tilde{\mathcal{P}}$ is not yet an equipartition. Suppose that $s := |V_1| = \ldots = |V_k|$ was the original size of the blocks in $\mathcal{P}$. Then we split each part of $\tilde{\mathcal{P}}$ into blocks of size $\frac{s}{4^k}$, moving leftover pieces into the exceptional set. Then we end up with $k \cdot 4^k$ many non-exceptional parts and the new exceptional set has size $|V_0'| \leq |V_0| + k2^{k-1} \cdot \frac{s}{4^k} \leq |V_0| + \frac{n}{2^k}$. □

Proof of Regularity Lemma. We begin with an arbitrary partition of the n vertices into $k_0 := \frac{1}{\varepsilon}$ many equal size blocks, requiring to move at most $\frac{1}{\varepsilon} \ll \frac{1}{2}\varepsilon n$ many nodes into the exceptional set. In the ith iteration, as long as the current partition is not ε-regular, we employ Lemma 6.3 and the number of partitions increases from k_i to $k_{i+1} \leq k_i 4^{k_i}$. As $q(\mathcal{P})$ increases by at least $\frac{1}{2}\varepsilon^5$, we terminate after at most $\frac{2}{\varepsilon^5}$ many calls. In each call the size of the exceptional set increases by $\frac{n}{2^{k_i}}$, but as $k_i \geq \frac{1}{\varepsilon}$, the total increase in size is generously bounded by $\frac{1}{2}\varepsilon n$. The argument will work as long as n stays bigger than the bound on k_i. That concludes the proof. □

The reader may have noticed that pessimistically it could happen that for $\Theta(\frac{1}{\varepsilon^5})$ times the number of partitions increases exponentially. In particular our bound on $T(\varepsilon)$ is a *tower of exponents* of height $\Theta(1/\varepsilon^5)$. Surprisingly, there is a complementary result of Gowers that in some graphs every ε-regular partition requires a number of partitions that is a tower of height polynomial in $1/\varepsilon$.

If the analysis is still arcane to the reader, maybe a simplifying thought experiment helps. Suppose we have numbers $x_1, \ldots, x_n \in [0,1]$ that are partitioned into blocks $S_1 \cup \ldots \cup S_k$. Consider the following random variable Z: pick $i \in [n]$ uniformly and let j be the index with $i \in S_j$. then Z is the average of numbers in S_j which is $\mathbb{E}_{\ell \sim S_j}[x_\ell]$. By Jensen's inequality, $\mathbb{E}[Z^2]$ would be non-decreasing if we would split one of the sets S_j into multiple sets. Moreover, if in a constant fraction of sets (weighted by size), a constant fraction of pairs of numbers differ by a constant amount, then the sets can be split into two parts so that the value $\mathbb{E}[Z]^2$ increases by a constant.

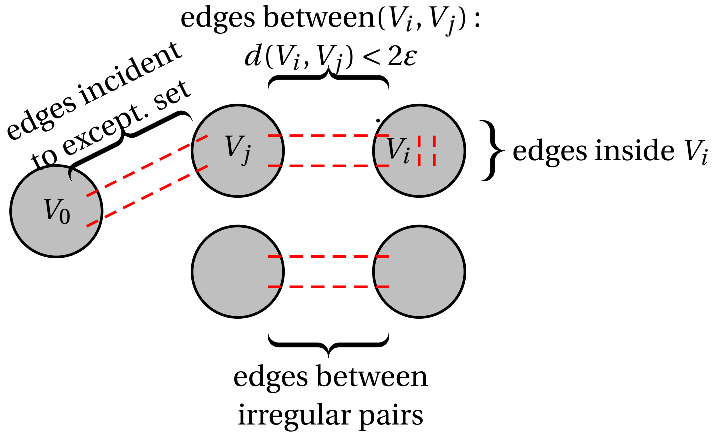
6.2 Application to testing triangle-freeness

We want to come back to our initial application where we wanted to be able to distinguish a triangle-free graph from a graph that is ε-far from being triangle free. It turns out that we can simply test for $O_\varepsilon(1)$ many random triples of nodes whether they form a triangle:

Lemma 6.4. *Let $G = (V, E)$ be a graph so that for every $H \subseteq E$ with $|H| \leq \varepsilon n^2$, $(V, E \setminus H)$ still contains at least one triangle. Then G itself contains δn^3 many triangle with $\delta := \delta(\varepsilon) > 0$.*

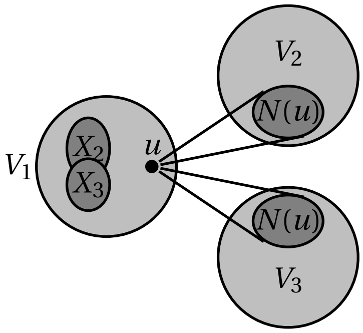
Proof. For the sake of a cleaner notation suppose the assumption is that at least $C\varepsilon n^2$ can be deleted without destroying all triangles, where $C > 0$ is a large constant. We invoke the Regularity Lemma and consider the partition $\mathcal{P} = (V_0, \ldots, V_k)$ that is ε-regular. We delete the following edges:

- All edges incident to the exceptional set V_0.
- All edges inside some block V_i.
- All edges between irregular pairs (V_i, V_j).
- All edges between regular pairs (V_i, V_j) where the density is $d(V_i, V_j) < 2\varepsilon$.



visualization of deleted edges

It is not hard to see that in each of the 4 categories, we delete at most $O(\varepsilon n^2)$ many edges. By assumption the remaining graph still has at least one single triangle. By construction, this triangle is running between partitions say V_1, V_2, V_3 where all pairs (V_i, V_j) are regular and the densities are $d(V_i, V_j) \geq 2\varepsilon$ for $1 \leq i < j \leq 3$. Recall that $|V_1| = |V_2| = |V_3| =: s \geq \frac{3}{4} \cdot \frac{n}{k}$. For $i \in \{2, 3\}$, let $X_i := \{u \in V_1 : |N(u) \cap V_i| \leq \varepsilon |V_i|\}$ be nodes with rather few neighbors. If $|X_i| \geq \varepsilon |V_1|$, then (X_i, V_i) was an ε-irregular part of (V_1, V_i). We call nodes $u \in V_1 \setminus (X_2 \cup X_3)$ *typical*. By regularity we know that the density between neighbors of a typical node $u \in V_1$ is $d(N(u) \cap V_2, N(u) \cap V_3) \geq 2\varepsilon - \varepsilon = \varepsilon$. But every edge between $N(u) \cap V_2$ and $N(u) \cap V_3$ forms a triangle together with u.



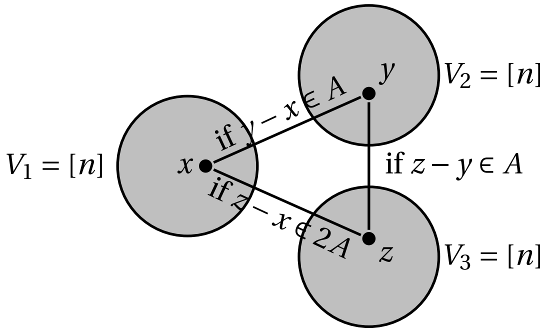
Overall the number of triangles between V_1, V_2, V_3 is at least $(1-2\varepsilon)|V_1|\cdot\varepsilon\cdot\varepsilon|V_2|\cdot\varepsilon|V_3| = \Omega(\varepsilon^3\frac{n^3}{k^3}) = \delta(\varepsilon)\cdot n^3$. □

6.3 Exercises

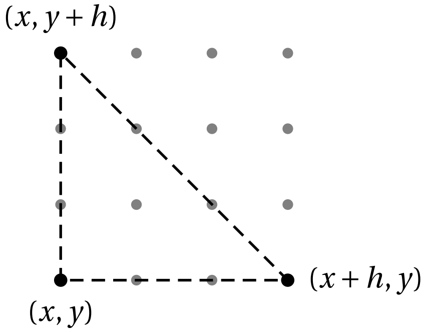
Exercise 6.1 (From Alon & Spencer [AS16])**.** Prove that the following is true. Fix $\varepsilon > 0$ and $r \in \mathbb{N}$. Suppose that $G = (V,E)$ is a graph with $n = |V|$ for which one needs to delete more than $\varepsilon^2 n$ edges to destroy all copies of K_r. Then G contains $C(\varepsilon, r)\cdot n^r$ many copies of K_r.

Exercise 6.2 (From Alon & Spencer [AS16])**.** Prove that the following is true. Fix $\varepsilon > 0$ and a graph H. Suppose that $G = (V,E)$ is a graph with $n = |V|$ for which one needs to delete more than $\varepsilon^2 n$ edges to destroy all copies of H. Then G contains $C(\varepsilon, H)\cdot n^{|V(H)|}$ many copies of H.
Hint. The *number of copies* of H in G is the number of injective maps $f : V(H) \to V(G)$ with $\{u,v\} \in E(H) \Rightarrow \{f(u), f(v)\} \in E(G)$. You can sort of follow the solution of the previous exercise, but note that some vertices of the remaining copy of H may end up in the same block V_i.

Exercise 6.3 (From Alon & Spencer [AS16])**.** Prove that for every $\varepsilon > 0$ there is an $n_0 := n_0(\varepsilon)$ so that every set $A \subseteq \{1,\ldots,n\}$ with $|A| \geq \varepsilon n$ and $n \geq n_0$ contains a 3-term arithmetic progression.
Hint. A 3-term arithmetic progression means that $\{a, a+b, a+2b\} \subseteq A$ for integers $a, b \in \mathbb{N}$. Analyze the triangles in the following graph $G = (V_1 \cup V_2 \cup V_3, E)$ where V_1, V_2, V_3 are copies of $[n]$:



Exercise 6.4. A *corner* in $\mathbb{Z}^2$ are three points $(x, y), (x+h, y), (x, y+h)$.



Prove that for every $\varepsilon > 0$, there is an n big enough so that any subset $A \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ with $|A| \geq \varepsilon n^2$ contains a corner.
Hint. Analyze the graph $G = (L_H \cup L_V \cup L_D, E)$ where L_H are the horizontal lines, L_V are the vertical ones and L_D are the diagonal ones and we have an edge in E if the crossing of the two lines is in A.

Chapter 7

Dependent Random Choice

The method of *dependent random choice* is a smart and slick probabilistic argument to find a small very well connected subgraph in a large graph with high enough average degee. In this chapter we follows the excellent exposition by Alon and Spencer [AS16]. For additional applications see the survey of Fox and Sudakov [FS11].

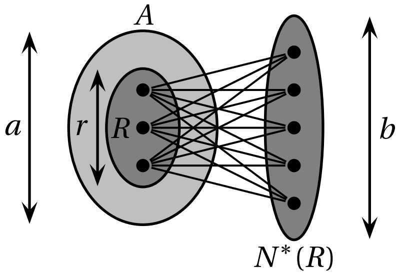
For $U \subseteq V$, we denote $N^*(U) := \bigcap_{v \in U} N(v)$ as the *common neighborhood* of U.

Theorem 7.1. *Let $a, b, n, r \in \mathbb{N}$ and let $G = (V, E)$ be a graph with $|V| = n$ vertices and average degree $d := 2|E|/n$. If*

$$\frac{d^r}{n^{r-1}} \geq b^r + a$$

then G contains a subset $A \subseteq V$ of size $|A| \geq a$ so that every $R \in \binom{A}{r}$ has $|N^*(R)| \geq b$.

The visualization of the claim is as follows:



Proof. Let $T \subseteq V$ be a multi-set of r vertices, picked uniformly at random with repetition. We set $A := \{v \in V \mid T \subseteq N(v)\} = N^*(T)$. Then it is not hard to see that the expected size satisfies

$$\begin{aligned}\mathbb{E}[|A|] &= \sum_{v\in V}\left(\frac{|N(v)|}{n}\right)^r = \frac{1}{n^{r-1}}\sum_{v\in V}\frac{|N(v)|^r}{n} \\ &\overset{\text{Jensen}}{\geq} \frac{1}{n^{r-1}}\Big(\underbrace{\frac{1}{n}\sum_{v\in V}|N(v)|}_{=d}\Big)^r = \frac{d^r}{n^{r-1}}.\end{aligned}$$

For the sampled set A we know by construction that it has at least r common neighbors (namely the nodes in T). But we require b many common neighbors at least for all r-tuples and in general we will have $b \gg r$. So, let $Y := |\{R \in \binom{A}{r} : |N^*(R)| < b\}|$ be the number of r tuples in our sample that have too few common neighbors. For the expected number of such tuples we know that

$$\mathbb{E}[Y] = \sum_{R\subseteq V:|R|=r \text{ and } |N^*(R)|<b} \underbrace{\Pr[T \subseteq N^*(R)]}_{=(|N^*(R)|/n)^r} \leq \binom{n}{r}\cdot\left(\frac{b}{n}\right)^r \leq b^r$$

We use the assumption and conclude that $\mathbb{E}[|A|-Y] \geq a$. In particular we can take the set A and for each tuple $R \in \binom{A}{r}$ with $|N^*(R)| < b$, we can drop one element. We will still be left with a set A' of size $\mathbb{E}[|A'|] \geq a$. This proves the claim. □

Note that the theorem requires that $d \geq n^{1-1/r}$, otherwise the statement is vacuous. In reverse if $d \geq C(a,b,r)\cdot n^{1-1/r}$ then indeed a set $A \subseteq V$ of size $|A| \geq a$ exists where the common neighborhood is at least b for every r-tuple.

As a second remark, consider the case that the graph G consists of two cliques of size $n/2$. With probability $1-\Theta(1)\cdot 2^{-r}$, the sample T contains nodes from each of the two cliques and hence the set A will be empty. In other words, the random variable $|A|$ is not at all well concentrated.

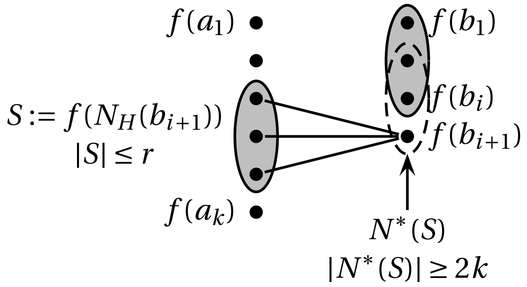
7.1 Turan numbers of bipartite graphs

One classical question in *extremal graph theory* is the following: One is given some "small" graph H. The question is how many edges can a "larger" graph $G = (V,E)$ on n vertices have if H is forbidden to appear as a subgraph. We denote this number by $\text{ex}(n,H)$. Note that formally H *appears as a subgraph* if there is an injective map $f : V(H) \to V(G)$ so that $\{u,v\} \in E(H) \Rightarrow \{f(u), f(v)\} \in E(G)$. For example, we seen the Kövari-Sos-Turan Theorem in Chapter 4.0.2 which shows that a forbidden K_r-subgraph limits the number of edges to $c_r \cdot n^{2-1/r}$. In the new notation that means $\text{ex}(n,K_r) \leq c_r \cdot n^{2-1/r}$. However, this does not fully answer the question for non-complete graphs. For example consider the case of the *k-cycle* denoted by C_k. Then forbidding the K_k also forbids the C_k and hence $\text{ex}(n, C_k) \leq c_k \cdot n^{2-1/k}$. But with the dependent random choice theorem we will be able to prove a better bound!

But first, we need a fairly useful embedding theorem:

Lemma 7.2. *Let $H = (A\dot{\cup}B, F)$ be a bipartite graph with $|A|, |B| \leq k$ and $\deg(b) \leq r\ \forall b \in B$. If $G = (V, E)$ contains $|U| = 2k$ vertices so that for all $S \subseteq U$ with $|S| = r$ one has $|N^*(S)| \geq 2k$, then G contains H as a subgraph.*

Proof. W.l.o.g. suppose that $A = \{a_1, \ldots, a_k\}$ and $B = \{b_1, \ldots, b_k\}$. We pick a map $f : A \cup B \to V$ defining the embedding. We can indeed embedd the left hand side A arbitrarily. We then embedd the vertices $b_1, \ldots, b_k$ of B one by one. Suppose that the embedding for $b_1, \ldots, b_i$ has already been chosen. Then for b_{i+1}, consider the at most r nodes $S := f(N_H(b_{i+1}))$. These have at least $2k$ common neighbors in G. We pick any one as $f(b_{i+1})$ that has not been used so far to embed A or $b_1, \ldots, b_i$. The embedding can look as follows:



□

Theorem 7.3. *Let $H = (A\dot{\cup}B, F)$ be a bipartite graph with $|A|, |B| \leq k$ and $\deg(b) \leq r \forall b \in B$. Then $\text{ex}(n, H) \leq c_k n^{2-1/r}$.*

Proof. Suppose that we have a graph G where the average degree is $d = cn^{1-1/r}$ where we can make c as large as needed. Then $\frac{d^r}{n^{r-1}} = c^r$. If for example $c \geq 4k$, then Theorem 7.1 gives us a set A of size $2k$ so that $S \in \binom{A}{r}$ has $|N^*(S)| \geq 2k$. Then the Embedding Lemma shows that H can be embedded into A. □

In particular this shows that for even k, one has $\text{ex}(n, C_k) \leq c_k n^{3/2}$. Recall that in an earlier chapter we have seen in an exercise that indeed $\text{ex}(n, C_4) = \Theta(n^{3/2})$. Now we have seen that forbidding a 1000-cycle instead of a 4-cycle leads to the same asymptotic upper bound on the number of edges.

7.2 Exercises

Exercise 7.1. Suppose k is odd. How large can $\text{ex}(n, C_k)$ be?

Chapter 8

Existence of Rigid Structures: The Kuperberg-Lovett-Peled Theorem

For a motivating example, note that many algorithms used in theoretical computer science are *randomized*, that means they draw some random bit string $\boldsymbol{x} \sim \{0,1\}^n$ that is used for the computation. Then the correctness (and possibly the running time) depend on that random string. For example, if we want to find a cut $S \subseteq V$ in a graph $G = (V,E)$ with $|\delta(S)| \geq |E|/2$, then a random cut that includes each node independently with probability $1/2$ will be enough in expectation. However, it is more desirable to have a *deterministic* algorithm that always terminates with the correct answer. One trivial way of *derandomizing* the algorithm is by trying out all 2^n possible random strings. Obviously this is terribly wasteful. Often, the randomized algorithm does not actually need n many *independent* random bits. They might already work if the bits are *pairwise independent* or t*-wise independent* for some $t > 2$. In our example of MaxCut, it actually suffices if bits are pairwise independent.

This is our motivation to define an *orthogonal array of strength t* as a subset $T \subseteq \{0,1\}^n$ of $0/1$ strings with the property that for all $1 \leq i_1 < \ldots < i_t \leq n$ and all $a_1, \ldots, a_t \in \{0,1\}$ one has

$$\Pr_{\boldsymbol{x} \sim T}[x_{i_1} = a_1, \ldots, x_{i_t} = a_t] = 2^{-t}. \qquad (*)$$

In other words, if we draw $\boldsymbol{x} \sim T$, then this gives us a t-wise independent random vector. For example $T = \{(0,\ldots,0),(1,\ldots,1)\}$ is a strength-1 orthogonal array, but it does not have strength-2.

In fact one can prove that an orthogonal array of strength t has size $|T| \geq \Omega_t(n^{t/2})$ for larger t. For the other direction it seems highly non-trivial to find a construction giving a strength-t orthogonal array of size, say $n^{O(t)}$. If one picks the set T randomly from $\{0,1\}^n$, then with concentration arguments one can only reason that the condition $(*)$ is *approximately* satisfied. But satisfying it exactly seems hard. Note that there are many related applications where one is looking for a *rigid structure*; for example one can ask for a subset T of permutations so that the distribution of any t indices $(\pi(i_1),\dots,\pi(i_t))$ looks exactly uniform if we draw $\pi \sim T$. A conventional approach for such applications is to look for a suitable *algebraic construction*. But if these are not known to exist or the parameters are not optimal, then until very recently there was no plan B. In this chapter, we want to discuss a more systematic *probabilistic* technique of Kuperberg, Lovett and Peled [KLP17] to show the existence of such rigid structures. We will showcase the technique to prove the following result:

Theorem 8.1. *For any $n,t \in \mathbb{N}$, there exists a strength-t orthogonal array $T \subseteq \{0,1\}^n$ of size $|T| \leq (\frac{cn}{t})^{ct}$ for some constant $c > 0$.*

We would like to remark that for our setting there is indeed an alternative algebraic construction for a strength-t orthogonal array $T \subseteq \{0,1\}^n$ of size $n^{O(t)}$, but that construction provides suboptimal bounds for the slight generalization of $T \subseteq \{0,\dots,q-1\}^n$, while the probabilistic proof that we present here gives tight bounds for all q. However, we restrict our attention to $q=2$ to keep the notation simple.

8.1 A matrix view on orthogonal arrays

We want to rephrase the problem of finding an orthogonal array as a matrix problem. Let $B := \{0,1\}^n$ be the set of all 0/1 strings and define $A := \{(I,\boldsymbol{a}) \mid I \in \binom{[n]}{t}, \boldsymbol{a} \in \{0,1\}^I\}$ as all combinations of t-tuples equipped with bits. We define a matrix $\boldsymbol{M} \in \{0,1\}^{B\times A}$ by letting

$$M_{\boldsymbol{x},(I,\boldsymbol{a})} := \begin{cases} 1 & \text{if } x_i = a_i \quad \forall i \in I \\ 0 & \text{otherwise} \end{cases} \qquad \forall \boldsymbol{x} \in B = \{0,1\}^n \ \forall (I,\boldsymbol{a}) \in A.$$

Then a subset $T \subseteq B$ of rows is a strength-t orthogonal array if and only if it has the same *row average* as the set of all rows, that means

$$\mathop{\mathbb{E}}_{b\sim T}[\boldsymbol{M}_b] = \mathop{\mathbb{E}}_{b\sim B}[\boldsymbol{M}_b].$$

Note that the "constraints" — in this case having all events on t coordinates uniform — are encoded as columns of $\boldsymbol{M}$. A natural idea is to sample T at random. We fix a large enough parameter $N \in \mathbb{N}$ that is our target size for T and set $p := \frac{N}{|B|}$.

Now sample a subset $T \subseteq B$ by including each row independently with probability p. If we abbreviate $\boldsymbol{X} := \sum_{b\in T} \boldsymbol{M}_b$ as the sum of the sampled rows, then

$$\mathbb{E}[\boldsymbol{X}] = \mathop{\mathbb{E}}_{T}\Big[\sum_{b\in T} \boldsymbol{M}_b\Big] = \frac{N}{|B|}\sum_{b\in B} \boldsymbol{M}_b \quad \text{and} \quad \mathop{\mathbb{E}}_{T}[|T|] = p|B| = N.$$

What we need to show is that indeed

$$\Pr\big[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}] \text{ and } |T| = N\big] > 0, \quad (**)$$

meaning that it is indeed possible to hit the expectation. Observe that the above probability will be exponentially small, which does not leave many candidate tools for the analysis. In the following we will give a specialized version of the *central limit theorem* that proves $(**)$. This chapter is organized that we first prove a general statement for any matrix $\boldsymbol{M}$ that satisfies certain properties. In Section 8.4 we will then argue that the matrix $\boldsymbol{M}$ stemming from orthogonal arrays indeed satisfies these properties. Starring at the definition of the matrix $\boldsymbol{M}$ for orthogonal arrays, we can see already some properties that must be useful: (i) we have $|B| \gg |A|$, that means we have a lot more rows than columns; (ii) the matrix is highly symmetric, meaning that there is no row that has a particular importance. In the next section, we formally extract the properties that are needed.

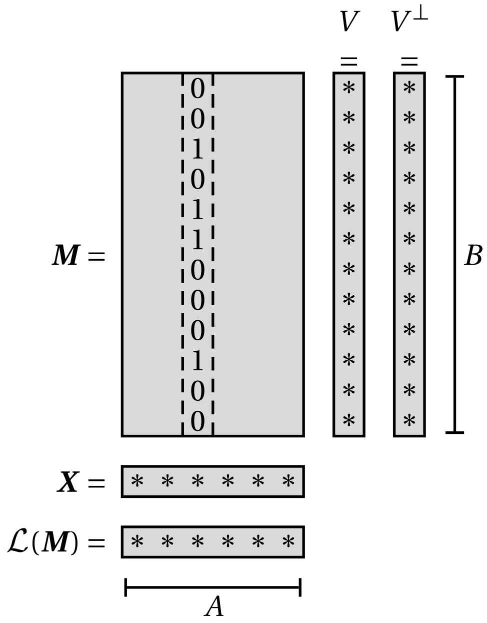
8.2 The Kuperberg-Lovett-Peled Theorem

We will now explain the matrix properties needed for the KLP-Theorem. We should remark that we state the Theorem slightly less general compared to the original paper to keep things more "concrete". Consider an arbitrary matrix $\boldsymbol{M} \in \{0,1\}^{B\times A}$. We denote B as the row indices and A as the column indices. Moreover, $\boldsymbol{M}_b$ with $b \in B$ will denote a row vector and $\boldsymbol{M}^a$ with $a \in A$ will denote a column vector. Our random experiment is to fix a large enough number $N \in \mathbb{N}$ and a probability $p := \frac{N}{|B|}$ and sample each row independently into a subset $T \subseteq B$ with that probability p. For later reference, let $T_b \in \{0,1\}$ be the indicator variable telling whether row b was sampled and let $\boldsymbol{X} := \sum_{b\in B} T_b \boldsymbol{M}_b$ be the sum of the sampled rows.

Let $V := \text{span}\{\boldsymbol{M}^a \mid a \in A\} \subseteq \mathbb{R}^B$ be the span of the column vectors. Note that typically $|A| \ll |B|$, hence the vector space V has a rather small dimension compared to its ambient space. Later, the orthogonal space $V^\perp := \{\boldsymbol{x} \in \mathbb{R}^B \mid \boldsymbol{x} \perp \boldsymbol{x}' \ \forall \boldsymbol{x}' \in V\}$ will also play a role. Let

$$\mathcal{L}(\boldsymbol{M}) := \Big\{\sum_{b\in B} \lambda_b \boldsymbol{M}_b \mid \lambda_b \in \mathbb{Z}\Big\} \subseteq \mathbb{Z}^A$$

be the *lattice* generated by the row vectors of $\boldsymbol{M}$. We can visualize the formats as follows:



Divisibility. In the end, we want to guarantee that there is a subset $T \subseteq B$ of size $|T| = N$ so that $\frac{1}{N}\sum_{b\in T}\boldsymbol{M}_b = \frac{1}{|B|}\sum_{b\in B}\boldsymbol{M}_b$ holds. A relaxed condition is certainly that

$$\mathbb{E}[\boldsymbol{X}] = \frac{N}{|B|}\sum_{b\in B}\boldsymbol{M}_b \overset{!}{\in} \mathcal{L}(\boldsymbol{M})$$

The set of N's that satisfy this, must be of the form $N \in \{c, 2c, 3c, \dots\}$ for some $c > 0$. That particular value c is what we call the *divisibility constant*.

Symmetry. The symmetry condition basically says that "all rows of $\boldsymbol{M}$ have to look the same". More formally, a *symmetry* of the matrix $\boldsymbol{M}$ is a *permutation* $\pi : B \to B$ on the rows so that $(M_{\pi(b),a})_{b\in B} \in V$. In other words, after permuting the row indices according to π, each vector is still in the column space. Note that this is actually more of a property of the vector space V and replacing $\boldsymbol{M}$ by a matrix with the same column space V would not change the set of symmetries.

Lemma 8.2. *Suppose that the column vectors of* $\boldsymbol{M}$ *are linearly independent. Then* $\pi : B \to B$ *is a symmetry if and only if there exists a bijective linear map* $\tau : \mathbb{R}^A \to \mathbb{R}^A$ *with*

$$\boldsymbol{M}_{\pi(b)} = \tau(\boldsymbol{M}_b) \quad \forall b \in B$$

Proof. Well, if that linear function τ exists, then it means that there is an invertible matrix $\boldsymbol{S} \in \mathbb{R}^{A\times A}$ with $\boldsymbol{M}_{\pi(b)} = \boldsymbol{M}_b\boldsymbol{S}$ (considering $\boldsymbol{M}_b$ and $\boldsymbol{M}_{\pi(b)}$ indeed as $1 \times A$

dimensional row vectors). Then in matrix form we can write $(\boldsymbol{M}_{\pi(b)})_{b\in B} = \boldsymbol{MS}$. For the column index $a \in A$ means that $(M_{\pi(b),a})_{b\in B} = \sum_{a'\in A} S_{a,a'}\boldsymbol{M}^{a'} \in V$ which is the desired linear combination. One can also reverse the argument to get the other direction. □

In particular, we will use that symmetries π preserve *linear dependencies*. More precisely, for a vector $\boldsymbol{y} \in \mathbb{R}^B$ one has

$$\sum_{b\in B} y_b \boldsymbol{M}_b = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{0} = \tau\Big(\sum_{b\in B} y_b \boldsymbol{M}_b\Big) = \sum_{b\in B} y_b \tau(\boldsymbol{M}_b) = \sum_{b\in B} y_b \boldsymbol{M}_{\pi(b)}.$$

As condition for the KLP-Theorem, we will require that for all rows $b_1, b_2 \in B$, there exists a symmetry π of V so that $\pi(b_1) = b_2$.

8.2.1 The theorem

Now we can formally state the KLP-Theorem[1]:

Theorem 8.3 (KLP-Theorem)**.** *Let $\boldsymbol{M} \in \{0,1\}^{B\times A}$ be a matrix with $V := \text{span}\{\boldsymbol{M}^a \mid a \in A\}$ and let $K \geq |A|$ be a parameter that is at least a large enough constant. Assume that $\boldsymbol{M}$ has the following properties:*

- (I) Divisibility: *c is the divisibility constant of V.*
- (II) Local Decodability*: There is an integer $m \in \{1,\dots,K\}$ so that for each column $a \in A$ there is a vector $\boldsymbol{y}^a \in \mathbb{Z}^B$ with $\|\boldsymbol{y}^a\|_1 \leq K$ and $(\boldsymbol{y}^a)^T\boldsymbol{M} = m\cdot\boldsymbol{e}_a$.*
- (III) Symmetry*: for any row indices $b_1, b_2 \in B$, there is a symmetry $\pi : B \to B$ of V so that $\pi(b_1) = b_2$.*
- (IV) Constant column vectors: *The all-ones-vector $\boldsymbol{1} \in \mathbb{R}^B$ lies in V.*

If N is a multiple of c and $K^{10} \leq N \leq \frac{|B|}{2}$, then one can sample each row with probability $p := \frac{N}{|B|}$ and

$$\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}], |T| = N] > 0.$$

Note that if $\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]$, then for any linear combination of columns we also have $\sum_{a\in A} y_a X_a = \mathbb{E}[\sum_{a\in A} y_a X_a]$. By the property (IV) in the KLP-Theorem, there is a linear combination $\boldsymbol{y} \in \mathbb{R}^A$ with $\sum_{a\in A} \boldsymbol{M}^a y_a = \boldsymbol{1} \in \mathbb{R}^B$, and hence we know that $\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}] \Rightarrow |T| = N$. Then it suffices to prove that $\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]] > 0$.

[1] The KLP paper also requires the "boundedness condition", which asks that V is spanned by short integer vectors, more precisely that $V = \text{span}\{\boldsymbol{x} \in V \cap \mathbb{Z}^B : \|\boldsymbol{x}\|_\infty \leq c_2\}$. But if the entries of $\boldsymbol{M}$ are in $\{0,1\}$ and $V = \text{span}\{\boldsymbol{M}^a \mid a \in A\}$, then this is automatically satisfied for $c_2 = 1$.

Actually dropping columns that are linearly dependent does not change the probability of the event $\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]$ or affect (III) and it can only loosen the conditions (II), hence we may assume w.l.o.g. for the proof that all column vectors $\boldsymbol{M}^a$ are linearly independent and $\dim(V) = |A|$. This also implies that the lattice $\mathcal{L}(\boldsymbol{M})$ has full rank, that means $\text{span}(\mathcal{L}(\boldsymbol{M})) = \mathbb{R}^A$.

Note that in the proof, the value K will be our *running parameter* and most bounds that we see are going to be some polynomial in K. However, we make no effort in optimizing the exponents of those polynomials.

8.2.2 An overview over the proof

We will now give an outline of the proof of the KLP Theorem and fill in details for definitions and proofs later.

(1) **Fourier analysis.** The proof is based on Fourier analysis argument. In particular we study the Fourier transform $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) := \mathbb{E}[\exp(2\pi i \langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)]$ of the random variable $\boldsymbol{X}$, where $\boldsymbol{\theta} \in \mathbb{R}^A$. Then the *Fourier inversion formula* tells us that

$$\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]] = \det(\mathcal{L}) \cdot \int_{D^*} \hat{\boldsymbol{X}}(\boldsymbol{\theta}) \cdot e^{-2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta}\rangle} d\boldsymbol{\theta}$$

Here D^* will denote the *Voronoi cell* of the lattice $\mathcal{L}^*$ which is the dual lattice to $\mathcal{L}$. Then the overall strategy is to prove that the latter integral evaluates to > 0.

(2) **Well-behaved Fourier coefficients.** We then prove that the Fourier coefficients $\hat{\boldsymbol{X}}(\boldsymbol{\theta})$ are well behaved. In particular it is not possible that $\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle$ is large for one row b and close to 0 for all others. More concretely one can use the symmetry condition to argue that $\max_{b\in B} |\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle| \leq K^3 \cdot \mathbb{E}_{b\in B}[\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle^2]^{1/2}$ and $\max_{b\in B}\{\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle\} \leq K^3\, \mathbb{E}_{b\in B}[\{\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle\}^2]^{1/2}$ for all $\boldsymbol{\theta} \in \mathbb{R}^A$, where $\{\cdot\}$ gives the distance to the nearest integer.

We will outline the first bound. Fix a $\boldsymbol{\theta}$ and suppose for some row $e \in B$ we want to bound $|\langle \boldsymbol{M}_e, \boldsymbol{\theta}\rangle|$ in comparison to the average $\beta := \mathbb{E}_{b\sim B}[\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle^2]^{1/2}$.

Let $E \subseteq B$ be the rows b where $|\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle| > K\beta$. By Markov's inequality $|E| \leq |B|/K^2$. Now suppose that $e \in E$ since otherwise we are done. Pick a random subset $S \subseteq B \setminus E$ of rows of size $|S| = K^2$. Using the *pigeonhole principle* one can show that there is a vector $\boldsymbol{y} \in \{-1, 0, 1\}^A$ with $\boldsymbol{y}^T \boldsymbol{M} = \boldsymbol{0}$ while $\text{supp}(\boldsymbol{y}) \subseteq S$ and $|\text{supp}(\boldsymbol{y})| \geq |S|/4$. That means for at least a quarter of rows b_0 in S there is a linear dependence of the form $\boldsymbol{M}_{b_0} = \pm \sum_{b\in S\setminus\{b_0\}} y_b \boldsymbol{M}_b$. Let $\pi_{b_0} : B \to B$ be the symmetry with $\pi_{b_0}(b_0) = e$. As symmetries preserve linear dependence we have

$$\boldsymbol{M}_e = \boldsymbol{M}_{\pi(b_0)} = \pm \sum_{b \in S/\{b_0\}} y_b \boldsymbol{M}_{\pi(b)}$$

Next, observe that even conditioning on the row b_0, the indices $\pi_{b_0}(S \setminus \{b_0\})$ are a uniform random subset of $B \setminus \pi_{b_0}(b_0)$ and in particular the probability that the indices of $\pi_{b_0}(S \setminus \{b_0\})$ intersect E is bounded by $\frac{(|S|-1)\cdot|E|}{|B|-1} \leq \frac{1}{8}$. Then with probability at least $\frac{1}{4} - \frac{1}{8} = \frac{1}{4}$ the experiment is successful and hence we can write $\boldsymbol{M}_e$ as a short linear combination of rows not from E, i.e.

$$|\langle \boldsymbol{M}_e, \boldsymbol{\theta} \rangle| \leq \Big| \sum_{b \in S \setminus \{b_0\}} y_b \langle \boldsymbol{M}_b, \boldsymbol{b} \rangle \Big| \leq \|\boldsymbol{y}\|_1 \cdot K \cdot \beta \leq K^3 \beta.$$

(3) **Fourier coefficients close to 0.** A convenient norm to use will be $\|\boldsymbol{\theta}\|_R := \mathbb{E}_{b \in B}[\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle^2]^{1/2}$. Using (2) one can show that for $\|\boldsymbol{\theta}\|_R \leq \frac{1}{8K^3}$ one has

$$\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \exp\big(2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta} \rangle - 2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta} \pm O(K^3 N \|\boldsymbol{\theta}\|_R^3)\big)$$

where $\boldsymbol{\Sigma}[\boldsymbol{X}] := p \cdot (1-p) \cdot \boldsymbol{M}^T \boldsymbol{M}$ is the *covariance matrix* of $\boldsymbol{M}$. What this means is that for $\|\boldsymbol{\theta}\|_R$ small enough, the Fourier coefficient $\hat{\boldsymbol{X}}(\boldsymbol{\theta})$ is very close to the Fourier coefficient of a Gaussian with the same expectation and covariance matrix and in particular it is positive. This fact itself is true for *any* sum of independent random variables (that's how the Central Limit Theorem is proven), but in our setting we can use (2) to argue that the threshold on the length of $\|\boldsymbol{\theta}\|_R$ can be chosen more generously.

(4) **Geometry of vectorspace V.** We can use the *Local Decodability* property to show that $\mathbb{Z}^n \setminus V$ have a $\|\cdot\|_\infty$-distance of at least $\frac{1}{K^3}$ to the subspace V.

(5) **Fourier coefficients far from the dual lattice.** Consider a Fourier coefficient $\boldsymbol{\theta} \in D^*$ that is so far from the dual lattice that we cannot apply (3). We the main inequality will be that (and there is an overlap in the cases)

$$|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp\Big(-\Theta(\tfrac{1}{K^{12}} N)\Big) \quad \forall \boldsymbol{\theta} \in D^* \text{ with } \|\boldsymbol{\theta}\|_R \geq \frac{1}{4K^6}$$

The outline of the argument is as follows. Fix a $\boldsymbol{\theta} \in D^*$ with $\|\boldsymbol{\theta}\|_R \geq \varepsilon := \frac{1}{4K^6}$ and write $\boldsymbol{M\theta} = \boldsymbol{n} + \boldsymbol{r}$ where $n_b = \lfloor \langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle \rceil \in \mathbb{Z}$ and $r_b = \pm\{\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle\}$. A straightforward estimate shows that $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp(-\Theta(N \cdot \mathbb{E}_{b \in B}[r_b^2]))$, which means that suffices to $\mathbb{E}_{b \in B}[r_b^2]$ is large. In other words we need to prove an implication of the form: $\boldsymbol{\theta}$ far from $\mathcal{L}^* \Rightarrow \boldsymbol{M\theta}$ far from $\mathbb{Z}^B$. Suppose for the sake of contradiction that $\mathbb{E}_{b\in B}[r_b^2]^{1/2} \leq \frac{1}{4\cdot K^6}$. From (2) we know that $\|\boldsymbol{r}\|_\infty \leq K^3 \cdot \mathbb{E}_{b\in B}[r_b^2]^{1/2} \leq \frac{1}{4K^3}$. As $\boldsymbol{M\theta} = \boldsymbol{n} + \boldsymbol{r} \in V$ and $\|\boldsymbol{r}\|_\infty < \frac{1}{K^3}$ we know that indeed $\boldsymbol{n}$ lies in the vectorspace V. Then we can write $\boldsymbol{n} = \boldsymbol{M\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^A$. Finally

$$\mathop{\mathbb{E}}_{b\in B}[r_b^2]^{1/2} \overset{\boldsymbol{M\theta}=\boldsymbol{M\alpha}+\boldsymbol{r}}{=} \mathop{\mathbb{E}}_{b\in B}[\langle \boldsymbol{M}_b, \boldsymbol{\theta} - \boldsymbol{\alpha}\rangle^2]^{1/2} = \|\boldsymbol{\theta} - \boldsymbol{\alpha}\|_R \overset{\boldsymbol{\theta}\in \text{ Voronoi cell}}{\geq} \|\boldsymbol{\theta}\|_R \geq \varepsilon = \frac{1}{4K^6}$$

and we have a contradiction.

(6) **Estimating** Pr$[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]]$**.** The final step involves a lot of estimates but now new ideas per se. If $\boldsymbol{Y} \in \mathbb{R}^A$ is the Gaussian random variable with identical expectation $\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$ and identical covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{Y}] = \boldsymbol{\Sigma}[\boldsymbol{X}]$. If $f_{\boldsymbol{Y}}$ is the density function of that Gaussian, then we can estimate that for a proper choice of $\varepsilon := \text{poly}(K) \cdot \ln(N)/\sqrt{N}$ one has

$$\frac{\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]]}{\det(\mathcal{L})} \geq \underbrace{f_{\boldsymbol{Y}}(\mathbb{E}[\boldsymbol{X}])}_{(*)} - \underbrace{\int_{\|\boldsymbol{\theta}\|_R \leq \varepsilon} |\hat{\boldsymbol{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{\text{small by (3)}} - \underbrace{\int_{\boldsymbol{\theta}\in D^*:\|\boldsymbol{\theta}\|_R > \varepsilon} |\hat{\boldsymbol{X}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{\text{small by (5)}}$$
$$- \underbrace{\int_{\|\boldsymbol{\theta}\|_R > \varepsilon} |\hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{\text{small by } (*)} \quad > 0$$

In $(*)$ we use standard estimates for Gaussians. This proves the KLP-Theorem.

8.3 Proof of the KLP-Theorem

The main idea behind the proof is to design a variant of the *central limit theorem* that provides super-fast convergence based on the properties of $\boldsymbol{M}$, in particular the symmmetry and the fact that $|B| \gg |A|$. The *Fourier transform* of the random variable $\boldsymbol{X}$ is the function

$$\hat{\boldsymbol{X}} : \mathbb{R}^A \to \mathbb{C} \quad \text{with} \quad \hat{\boldsymbol{X}}(\boldsymbol{\theta}) := \mathbb{E}\left[\exp(2\pi i \langle \boldsymbol{X}, \boldsymbol{\theta}\rangle)\right]$$

We abbreviate $\mathcal{L} := \mathcal{L}(\boldsymbol{M})$ as the lattice spanned by row vectors of $\boldsymbol{M}$ and denote

$$\mathcal{L}^* := \left\{\boldsymbol{y} \in \mathbb{R}^A \mid \langle \boldsymbol{y}, \boldsymbol{z}\rangle \in \mathbb{Z} \quad \forall \boldsymbol{z} \in \mathcal{L}\right\}.$$

as the *dual lattice*[2]. We assumed that $\boldsymbol{M}$ has rank $|A|$, hence both $\mathcal{L}$ and $\mathcal{L}^*$ are full rank lattices and $\det(\mathcal{L}) \cdot \det(\mathcal{L}^*) = 1$. For any dual lattice vector $\boldsymbol{y} \in \mathcal{L}^*$ we

[2] In the KLP paper the dual lattice is denoted by L.

have $\langle \boldsymbol{M}_b, \boldsymbol{y} \rangle \in \mathbb{Z}$ and hence $\langle \boldsymbol{X}, \boldsymbol{y} \rangle \in \mathbb{Z}$ for any outcome of $\boldsymbol{X}$. That implies that $e^{2\pi i \langle \boldsymbol{X}, \boldsymbol{y} \rangle} = 1$ and the Fourier transform $\hat{\boldsymbol{X}}$ is $\mathcal{L}^*$*-periodic*, meaning that

$$\hat{\boldsymbol{X}}(\boldsymbol{\theta} + \boldsymbol{y}) = \hat{\boldsymbol{X}}(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \mathbb{R}^A \ \forall \boldsymbol{y} \in \mathcal{L}^*$$

Actually we will have $\boldsymbol{X} \in \mathcal{L}$ at any time. To understand the convergence of $\boldsymbol{X}$, it will be crucial to consider its *covariance matrix* $\Sigma[\boldsymbol{X}] \in \mathbb{R}^{A \times A}$ which is given by

$$\begin{aligned} \boldsymbol{\Sigma}[\boldsymbol{X}] \quad &:= \quad \mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\right)^T \left(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\right)\right] \\ &\overset{\text{independence}}{=} \quad \sum_{b \in B} \left(\mathbb{E}[(T_b \boldsymbol{M}_b)^T (T_b \boldsymbol{M}_b)] - \mathbb{E}[T_b \boldsymbol{M}_b]^T \, \mathbb{E}[T_b \boldsymbol{M}_b] \right) \\ &= \quad \sum_{b \in B} (p - p^2) \boldsymbol{M}_b^T \boldsymbol{M}_b = p(1-p) \underbrace{\boldsymbol{M}^T \boldsymbol{M}}_{=: \boldsymbol{R}}. \end{aligned}$$

We can use the scalar $\boldsymbol{R}$ of the covariance matrix to define a useful *norm*

$$\|\boldsymbol{\theta}\|_{\boldsymbol{R}} := \left(\frac{1}{|B|} \cdot \boldsymbol{\theta}^T \boldsymbol{R} \boldsymbol{\theta} \right)^{1/2} = \left(\mathop{\mathbb{E}}_{b \sim B} [\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle^2] \right)^{1/2}$$

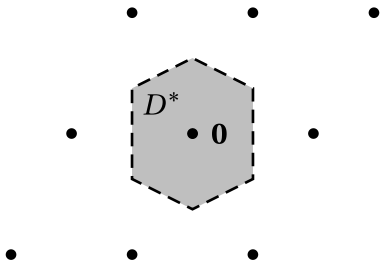
We define the corresponding *norm ball*

$$\mathcal{B}_{\boldsymbol{R}}(\varepsilon) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^A \mid \|\boldsymbol{\theta}\|_{\boldsymbol{R}} \leq \varepsilon \right\}$$

Moreover, let

$$D^* := \left\{ \boldsymbol{\theta} \in \mathbb{R}^A \mid \|\boldsymbol{\theta}\|_R < \|\boldsymbol{\theta} - \boldsymbol{y}\|_R \quad \boldsymbol{y} \in \mathcal{L}^* \setminus \{\boldsymbol{0}\} \right\}$$

be the *Voronoi cell* of the dual lattice with respect to the norm $\|\cdot\|_{\boldsymbol{R}}$. Recall that the Voronoi cell contains all points that are closer to $\boldsymbol{0}$ than to any other $\mathcal{L}^*$-lattice point in terms of $\|\cdot\|_{\boldsymbol{R}}$-distance.



visualization of $\mathcal{L}^*$ and D^*

An important property is that D^* induces a *tiling* of the whole space, meaning that $D^* + \boldsymbol{y}$ for $\boldsymbol{y} \in \mathcal{L}^*$ partitions the space (apart from the set of measure 0 that belongs to the boundary between translates).

Next, we recall the reason why the Fourier transform is useful when talking about probabilities — the reason is that with the *Inverse Fourier Transform*, the probabilities can be recovered from the Fourier coefficients:

Lemma 8.4 (Fourier inversion formula)**.** *For any $\boldsymbol{\lambda}\in\mathcal{L}$ one has*

$$\Pr[\boldsymbol{X}=\boldsymbol{\lambda}] = \det(\mathcal{L})\cdot\int_{D^*}\hat{\boldsymbol{X}}(\boldsymbol{\theta})\cdot e^{-2\pi i\langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle}\,d\boldsymbol{\theta}$$

Proof. We apply the usual "swap-and-cancel trick" in Fourier analysis to obtain

$$\begin{aligned}\int_{D^*}\hat{\boldsymbol{X}}(\boldsymbol{\theta})\cdot\exp(-2\pi i\langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle)\,d\boldsymbol{\theta} &= \int_{D^*}\mathop{\mathbb{E}}_{\boldsymbol{X}}\big[\exp(2\pi i\langle\boldsymbol{X},\boldsymbol{\theta}\rangle)\big]\cdot\exp(-2\pi i\langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle)\,d\boldsymbol{\theta}\\ &= \mathop{\mathbb{E}}_{\boldsymbol{X}}\Big[\underbrace{\int_{D^*}\exp\big(2\pi i\langle\boldsymbol{X}-\boldsymbol{\lambda},\boldsymbol{\theta}\rangle\big)\,d\boldsymbol{\theta}}_{=0\text{ if }\boldsymbol{X}-\boldsymbol{\lambda}\in\mathcal{L}\backslash\{\mathbf{0}\}}\Big]\\ &= \Pr[\boldsymbol{X}=\boldsymbol{\lambda}]\cdot\underbrace{\int_{D^*}\underbrace{\exp(2\pi i\cdot 0)}_{=1}\,d\boldsymbol{\theta}}_{=\det(\mathcal{L}^*)} = \frac{1}{\det(\mathcal{L})}\cdot\Pr[\boldsymbol{X}=\boldsymbol{\lambda}].\end{aligned}$$

using that $\mathrm{vol}_n(D^*)=\det(\mathcal{L}^*)=\frac{1}{\det(\mathcal{L})}$ for any full rank dual lattice. Here we use the following:

Claim. Let $\boldsymbol{s}\in\mathcal{L}/\{\mathbf{0}\}$. Then $\int_{D^*}e^{2\pi i\langle\boldsymbol{s},\boldsymbol{\theta}\rangle}\,d\boldsymbol{\theta}=0$.

Proof of claim. As mentioned earlier, $\mathcal{L}^*+D^*$ gives a tiling of $\mathbb{R}^n$. If $P\subseteq\mathbb{R}^n$ is another region so that $\mathcal{L}^*+P$ is a tiling, then the integral of any $\mathcal{L}^*$-periodic function over D^* and P will give the same value. Let $\boldsymbol{u}_1,\dots,\boldsymbol{u}_{|A|}\in\mathcal{L}^*$ be a basis of the lattice $\mathcal{L}^*$ and suppose we picked indices so that $\langle\boldsymbol{s},\boldsymbol{u}_1\rangle\neq 0$. Our choice for a region is the *fundamental parallelepiped* $P:=\{\sum_{i=1}^{|A|}z_i\boldsymbol{u}_i\mid 0\le z_i<1\}$. We will write $\boldsymbol{z}=(z_1,\bar{\boldsymbol{z}})$. Then

$$\begin{aligned}\int_{D^*}e^{2\pi i\langle\boldsymbol{s},\boldsymbol{\theta}\rangle}\,d\boldsymbol{\theta} &= \int_P e^{2\pi i\langle\boldsymbol{s},\boldsymbol{\theta}\rangle}\,d\boldsymbol{\theta} = \det(\mathcal{L}^*)\int_{[0,1[^n}e^{2\pi i\langle\sum_{i=1}^{|A|}z_i\boldsymbol{u}_i,\boldsymbol{s}\rangle}\,d\boldsymbol{z}\\ &= \det(\mathcal{L}^*)\underbrace{\int_0^1 e^{2\pi i z_1\langle\boldsymbol{u}_1,\boldsymbol{s}\rangle}\,dz_1}_{=0}\cdot\Big(\int_{\bar{\boldsymbol{z}}\in[0,1[^{n-1}}e^{2\pi i\langle\sum_{i=2}^{|A|}z_i\boldsymbol{u}_i,\boldsymbol{s}\rangle}\,d\bar{\boldsymbol{z}}\Big)\end{aligned}$$

Here the crucial observation is that $\int_0^1 e^{2\pi i z_1\langle\boldsymbol{u}_1,\boldsymbol{s}\rangle}\,dz=0$ since we integrate exactly $|\langle\boldsymbol{u}_1,\boldsymbol{s}\rangle|\in\mathbb{N}$ times over the complex unit circle and antipodal values cancel each other out. □

We sampled rows independently, which we can use to get an explicit expression for the Fourier coefficients:

Lemma 8.5. *For any $\boldsymbol{\theta}\in\mathbb{R}^A$ one has*

$$\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \prod_{b\in B}\Big(1-p+p\cdot\exp\big(2\pi i\langle\boldsymbol{M}_b,\boldsymbol{\theta}\rangle\big)\Big)$$

Proof. We use the independence of the random variables T_b to write

$$\begin{aligned}\hat{\boldsymbol{X}}(\boldsymbol{\theta}) &= \mathbb{E}\Big[\exp\Big(2\pi i\langle \sum_{b\in B} T_b \boldsymbol{M}_b, \boldsymbol{\theta}\rangle\Big)\Big] = \prod_{b\in B}\mathbb{E}\Big[\exp\big(2\pi \boldsymbol{i}\cdot T_b\cdot\langle \boldsymbol{M}_b,\boldsymbol{\theta}\rangle\big)\Big] \\ &\overset{T_b\in\{0,1\}}{=} \prod_{b\in B}\Big((1-p)\cdot 1 + p\cdot\exp(2\pi \boldsymbol{i}\,\langle \boldsymbol{M}_b,\boldsymbol{\theta}\rangle)\Big),\end{aligned}$$

which gives the claim. □

From the formula in Lemma 8.5 we can already obtain an important observation for later: If for a fixed coefficient vector $\boldsymbol{\theta}$ we have a non trivial fraction of $b \in B$ satisfy that $\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle$ has a significant distance from the nearest integer, then the size $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|$ will be negligibly small.

8.3.1 A basis for $V^\perp$

Reall that $V^\perp \subseteq \mathbb{R}^A$ is the space that is orthogonal to the space V that is spanned by the column vectors $\boldsymbol{M}^a$. For later, we will need a "short" basis of $V^\perp$:

Lemma 8.6. *The space $V^\perp$ has a basis of integer vectors of $\|\cdot\|_1$-length at most K^3.*

Proof. By a slight abuse of notation, we denote $\boldsymbol{e}_a \in \mathbb{R}^A$ as a unit vector in $\mathbb{R}^A$ and $\boldsymbol{e}_b \in \mathbb{R}^B$ as a unit vector in $\mathbb{R}^B$. By property (II), we know that there are vectors $\boldsymbol{y}^a \in \mathbb{Z}^B$ with $m\cdot \boldsymbol{e}_a = (\boldsymbol{y}^a)^T\boldsymbol{M}$ so that $1 \le m \le K$ and $\|\boldsymbol{y}^a\|_1 \le K$. Now consider the vectors

$$\boldsymbol{u}^b := \Big(m\cdot \boldsymbol{e}_b - \sum_{a'\in A} M_{ba'}\boldsymbol{y}^{a'}\Big) \in \mathbb{R}^B \quad \forall b \in B.$$

We claim that $\{\boldsymbol{u}^b\}_{b\in B} \subseteq V^\perp$ as for any column vector $\boldsymbol{M}^a$ one has

$$\langle \boldsymbol{M}^a, \boldsymbol{u}^b\rangle = m\cdot \underbrace{\langle \boldsymbol{e}_b, \boldsymbol{M}^a\rangle}_{=M_{ba}} - \sum_{a'\in A} M_{ba'}\cdot \underbrace{\langle \boldsymbol{y}^{a'}, \boldsymbol{M}^a\rangle}_{=m \text{ if } a=a',0 \text{ o.w.}} = 0.$$

The length of the vectors is

$$\|\boldsymbol{u}^b\|_1 \le \underbrace{m}_{\le K}\cdot\underbrace{\|\boldsymbol{e}_b\|_1}_{\le 1} + \sum_{a'\in A}\underbrace{|M_{ba'}|}_{\le 1}\cdot\underbrace{\|\boldsymbol{y}^{a'}\|_1}_{\le K} \le K + K^2 \le K^3$$

As $\dim(\text{span}\{\boldsymbol{e}_b \mid b \in B\}) = |B|$ and $\dim(\text{span}\{\boldsymbol{y}^a \mid a \in A\}) \le |A|$, we clearly have $\dim(\text{span}\{\boldsymbol{u}^b \mid b \in B\}) \ge |B| - |A|$. Then any maximally linear independent subset of $\{\boldsymbol{u}^b\}_{b\in B}$ is a suitable basis of $V^\perp$. □

8.3.2 Rows as linear combinations of few other rows

We start with a small auxiliary lemma. In the following, for an index set $S \subseteq B$, we denote $\boldsymbol{M}_S$ as the submatrix of $\boldsymbol{M}$ that contains only the rows in S.

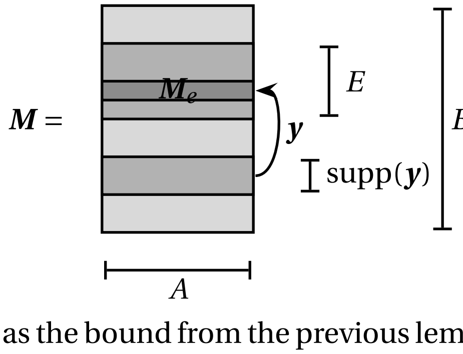
Lemma 8.7. *For any subset $S \subseteq B$ of $|S| \geq \Theta(K\log(K))$ rows, there is a non-zero vector $\boldsymbol{y} \in \{-1,0,1\}^S$ satisfying $\boldsymbol{y}^T\boldsymbol{M}_S = \boldsymbol{0}$ and $|supp(\boldsymbol{y})| \geq \frac{|S|}{4}$.*

Proof. The proof is a *pigeonhole principle* argument. First, note that for $\boldsymbol{y} \in \{0,1\}^S$ one has $\|\boldsymbol{y}^T\boldsymbol{M}_S\|_\infty \leq |S|$ and hence there are only $(|S|+1)^{|A|}$ many outcomes for $2^{|S|}$ many assignments $\boldsymbol{y} \mapsto \boldsymbol{y}^T\boldsymbol{M}_S$. It is not hard to check that if $|S| \geq O(|A|\log|A|)$ with a large enough hidden constant, then $2^{|S|} \geq 2^{0.99|S|} \cdot (|S|+1)^{|A|}$ and there will even be $2^{0.99|S|}$ many $\boldsymbol{y}$'s with $\boldsymbol{y}^T\boldsymbol{M}_S$ being identical. Then another simple calculation shows that there must be $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \{0,1\}^S$ with $\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_1 \geq \frac{|S|}{4}$ and $\boldsymbol{y}^T\boldsymbol{M}_S = \tilde{\boldsymbol{y}}^T\boldsymbol{M}_S$. Then $\boldsymbol{y} - \tilde{\boldsymbol{y}}$ satisfies the claim. □

As we will see, any row $\boldsymbol{M}_e$ can be written as an integer combination of a few other rows, even if some not too large subset of rows is forbidden. The proof uses the vector $\boldsymbol{y}$ with $\boldsymbol{y}^T\boldsymbol{M} = \boldsymbol{0}$ from the last Lemma 8.7. Then we can use the symmetry assumption to embed the set S randomly. Then with positive probability one will have $y_e \neq 1$ and $\text{supp}(\boldsymbol{y}) \cap (E \setminus \{e\}) = \emptyset$.

Lemma 8.8. *For any subset $E \subseteq B$ of at least $|E| \leq \frac{|B|}{K^2}$ many rows and any $e \in E$, there is a $\boldsymbol{y} \in \mathbb{Z}^B$ with $\boldsymbol{M}_e = \boldsymbol{y}^T\boldsymbol{M}$, $supp(\boldsymbol{y}) \subseteq B \setminus E$ and $\|\boldsymbol{y}\|_1 \leq K^2$.*

A poor visualization might be as follows:



Proof. Set $s := \Theta(K\log(K))$ as the bound from the previous lemma. Fix a set $E \subseteq B$ with $|E| \leq \frac{|B|}{8s}$ and $e \in E$ and note that $8s \leq K^2$ as we may assume that K is at least a big enough constant. We know that for every row $b \in B$ there is a *symmetry* $\pi_b : B \to B$ with $\pi_b(b) = e$. We pick a subset $S \subseteq B$ of fixed size $|S| = s$ uniformly at random and denote $\boldsymbol{y}^S \in \{-1,0,1\}^B$ as the non-zero vector with $\boldsymbol{y}^S\boldsymbol{M} = 0$ and $\text{supp}(\boldsymbol{y}^S) \subseteq S$ that exists by the previous Lemma 8.7. Now pick $b_0 \in S$ uniformly at random. Then

$$\Pr_{S,b_0}[y^S_{b_0} \neq 0] \geq \frac{1}{4}$$

as b_0 is picked uniformly from S. But even conditioning on a fixed b_0 we have

$$\Pr_S[\underbrace{\pi_{b_0}(S \setminus \{b_0\})}_{\text{unif. from } B/\{b_0\}} \cap E \neq \emptyset \mid b_0] \leq \frac{s|E|}{|B|} \leq \frac{1}{8}.$$

Then by the union bound $\Pr_{S,b_0}[y^S_{b_0} \neq 0 \text{ and } \pi_{b_0}(S) \cap E = \{e\}] \geq \frac{1}{4} - \frac{1}{8} > 0$. Let us fix an outcome of (S, b_0) and $\boldsymbol{y} := \boldsymbol{y}^S$ attaining this event — wl.o.g. with $y_{b_0} = 1$. Then

$$\boldsymbol{M}_{b_0} = \sum_{b \in S \setminus \{b_0\}} (-y_b)\boldsymbol{M}_b$$

and applying the symmetry permutation to the rows gives

$$\boldsymbol{M}_e = \boldsymbol{M}_{\pi(b_0)} = \sum_{b \in S \setminus \{b_0\}} (-y_b) \cdot \boldsymbol{M}_{\pi(b)}$$

which means we can write $\boldsymbol{M}_e$ as a short integer combination of rows with indices $\pi(b) \notin E$. □

8.3.3 Well behaved Fourier coefficients

We need a technical lemma that tells us that the inner product $\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle$ for a particular row b is always bounded in terms of the inner product for the average row. Basically this is where symmetry as well as the other properties come into play. In the following, for a number $\alpha \in \mathbb{R}$, let $\{\alpha\} := \min_{z \in \mathbb{Z}} |\alpha - z|$ be the distance to the nearest integer.

Lemma 8.9. *The following holds for every $\boldsymbol{\theta} \in \mathbb{R}^A$:*

(i) One has $\max_{b \in B} |\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle| \leq K^3 \cdot (\mathbb{E}[\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle^2])^{1/2} = K^3 \cdot \|\boldsymbol{\theta}\|_R$.

(i) Write $r_b := \{\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle\}$ as the distance to the nearest integer. Then $\max_{b \in B} |r_b| \leq K^3 \cdot \left(\mathbb{E}_{b \in B}[r_b^2]\right)^{1/2}$.

Proof. For a fixed vector $\boldsymbol{\theta} \in \mathbb{R}^A$, set $\beta := (\mathbb{E}_{b \sim B}[\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle^2])^{1/2}$ as a sort of a "geometric average". Consider the set of indices $E := \{b \in B \mid |\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle| \geq K\beta\}$ where the inner product $|\langle \boldsymbol{M}_b, \boldsymbol{\theta} \rangle|$ is significantly above that average. Then $\beta \geq (\frac{1}{|B|} \cdot |E| \cdot (K\beta)^2)^{1/2} \Rightarrow |E| \leq \frac{|B|}{K^2}$. Fix any $e \in E$ (if there is none, we are done with (i)).

Then by Lemma 8.8, there is a vector $\boldsymbol{y}$ with support disjoint to E so that $\boldsymbol{M}_e = \sum_{b\in B\setminus E} y_b \boldsymbol{M}_b$ and $\|\boldsymbol{y}\|_1 \leq K^2$. Then we can bound the inner product as

$$|\langle \boldsymbol{M}_e, \boldsymbol{\theta}\rangle| \leq \underbrace{\sum_{b\in B\setminus E} |y_b|}_{\leq K^2} \cdot \underbrace{|\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle|}_{\leq K\beta} \leq K^3\beta$$

and the claim (i) is true.

The proof for (ii) is quite similar. Set $\beta := \mathbb{E}_{b\sim B}[r_b^2]^{1/2}$ as the "geometric average" of the remainders of inner products. Again, let $E := \{b \in B \mid |r_b| \geq K\beta\}$ be the rows where the remainder significantly exceeds the remainder. As before, we conclude that $|E| \leq \frac{|B|}{K^2}$ and fix a row $e \in E$. Again, take a $\boldsymbol{y} \in \mathbb{Z}^B$ with $\boldsymbol{M}_e = \sum_{b\in E\setminus B} y_b \boldsymbol{M}_b$ and $\text{supp}(\boldsymbol{y}) \cap E = \emptyset$. Then

$$\langle \boldsymbol{M}_e, \boldsymbol{\theta}\rangle = \sum_{b\in B\setminus E} y_b \langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle$$

holds and taking the remainders on both sides, we obtain

$$|r_e| \leq \underbrace{\sum_{b\in B\setminus E} |y_b|}_{\leq K^2} \cdot \underbrace{|r_b|}_{\leq K\beta} \leq K^3\beta$$

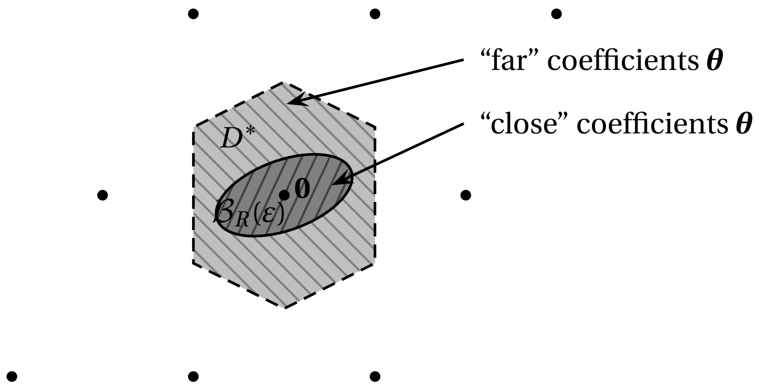
as claimed. □

8.3.4 The Fourier Transform near the origin

The next step is to show that the Fourier coefficients $\hat{\boldsymbol{X}}(\boldsymbol{\theta})$ close to the origin can be very well approximated by a Gaussian with same expectation and covariance matrix as $\boldsymbol{X}$. Here, "close" will mean that $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{R}}(\varepsilon)$ with a choice of

$$\varepsilon := \frac{\text{poly}(K)\cdot \ln(N)}{\sqrt{N}},$$

but we keep the following lemmas general.


"far" coefficients $\boldsymbol{\theta}$
"close" coefficients $\boldsymbol{\theta}$
D^*
$\mathbf{0}$
$\mathcal{B}_R(\varepsilon)$

visualization of $\mathcal{L}^*$ and D^*

Lemma 8.10. *For all $0 < \varepsilon \leq \frac{1}{8K^3}$ and all $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{R}}(\varepsilon)$ one has*

$$\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \exp\Big(2\pi i \cdot \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta}\rangle - 2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta} + \delta\Big),$$

where $|\delta| \leq O(K^3 \cdot \|\boldsymbol{\theta}\|_{\boldsymbol{R}}^3 N)$.

Proof. First note that the radius of the ball is chosen small enough so that all the inner products can be bounded by

$$\max_{b\in B} |\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle| \overset{\text{Lemma 8.9}}{\leq} K^3 \cdot \underbrace{\|\boldsymbol{\theta}\|_{\boldsymbol{R}}}_{\leq \frac{1}{8K^3}} \leq \frac{1}{8} \qquad (*)$$

Let us abbreviate $x_b := 2\pi \langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle$ as the “number of rotations” for the Fourier coefficient, that means $|x_b| \leq \frac{2\pi}{8} = \frac{\pi}{4}$. We can now get a more handy expression for $\hat{\boldsymbol{X}}(\boldsymbol{\theta})$. Here we use the standard estimates of $\exp(z) = 1 + z + \frac{1}{2}z^2 + O(|z|^3)$ and $1 + z = \exp(z - \frac{1}{2}z^2 + O(|z|^3))$ for $z \in \mathbb{C}$ for $|z| \leq 1$.

$$\begin{aligned}
\hat{\boldsymbol{X}}(\boldsymbol{\theta}) &\overset{\text{Lem 8.5}}{=} \prod_{b\in B} \big(1 + p\cdot(e^{i\cdot x_b} - 1)\big) \\
&= \prod_{b\in B} \exp\Big(p\cdot(e^{i\cdot x_b} - 1) - \frac{1}{2}p^2 \underbrace{(e^{i\cdot x_b} - 1)^2}_{=-x_b^2 + O(|x_b^4|)} + p^3 \cdot \underbrace{O(|e^{i\cdot x_b} - 1|^3)}_{O(|x_b|^3)}\Big) \\
&= \prod_{b\in B} \exp\Big(p\cdot\Big(i\cdot x_b + \frac{1}{2}(i\cdot x_b)^2 + O(|x_b|^3)\Big) + \frac{1}{2}p^2 x_b^2 + O(p^2|x_b|^3)\Big) \\
&= \exp\Big(ip\sum_{b\in B} x_b - \frac{1}{2}p(1-p)\sum_{b\in B} x_b^2 + O\Big(p\sum_{b\in B}|x_b|^3\Big)\Big) \\
&= \exp\Big(2\pi i \cdot \underbrace{p\sum_{b\in B}\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle}_{=\langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta}\rangle} - 2\pi^2 \underbrace{p(1-p)\sum_{b\in B}\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle^2}_{=\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}} + O\Big(p\sum_{b\in B}|x_b|^3\Big)\Big)
\end{aligned}$$

This already corresponds to the claimed expression for $\hat{\boldsymbol{X}}(\boldsymbol{\theta})$, just that we need to justify the bound on the error term. In fact,

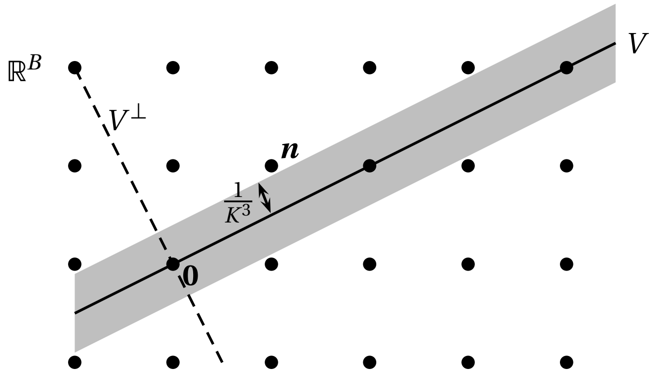
$$p\sum_{b\in B}|x_b|^3 \leq 8\pi^3 p \cdot \underbrace{\max\{|\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle| \mid b \in B\}}_{\leq K^3\|\boldsymbol{\theta}\|_{\boldsymbol{R}} \text{ by } (*)} \cdot \underbrace{\sum_{b\in B}\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle^2}_{=|B|\cdot\|\boldsymbol{\theta}\|_{\boldsymbol{R}}^2} \leq 8\pi^3 \underbrace{p|B|}_{=N} \cdot K^3 \cdot \|\boldsymbol{\theta}\|_{\boldsymbol{R}}^3.$$

□

8.3.5 The Fourier Transform far from $\mathcal{L}^*$

The next step is to show that the Fourier coefficients $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|$ decay rapidly when moving away from the dual lattice. First, a technical lemma. Recall that $V = \text{span}\{\boldsymbol{M}^a \mid a \in A\}$ is the span of the columns of $\boldsymbol{M}$. Then for any $\boldsymbol{\theta} \in \mathbb{R}^A$, the vector $(\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle)_{b\in B} = \boldsymbol{M}\boldsymbol{\theta}$ lies by definition in V. Now suppose that $\boldsymbol{M}\boldsymbol{\theta}$ lies very close to an integer vector $\boldsymbol{n}$. Then it is not apriori clear that also $\boldsymbol{n}$ lies in V (for a general matrix, it might not). However, our matrix $\boldsymbol{M}$ is nicely behaved, so that integer vectors are separated from V:

Lemma 8.11. *Any vector $\boldsymbol{n} \in \mathbb{Z}^B \setminus V$ and $\boldsymbol{\theta} \in \mathbb{R}^A$ one has $\|\boldsymbol{n} - \boldsymbol{M}\boldsymbol{\theta}\|_\infty \geq \frac{1}{K^3}$.*

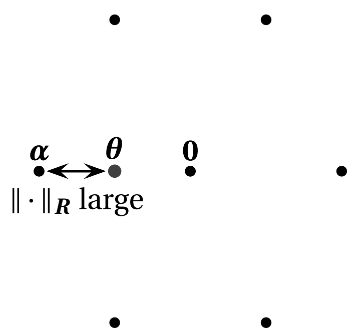


Proof. Fix $\boldsymbol{n} \in \mathbb{Z}^B \setminus V$ and let $\boldsymbol{n} + \boldsymbol{r} \in V$ be the vector that minimizes $\|\boldsymbol{r}\|_\infty$. We know from Lemma 8.6 that $V^\perp$ has a basis of short integer vectors of $\|\cdot\|_1$-length at most K^3. Let $\boldsymbol{u} \in V^\perp \cap \mathbb{Z}^B$ be such a basis vector that is not orthogonal to $\boldsymbol{n}$ (and hence not to $\boldsymbol{r}$). Then

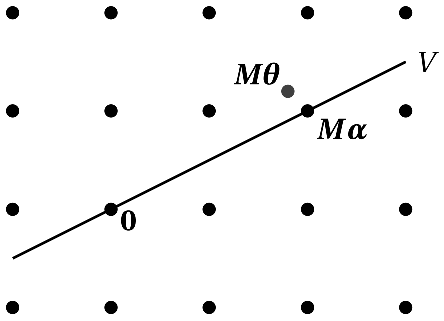
$$1 \leq |\underbrace{\langle \boldsymbol{u}, \boldsymbol{n}+\boldsymbol{r}\rangle}_{=0} - \underbrace{\langle \boldsymbol{u}, \boldsymbol{n}\rangle}_{\in\mathbb{Z}\setminus\{0\}}| = |\langle \boldsymbol{u}, \boldsymbol{r}\rangle| \leq \underbrace{\|\boldsymbol{u}\|_1}_{\leq K^3} \cdot \|\boldsymbol{r}\|_\infty$$

and rearranging gives $\|\boldsymbol{r}\|_\infty \geq \frac{1}{K^3}$. □

Now we can show that the Fourier coefficients $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|$ decay exponentially fast when moving away from dual lattice points. We want to give the proof intuition first. For a vector $\boldsymbol{\theta} \in D^*$ that is far enough from the origin, let $r_b := \{\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle\}$ be the remainder. We know that if one r_b is large, then the average of r_b's is somewhat large and it is not hard to then calculate that $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp(-\Theta(N) \cdot \mathbb{E}_{b\sim B}[r_b^2])$. So suppose for the sake of contradiction that $r_b \approx 0$ for all $b \in B$. In matrix notation, that means the vector $\boldsymbol{M}\boldsymbol{\theta}$ is very close to $\mathbb{Z}^B$. By the previous Lemma, close lattice point will be in V and hence can be written in the form $\boldsymbol{M}\boldsymbol{\alpha} \in \mathbb{Z}^B$ with $\boldsymbol{\alpha} \in \mathbb{R}^A$. In other words $\langle \boldsymbol{M}_b, \boldsymbol{\alpha}\rangle \in \mathbb{Z}$ for each b and hence $\boldsymbol{\alpha} \in \mathcal{L}^*$ lies in the dual lattice. Then their distance $\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|_{\boldsymbol{R}}^2 = \mathbb{E}_{b\sim B}[r_b^2]$ is large by assumption and we have a contradiction.



visualization of $\mathcal{L}^* \subseteq \mathbb{R}^A$



visualization of $V \subseteq \mathbb{R}^B$ and $\mathbb{Z}^B$

Lemma 8.12. *For* $0 < \varepsilon < \frac{1}{4K^6}$ *and* $\boldsymbol{\theta} \in D^* \setminus \mathcal{B}_{\boldsymbol{R}}(\varepsilon)$ *one has* $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp(-\frac{\varepsilon^2}{2}N)$.

Proof. Let us write $\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle = n_b + r_b$ with $n_b \in \mathbb{Z}$ and $|r_b| \leq \frac{1}{2}$.

Claim. $\mathbb{E}_{b\in B}[r_b^2]^{1/2} \geq \varepsilon$.

Proof of Claim. Written in matrix-vector notation, we have $\boldsymbol{M\theta} = (\boldsymbol{n}+\boldsymbol{r}) \in V$. We know by Lemma 8.9 that

$$\mathop{\mathbb{E}}_{b\in B}[r_b^2]^{1/2} \geq \frac{\|\boldsymbol{r}\|_\infty}{K^3},$$

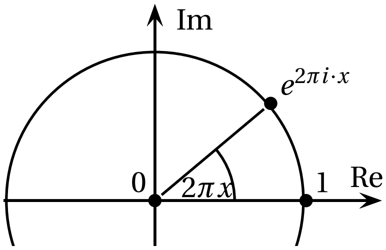
hence we are done if $\|\boldsymbol{r}\|_\infty \geq K^3\varepsilon$. So suppose for the sake of contradiction that this is not the case and $\|\boldsymbol{r}\|_\infty \leq K^3\varepsilon \leq \frac{1}{4K^3}$. However, for that case we know from Lemma 8.11 that close enough integer vectors are indeed in the subspace, that means $\boldsymbol{n} \in V$. By linear independence of $\boldsymbol{M}$'s columns, there is a unique $\boldsymbol{\alpha} \in \mathbb{R}^A$ with $\boldsymbol{M\alpha} = \boldsymbol{n}$. Note that $\boldsymbol{\alpha} \in \mathcal{L}^*$ since otherwise there had to be a generator $\boldsymbol{M}_b$ of the primal lattice $\mathcal{L}$ with $n_b = \langle \boldsymbol{M}_b, \boldsymbol{\alpha}\rangle \notin \mathbb{Z}$. From the properties of the Voronoi cell D^* of the dual lattice and the assumption $\boldsymbol{\theta} \in D^*$, we know that $\|\boldsymbol{\theta}-\boldsymbol{\alpha}\|_{\boldsymbol{R}} \geq \|\boldsymbol{\theta}\|_{\boldsymbol{R}}$. Then

$$\mathop{\mathbb{E}}_{b\in B}[r_b^2]^{1/2} \overset{\boldsymbol{M\theta}=\boldsymbol{M\alpha}+\boldsymbol{r}}{=} \frac{1}{|B|}\sum_{b\in B}\langle \boldsymbol{M}_b, \boldsymbol{\theta}-\boldsymbol{\alpha}\rangle^2 = \|\boldsymbol{\theta}-\boldsymbol{\alpha}\|_{\boldsymbol{R}} \geq \|\boldsymbol{\theta}\|_{\boldsymbol{R}} \geq \varepsilon. \quad \square$$

Now, bounding the Fourier coefficient is fairly straightforward:

$$\begin{aligned}
|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| &= \Big|\prod_{b\in B}\Big(1-p+p\cdot e^{2\pi i\langle \boldsymbol{M}_b,\boldsymbol{\theta}\rangle}\Big)\Big| \overset{\text{cyclicity of exp}}{=} \Big|\prod_{b\in B}\Big(1+p\cdot(e^{2\pi i\cdot r_b}-1)\Big)\Big| \\
&\leq \prod_{b\in B}\Big(1+p\cdot\underbrace{\mathrm{Re}(e^{2\pi i r_b}-1)}_{\leq -r_b^2}\Big) \overset{1-x\leq e^{-x/2}\forall 0\leq x\leq\frac{1}{2}}{\leq} \prod_{b\in B}\exp\Big(-\frac{p}{2}r_b^2\Big) \\
&= \exp\Big(-\frac{1}{2}\underbrace{p|B|}_{=N}\underbrace{\mathop{\mathbb{E}}_{b\in B}[r_b^2]}_{\geq\varepsilon^2}\Big)
\end{aligned}$$

Here we have used that $\mathrm{Re}(e^{2\pi i \cdot x}) = 1 - \Theta(x^2)$ for small x.



□

8.3.6 The main proof

First we need to show a quick lemma that says that we can put some sizable ball in the Voronoi cell:

Lemma 8.13. *One has $\mathcal{B}_{\boldsymbol{R}}(\frac{1}{2K^3}) \subseteq D^*$.*

Proof. We will show that any dual lattice point $\boldsymbol{\theta} \in \mathcal{L}^* \setminus \{\boldsymbol{0}\}$ has a length of $\|\boldsymbol{\theta}\|_{\boldsymbol{R}} \geq \frac{1}{K^3}$. Then every point with $\|\cdot\|_{\boldsymbol{R}}$-norm less than $\frac{1}{2K^3}$ must be closer to the origin than to any other lattice point. For every dual lattice point $\boldsymbol{\theta}$ there must be a generator $\boldsymbol{M}_b$ of the lattice that is not orthogonal and has

$$1 \leq |\langle \boldsymbol{M}_b, \boldsymbol{\theta}\rangle| \overset{\text{Lem. 8.9}}{\leq} K^3 \cdot \|\boldsymbol{\theta}\|_{\boldsymbol{R}}$$

Rearranging for $\|\boldsymbol{\theta}\|_{\boldsymbol{R}}$ gives the claim. □

Now we discuss the remaining main proof. Consider a second random vector $\boldsymbol{Y} \in \mathbb{R}^A$ that is a *random Gaussian* with expectation $\mathbb{E}[\boldsymbol{X}]$ and covariance matrix $\boldsymbol{\Sigma}[\boldsymbol{X}]$. Note that the *density function* of that Gaussian will be

$$f_{\boldsymbol{Y}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{|A|/2}\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}} \cdot \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}])^T \boldsymbol{\Sigma}[\boldsymbol{X}]^{-1}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{X}])\Big)$$

A well known fact that we leave as exercise is that the Fourier transform of that Gaussian is

$$\hat{\boldsymbol{Y}}(\boldsymbol{\theta}) = \mathbb{E}\left[\exp\left(2\pi i \langle \boldsymbol{Y}, \boldsymbol{\theta}\rangle\right)\right] = \exp\Big(2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta}\rangle - 2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}\Big) \quad \forall \boldsymbol{\theta} \in \mathbb{R}^A$$

The continuous version of the Fourier inversion formula tells us that

$$f_{\boldsymbol{Y}}(\boldsymbol{x}) = \int_{\mathbb{R}^A} \hat{\boldsymbol{Y}}(\boldsymbol{\theta}) \cdot \exp\left(-2\pi i \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle\right) d\boldsymbol{\theta} \quad \forall \boldsymbol{x} \in \mathbb{R}^A$$

is the recovery of the density function from the Fourier coefficients (again, we skip a proof of this fact known from your probability 101 course). We know that the density of the Gaussian is significant around the expectation, in this case $f_{\boldsymbol{Y}}(\mathbb{E}[\boldsymbol{X}]) = \frac{1}{(2\pi)^{|A|/2}\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}$. In particular we can relate $\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]]$ to that value. More generally,

$$\begin{aligned} &\Big|\Pr[\boldsymbol{X} = \boldsymbol{\lambda}] - \det(\mathcal{L}) \cdot f_{\boldsymbol{Y}}(\boldsymbol{\lambda})\Big| \\ \leq\ & \det(\mathcal{L}) \cdot \Big[\underbrace{\int_{\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{=I_1} + \underbrace{\int_{D^*\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{X}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{=I_2} + \underbrace{\int_{\mathbb{R}^A\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta}}_{=I_3}\Big] \end{aligned}$$

It remains to bound the integrals I_1, I_2, I_3 and hope that the error is less than $\det(\mathcal{L}) \cdot f_{\boldsymbol{Y}}(\mathbb{E}[\boldsymbol{X}])$.

Bounding I_2. First, we already learned in Lemma 8.12 that Fourier coefficients $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|$ decay exponentially outside of $\mathcal{B}_{\boldsymbol{R}}(\varepsilon)$ (but inside of the Voronoi cell).

Lemma 8.14 (Bounding I_2)**.** *If $0 < \varepsilon < \frac{1}{4K^6}$, then one has $\int_{D^*\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{X}}(\boldsymbol{\theta})| d\boldsymbol{\theta} \leq \frac{\exp(-\frac{\varepsilon^2}{2}N)}{\det(\mathcal{L})}$.*

Proof. We use the bound on the Fourier coefficient $|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|$ from Lemma 8.12 to get

$$\int_{D^*\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} \underbrace{|\hat{\boldsymbol{X}}(\boldsymbol{\theta})|}_{\leq\exp(-\frac{\varepsilon^2}{2}N)} \ d\boldsymbol{\theta} \leq \exp\Big(-\frac{\varepsilon^2}{2}N\Big) \cdot \underbrace{\mathrm{vol}_n(D^*)}_{=\det(\mathcal{L}^*)=\frac{1}{\det(\mathcal{L})}} \leq \frac{\exp(-\frac{\varepsilon^2}{2}N)}{\det(\mathcal{L})}$$

□

Bounding I_3. We will use a short auxiliary lemma to bound exponentially decaying integrals:

Lemma 8.15. *Let $\boldsymbol{S} \in \mathbb{R}^{A\times A}$ be a symmetric, positive semidefinite matrix. Then for $r \geq 10\sqrt{|A|}$*

$$\int_{\{\boldsymbol{x}\in\mathbb{R}^A | \boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x}\geq r^2\}} \exp(-\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x}) d\boldsymbol{x} \leq \exp(-\Theta(r^2))$$

Proof. We use the integral transformation formula[3]

$$\begin{aligned}\int_{\{\boldsymbol{x}\in\mathbb{R}^A \mid \boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x}\geq r^2\}} \exp(-\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x})d\boldsymbol{x} &= \frac{1}{\sqrt{\det(2\boldsymbol{S})}}\int_{\|\boldsymbol{x}\|_2\geq\sqrt{2}r} \exp(-\|\boldsymbol{x}\|_2^2/2) \\ &= \frac{(2\pi)^{|A|/2}}{\sqrt{\det(2\boldsymbol{S})}} \Pr_{\boldsymbol{x}\sim N_A(0,1)}\Big[\|\boldsymbol{x}\|_2 \geq \underbrace{\sqrt{2}r}_{\geq\mathbb{E}[\|\boldsymbol{x}\|_2]+r}\Big] \\ &\leq 2\cdot\frac{(2\pi)^{|A|/2}}{\sqrt{\det(2\boldsymbol{S})}}\exp(-r^2/2) \leq \frac{\exp(-\Theta(r^2))}{\sqrt{\det(\boldsymbol{S})}}\end{aligned}$$

where $N_A(0,1)$ is the distribution of standard Gaussian in $\mathbb{R}^A$. □

The next step is to bound the Fourier coefficients of the Gaussian outside or $\mathcal{B}_{\boldsymbol{R}}(\varepsilon)$, which quickly follows from integrating:

Lemma 8.16 (Bounding I_3)**.** *If $\varepsilon \geq C\frac{\sqrt{K}}{\sqrt{N}}$ for a large enough constant $C > 0$, then*

$$\int_{\mathbb{R}^A\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta} \leq \frac{\exp(-\Theta(\varepsilon^2 N))}{\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}$$

Proof. First recall that the covariance matrix is

$$\boldsymbol{\Sigma}[\boldsymbol{X}] = p(1-p)\cdot\boldsymbol{R} \geq \frac{N}{2|B|}\boldsymbol{R}$$

and hence $\|\boldsymbol{\theta}\|_{\boldsymbol{R}}^2 = \frac{1}{|B|}\cdot\boldsymbol{\theta}^T\boldsymbol{R}\boldsymbol{\theta} \leq \frac{2}{N}\cdot\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}$ which we use in $(*)$. We can write

$$\begin{aligned}\int_{\mathbb{R}^A\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{Y}(\boldsymbol{\theta})|d\boldsymbol{\theta} &= \int_{\mathbb{R}^A\setminus\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} \left|\exp\left(2\pi i\,\langle\mathbb{E}[\boldsymbol{X}],\boldsymbol{\theta}\rangle - 2\pi^2\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}\right)\right| d\boldsymbol{\theta} \\ &\overset{|e^{2\pi i\alpha}|\leq 1\ \&\ (*)}{\leq} \int_{\mathbb{R}^A\setminus\{\boldsymbol{\theta}\mid\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}\geq\frac{1}{2}\varepsilon^2 N\}} \exp(-2\pi^2\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &\overset{\text{Lem. 8.15}}{\leq} \frac{\exp(-\Theta(\varepsilon^2 N))}{\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}]))}}\end{aligned}$$

where the last inequality uses that indeed $\varepsilon^2 N \geq \Theta(|A|)$. □

[3] Recall that the transformation formula is $\int_{\mathbb{R}^n} f(\boldsymbol{Q}\boldsymbol{x})d\boldsymbol{x} = \frac{1}{\det(\boldsymbol{Q})}\int_{\mathbb{R}^n} f(\boldsymbol{x})d\boldsymbol{x}$ where $\boldsymbol{Q}$ is a regular square matrix and we apply it with $\boldsymbol{Q} := \sqrt{2}\cdot\boldsymbol{S}^{1/2}$ and $f(\boldsymbol{x}) = \mathbf{1}_{\|\boldsymbol{x}\|_2\geq\sqrt{2}r}\cdot\exp(-\|\boldsymbol{x}\|_2^2/2)$.

Bounding I_1. Again, a short lemma:

Lemma 8.17. *For any positive semidefinite matrix $\boldsymbol{S} \in \mathbb{R}^{A \times A}$ one has*

$$\int_{\mathbb{R}^A} \exp(-\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x}) \cdot |\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x}|^{3/2} d\boldsymbol{x} \leq \frac{(2\pi)^{|A|/2}}{\sqrt{\det(2\boldsymbol{S})}} \cdot O(|A|^{3/2})$$

Proof. We have

$$\begin{aligned}
\int_{\mathbb{R}^A} \exp(-\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x}) \cdot |\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x}|^{3/2} d\boldsymbol{x} \quad &\overset{\text{transformation}}{=} \quad \frac{O(1)}{\sqrt{\det(2\boldsymbol{S})}} \int_{\mathbb{R}^A} \exp(-\|\boldsymbol{x}\|_2^2/2) \cdot \|\boldsymbol{x}\|_2^3 d\boldsymbol{x} \\
&= \quad \frac{O(1) \cdot (2\pi)^{|A|/2}}{\sqrt{\det(2\boldsymbol{S})}} \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_A(0,1)} [\|\boldsymbol{x}\|_2^3]}_{O(|A|^{3/2})}
\end{aligned}$$

□

Finally, we bound the difference between $\boldsymbol{X}$ and the Gaussian $\boldsymbol{Y}$ that comes from the Fourier coefficients close to the origin. While we will be able to make the error terms I_2, I_3 exponentially small compared to the Gaussian density, the error I_1 will actually be of the form $\frac{\text{poly}(K)}{\sqrt{N}}$ times the Gaussian density. Hence this is the only significant error term.

Lemma 8.18 (Bounding I_1)**.** *One has*

$$\int_{\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta} \leq O\Big(\frac{K^{4.5}}{\sqrt{N} \cdot (2\pi)^{|A|/2} \sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}\Big)$$

Proof. First note that by Lemma 8.10, for an individual $\boldsymbol{\theta} \in \mathcal{B}_{\boldsymbol{R}}(\varepsilon)$ there is a δ with $|\delta| \leq O(K^3 \|\boldsymbol{\theta}\|_R^3 N)$ so that

$$\begin{aligned}
|\hat{\boldsymbol{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})| \quad &= \quad \Big| \exp\Big(2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta} \rangle - 2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta} + \delta\Big) - \exp\Big(2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta} \rangle - 2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta}\Big) \Big| \\
&\leq \quad \underbrace{|e^\delta - 1|}_{\leq 2|\delta|} \cdot \exp\big(-2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta}\big) \cdot \underbrace{|\exp(2\pi i \langle \mathbb{E}[\boldsymbol{X}], \boldsymbol{\theta} \rangle)|}_{\leq 1}
\end{aligned}$$

Then plugging this in, we get

$$\begin{aligned}
\int_{\mathcal{B}_{\boldsymbol{R}}(\varepsilon)} |\hat{\boldsymbol{X}}(\boldsymbol{\theta}) - \hat{\boldsymbol{Y}}(\boldsymbol{\theta})| d\boldsymbol{\theta} \quad &\leq \quad O(K^3 N) \cdot \int_{\mathbb{R}^A} \|\boldsymbol{\theta}\|_R^3 \cdot \exp(-2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\overset{(*)}{=} \quad O\Big(\frac{K^3 N}{N^{3/2}}\Big) \cdot \int_{\mathbb{R}^A} |\boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta}|^{3/2} \exp(-2\pi^2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}[\boldsymbol{X}] \boldsymbol{\theta}) d\theta \\
&\overset{\text{Lemma 8.17}}{\leq} \quad O\Big(\frac{K^3}{\sqrt{N}}\Big) \cdot \frac{(2\pi)^{|A|/2}}{\sqrt{\det(4\pi^2 \boldsymbol{\Sigma}[\boldsymbol{X}])}} \cdot O(|A|^{3/2}) \\
&\leq \quad \frac{O(K^{4.5})}{\sqrt{N} \cdot (2\pi)^{|A|/2} \sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}
\end{aligned}$$

applying Lemma 8.17 with $\boldsymbol{S} := 2\pi^2\boldsymbol{\Sigma}[\boldsymbol{X}]$. Note that we also have been using that

$$\|\boldsymbol{\theta}\|_R = \Big(\frac{1}{|B|}\boldsymbol{\theta}^T\boldsymbol{R}\boldsymbol{\theta}\Big)^{1/2} = \Theta(1)\cdot\Big(\frac{1}{p|B|}\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}\Big)^{1/2} = \Theta(1)\cdot\Big(\frac{1}{N}\boldsymbol{\theta}^T\boldsymbol{\Sigma}[\boldsymbol{X}]\boldsymbol{\theta}\Big)^{1/2} \qquad (*)$$

□

Finishing the main proof

First note that we assume that $\boldsymbol{M} \in \{0,1\}^{B\times A}$ and hence $\boldsymbol{\Sigma}[\boldsymbol{X}] \in \mathbb{Z}^{A\times A}$ and thus $\det(\boldsymbol{\Sigma}[\boldsymbol{X}]) \geq 1$ using that $\boldsymbol{M}$ has full column rank. On the other hand entries in the covariance matrix are bounded by $\|\boldsymbol{\Sigma}[\boldsymbol{X}]\|_\infty = p(1-p)\cdot\max_{a,a'\in A}|\langle\boldsymbol{M}^a,\boldsymbol{M}^{a'}\rangle| \leq p|B| = N$ and then by Hadamard's inequality generously we get $\det(\boldsymbol{\Sigma}[\boldsymbol{X}]) \leq (KN)^K$. For the lattice $\mathcal{L}$ it will suffice that $\det(\mathcal{L}) \geq 1$ by integrality. Then

$$\frac{\Pr[\boldsymbol{X} = \mathbb{E}[\boldsymbol{X}]]}{\det(\mathcal{L})} \geq f_{\boldsymbol{Y}}(\mathbb{E}[\boldsymbol{X}]) - I_1 - I_2 - I_3$$

$$\geq \frac{1}{(2\pi)^{|A|/2}\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}} - \underbrace{\frac{O(K^{4.5})}{\sqrt{N}\cdot(2\pi)^{|A|/2}\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}}_{\text{small if } N\geq K^{10}} - \underbrace{\frac{\exp(-\frac{\varepsilon^2}{2}N)}{\det(\mathcal{L})}}_{\text{small if } \varepsilon\geq\frac{4K}{\sqrt{N}}\ln(KN)} - \underbrace{\frac{\exp(-\Theta(\varepsilon^2 N))}{\sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}}}_{\text{small if } \varepsilon\geq\frac{K}{\sqrt{N}}} > 0$$

For the estimate of I_2, note that $\exp(-\frac{\varepsilon^2}{2}N) \ll \sqrt{\det(\boldsymbol{\Sigma}[\boldsymbol{X}])}$ as long $\varepsilon^2 N \geq 4K\ln(KN)$. Overall, a choice of $\varepsilon := \frac{8K}{\sqrt{N}}\ln(N)$ and $N \geq K^{10}$ works.

8.4 Application to orthogonal arrays

Now we will fill-in the formal details, how to apply the Kuperberg-Lovett-Peled Theorem to show existence of t-wise orthogonal arrays. We fix n and t. Again, let $B := \{0,1\}^n$ be the set of all 0/1 strings as row vectors. Similar to before, for an index set $I \subseteq [n]$ and assignments $\boldsymbol{z} \in \{0,1\}^I$ we define a column vector $\boldsymbol{M}^{(I,\boldsymbol{z})} \in \{0,1\}^B$ with

$$M_{(I,\boldsymbol{z}),\boldsymbol{x}} := \begin{cases} 1 & x_i = z_i \ \forall i \in I \\ 0 & \text{otherwise.} \end{cases} \qquad \forall \boldsymbol{x} \in B$$

Recall that we had previously used the columns $A := \{(I,\boldsymbol{z}) \mid |I| = t, \boldsymbol{z} \in \{0,1\}^I\}$ and the matrix $\boldsymbol{M} := (\boldsymbol{M}^a)_{a\in A}$. It will be convenient to study in parallel the alternative set of column indices $\tilde{A} := \{(I,\mathbf{1}) \mid |I| \leq t\}$ inducing a matrix $\tilde{\boldsymbol{M}} := (\boldsymbol{M}^a)_{a\in\tilde{A}}$. Both choices are actually equivalent as $\boldsymbol{x} \sim T$ is going to be a uniform distribution on all subsets of t variables if and only if $\Pr_{\boldsymbol{x}\sim T}[\bigwedge_{i\in I}(x_i = 1)] = 2^{-|I|}$ for all $|I| \leq t$. We could in principle use either matrix $\boldsymbol{M}$ or $\tilde{\boldsymbol{M}}$, but some properties will be easier to justify for $\boldsymbol{M}$ and other easier for $\tilde{\boldsymbol{M}}$ — and this is admissible as the space spanned by their columns is the same.

Lemma 8.19. *One has* $V := \mathit{span}\{\boldsymbol{M}^a \mid a \in A\} = \mathit{span}\{\boldsymbol{M}^a \mid a \in \tilde{A}\}$.

Proof. For $|I| \leq t$ one can fix any $I \subseteq J \subseteq [n]$ with $|J| = t$ and simply sum up over the matching atomic events to get $\boldsymbol{M}^{(I,\mathbf{1})} = \sum_{\boldsymbol{z}\in\{0,1\}^J : z_i = 1 \forall i \in I} \boldsymbol{M}^{(J,\boldsymbol{z})}$. In reverse, for $(I, \boldsymbol{z}) \in A$ one can abbreviate $I_0 := \{i \in I \mid z_i = 0\}$ and $I_1 := \{i \in I \mid z_i = 1\}$ and use the *inclusion-exclusion formula* to get

$$\boldsymbol{M}^{(I,\boldsymbol{z})} = \sum_{J \subseteq I_0} (-1)^{|J|} \boldsymbol{M}^{(I_1 \cup J, \mathbf{1})}$$

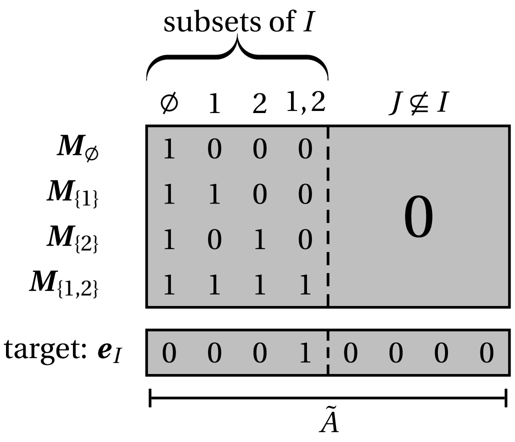
That means every column in $\boldsymbol{M}$ and $\tilde{\boldsymbol{M}}$, resp., can be written as a linear combination of rows of the other matrix. The claim follows. □

All-ones function. First of all, note that the matrix $\tilde{\boldsymbol{M}}$ also includes the all-ones column vector $\boldsymbol{M}^{(\emptyset,\emptyset)}$, which satisfies the last condition of the KLP-Theorem.

Local decodability. This is the only part for orthogonal arrays that need some insight:

Lemma 8.20. *For each* $a \in \tilde{A}$, *there is a* $\boldsymbol{y} \in \mathbb{Z}^B$ *with* $\boldsymbol{e}_a = \boldsymbol{y}^T \boldsymbol{M}$ *and* $\|\boldsymbol{y}\|_1 \leq 2^t$.

Proof. For convenience reasons, we index the columns by $|I| \leq t$. By a slight abuse of notation, we can also denote $\boldsymbol{M}_J := \boldsymbol{M}_{\mathbf{1}_J} \in \{0,1\}^{\tilde{A}}$ as the row induced by the characteristic vector of J. We fix an index set I with $|I| \leq t$ and claim that there is a short integer vector $\boldsymbol{y}$ so that $\boldsymbol{e}_I = \sum_{J \subseteq I} y_J \boldsymbol{M}_J$, where $\boldsymbol{e}_I \in \{0,1\}^{\tilde{A}}$ is the target unit vector. Note that all those rows M_J with $J \subseteq I$ are 0 on columns indexed by non-subsets of I.



$$\begin{array}{r|cccc:cccc|}
 & \emptyset & 1 & 2 & 1,2 & \multicolumn{4}{c}{J \not\subseteq I} \\ \hline
\boldsymbol{M}_{\emptyset} & 1 & 0 & 0 & 0 & & & & \\
\boldsymbol{M}_{\{1\}} & 1 & 1 & 0 & 0 & \multicolumn{4}{c|}{\mathbf{0}} \\
\boldsymbol{M}_{\{2\}} & 1 & 0 & 1 & 0 & & & & \\
\boldsymbol{M}_{\{1,2\}} & 1 & 1 & 1 & 1 & & & & \\ \hline
\text{target: } \boldsymbol{e}_I & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline
\end{array}$$

Example: rows $\boldsymbol{M}_J$ with $J \subseteq I$ for $t = 2$

Then the inclusion exclusion formula gives

$$\boldsymbol{e}_I = \sum_{J \subseteq I} \underbrace{(-1)^{|I \setminus J|}}_{=:y_J} \boldsymbol{M}_J$$

and clearly $\|\boldsymbol{y}\|_1 \leq 2^{|I|} \leq 2^t$. □

Divisibility. First a quick consequence of the previous lemma:

Corollary 1. *One has $\mathcal{L}(\tilde{\boldsymbol{M}}) = \mathbb{Z}^{\tilde{A}}$.*

Proof. Lemma 8.20 implies in particular that $\boldsymbol{e}_a \in \mathcal{L}(\tilde{\boldsymbol{M}})$. Then indeed, integer combinations of the rows of $\tilde{\boldsymbol{M}}$ must give the whole $\mathbb{Z}^{\tilde{A}}$. □

Note that in each column $M^{(I,\mathbf{1})}$ exactly a $2^{-|I|}$ fraction of entries are 1's. That means as long as N is an integer multiple of 2^t we can be sure that $\frac{N}{|B|} \sum_{b \in B} M_b^{(I,1)} \in \mathbb{Z}$. Then $\tilde{\boldsymbol{M}}$ satisfies the divisibility condition with constant 2^t.

Symmetry. Now we come to the symmetry condition:

Lemma 8.21. *For every $b_1, b_2 \in B$, there is permutation $\pi : B \to B$ so that $\pi(b_1) = b_2$ and $(M_{\pi(b),a})_{b \in B} \in V$ for all $a \in A$.*

Proof. For a vector $x \in \{0,1\}^n$ we define a permutation on the row indices

$$\pi_x : B \to B \quad \text{with} \quad \pi_x(b) = b \oplus x \quad \text{for} \quad b \in B = \{0,1\}^n$$

where $\oplus$ is the addition modulo 2. Then $\pi_{b_1 \oplus b_2}(b_1) = b_2$, that means the set of permutations $(\pi_x)_{x \in \{0,1\}^n}$ can shuffle any row index to any other row index. Next, note that if we apply the permutation π_x to the entries of a column vector we obtain

$$(M_{\pi_x(b),(I,z)})_{b\in B} = (M_{x\oplus b,(I,z)})_{b\in B} \overset{(*)}{=} (M_{b,(I,z\oplus x)})_{b\in B} = M^{(I,z\oplus x)} \in V$$

Note that this vector is again a column vector of $\boldsymbol{M}$ and hence also in the vector space V. Here we have used in $(*)$ that $x_i \oplus b_i = z_i \Leftrightarrow b_i = z_i \oplus x_i$. $\square$

Summary. Finally, we can apply the KLP-Theorem with matrix $\tilde{\boldsymbol{M}}$ and parameter $K := \max\{\dim(V), 2^t\} = n^{\Theta(t)}$.

8.5 Open problems

The KLP-Theorem is purely existential in the sense that it does not provide an algorithm that could find the set T in time polynomial in the size of $\boldsymbol{M}$. It is an interesting open problem, whether there is a polynomial time algorithm (possibly randomized) for the same task.

To understand, what properties of a matrix are needed to make the presented arguments work, consider the following result of Alon and Vu [AV97]: There is a $0/1$ matrix $\boldsymbol{M}$ with $n^{\Theta(n)}$ rows and n columns that all have the same number of 1-entries. Still no proper subset T will have the same row average as the whole matrix.

It would be interesting whether there is a simpler form of the KLP-Theorem with fewer assumptions or assumptions that are easier to verify.

8.6 Exercises

Exercise 8.1. Fix $m \in \mathbb{N}$ and an even integer n with $n \geq C \cdot m^C$ where $C > 0$ is a large enough constant. Consider a random matrix $\boldsymbol{A} \in \{-1,1\}^{m\times n}$ where each entry is picked uniformly and independently drawn from $\{-1,1\}$. In this exercise we want to give a Fourier analysis proof for the fact that with high probability there is an $\boldsymbol{x} \in \{-1,1\}^n$ with $\boldsymbol{Ax} = \boldsymbol{0}$. For this sake, draw $\boldsymbol{x} \in \{-1,1\}^n$ uniformly at random and abbreviate $\boldsymbol{X} := \boldsymbol{Ax} \in \mathbb{Z}^m$. We will study the Fourier coefficients $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}[\exp(2\pi i\,\langle \boldsymbol{X},\boldsymbol{\theta}\rangle)]$ where $\boldsymbol{\theta} \in \mathbb{R}^m$.

(a) Show that $\Pr_{\boldsymbol{x}}[\boldsymbol{X} = \boldsymbol{0}] = 2^m \int_{[-\frac{1}{4},\frac{1}{4}[^m} \hat{\boldsymbol{X}}(\boldsymbol{\theta})\,d\boldsymbol{\theta}$.

(b) Show that $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \prod_{j=1}^n \cos(2\pi\,\langle \boldsymbol{A}^j,\boldsymbol{\theta}\rangle)$ for all $\boldsymbol{\theta} \in \mathbb{R}^m$.

(c) Show that for every outcome of $\boldsymbol{A}$ and every $\|\boldsymbol{\theta}\|_2 \leq \frac{1}{8\sqrt{m}}$ one has $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) = \exp(-2\pi^2\boldsymbol{\theta}^T(\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{\theta} + \delta(\boldsymbol{\theta}))$ with an error term of $|\delta(\boldsymbol{\theta})| \leq O(\sum_{j=1}^n \langle \boldsymbol{A}^j, \boldsymbol{\theta}\rangle^4)$.

(d) Show that for every outcome of $\boldsymbol{A}$ and every $\|\boldsymbol{\theta}\|_2 \leq \frac{1}{8\sqrt{m}}$ one has $\hat{\boldsymbol{X}}(\boldsymbol{\theta}) \geq \exp(-\Theta(nm) \cdot \|\boldsymbol{\theta}\|_2^2) > 0$.

(e) Prove that for every fixed $\boldsymbol{\theta} \in [-\frac{1}{4}, \frac{1}{4}[^m$ one has $\Pr_{\boldsymbol{A}}[|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp(-\Theta(\frac{n}{\text{poly}(m)}) \cdot \|\boldsymbol{\theta}\|_2^2)] \geq 1 - \exp(-n/\text{poly}(m))$.

(f) Prove that $\Pr_{\boldsymbol{A}}[|\hat{\boldsymbol{X}}(\boldsymbol{\theta})| \leq \exp(-\Theta(\frac{n}{\text{poly}(m)}) \cdot \|\boldsymbol{\theta}\|_2^2)\ \forall \boldsymbol{\theta} \in [\frac{1}{4}, \frac{1}{4}[^m] \geq 1 - \exp(-n/\text{poly}(m))$.

(g) Show that with high probability over the choice of $\boldsymbol{A}$ one has $\int_{\boldsymbol{\theta} \in D^+} \hat{\boldsymbol{X}}(\boldsymbol{\theta}) d\boldsymbol{\theta} > (\frac{1}{\text{poly}(m,n)})^m$ and $\int_{\boldsymbol{\theta} \in D^-} |\hat{\boldsymbol{X}}(\boldsymbol{\theta})| d\boldsymbol{\theta} \leq \exp(-n/\text{poly}(m))$ where $D^+ := \{\boldsymbol{\theta} \in [-\frac{1}{4}, \frac{1}{4}[^m \colon \hat{\boldsymbol{X}}(\boldsymbol{\theta}) \geq 0\}$ and $D^- := \{\boldsymbol{\theta} \in [-\frac{1}{4}, \frac{1}{4}[^m \colon \hat{\boldsymbol{X}}(\boldsymbol{\theta}) < 0\}$.

(h) Prove that with high probability over the choice of $\boldsymbol{A}$ one has $\Pr_{\boldsymbol{x}}[\boldsymbol{X} = \boldsymbol{0}] > 0$.

Hint. You may use the inequality $\exp(-\frac{1}{2}z^2 - z^4) \leq \cos(z) \leq \exp(-\frac{1}{2}z^2 + z^4)$ for $|z| \leq \frac{1}{8}$ without proving it.

Bibliography

[AKS81] Miklós Ajtai, János Komlós, and Endre Szemerédi. A dense infinite Sidon sequence. *European J. Combin.*, 2(1):1–11, 1981.

[Alo96] Noga Alon. Independence numbers of locally sparse graphs and a Ramsey type problem. *Random Structures Algorithms*, 9(3):271–278, 1996.

[AS16] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, fourth edition, 2016.

[AV97] Noga Alon and Van H Vu. Anti-hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs. *Journal of Combinatorial Theory, Series A*, 79(1):133 – 160, 1997.

[BRSW06] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2-source dispersers for sub-polynomial entropy and ramsey graphs beating the frankl-wilson construction. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA, May 21-23, 2006*, pages 671–680, 2006.

[CFS15] David Conlon, Jacob Fox, and Benny Sudakov. Recent developments in graph ramsey theory. In *Surveys in Combinatorics 2015*, pages 49–118. 2015.

[CGH10] Karthekeyan Chandrasekaran, Navin Goyal, and Bernhard Haeupler. Deterministic algorithms for the lovÁsz local lemma. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 992–1004, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[EL75] P. Erdős and L. Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. pages 609–627. Colloq. Math. Soc. János Bolyai, Vol. 10, 1975.

[FS11] Jacob Fox and Benny Sudakov. Dependent random choice. *Random Struct. Algorithms*, 38(1-2):68–99, 2011.

[KLP17] Greg Kuperberg, Shachar Lovett, and Ron Peled. Probabilistic existence of regular combinatorial structures. *Geometric and Functional Analysis*, 27(4):919–972, Jul 2017.

[Mat02] Jiri Matousek. *Lectures on discrete geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[MSS15] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families ii: Mixed characteristic polynomials and the kadison-singer problem. *Annals of Mathematics*, 182(1):327–350, 2015.

[MT10] Robin A. Moser and Gábor Tardos. A constructive proof of the general Lovász local lemma. *J. ACM*, 57(2):Art. 11, 15, 2010.

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and computing*. Cambridge University Press, Cambridge, second edition, 2017. Randomization and probabilistic techniques in algorithms and data analysis.

[She83] James B. Shearer. A note on the independence number of triangle-free graphs. *Discrete Math.*, 46(1):83–87, 1983.

[TV10] Terence Tao and Van H. Vu. *Additive combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010. Paperback edition [of MR2289012].