Partial coloring from volume lower bounds

Talk by Vishesh Jain
Notes by Sanjana Das
April 19, 2024

§1 Motivation

One piece of motivation for what we'll discuss today comes from Spencer's theorem on low-discrepancy colorings (or signings — we'll use the two words interchangeably).

Theorem 1.1 (Spencer)
Given sets $S_1, \dots, S_m \subseteq [n]$, there exists a signing $x \in \{\pm 1\}^n$ such that

$$\max_{j \in [m]} \left| \sum_{i \in S_j} x_i \right| \lesssim \sqrt{n \log\left(\frac{m}{n} + 2\right)}.$$

(The $+2$ is not important; it's just to not have issues when $m \leq n$, where we want a bound of $\sqrt{n}$.)

In words, we've got a set system over a common ground set $[n]$, and we're trying to get a low-discrepancy signing. A *random* assignment would have a discrepancy of $\sqrt{n \log m}$, and Spencer's theorem says that we can actually do significantly better than this — we can instead get $\sqrt{n \log(m/n)}$. We're most interested in the case where $m = n$, in which case the log factor entirely disappears.

The proof of Spencer's theorem involves the *partial coloring method* — first, instead of considering full signings $x \in \{\pm 1\}^n$, we try to find *partial* signings $x \in \{-1, 0, 1\}^n$.

Theorem 1.2 (Partial coloring)
Given sets $S_1, \dots, S_m \subseteq [n]$, there exists a 'partial signing' $x \in \{-1, 0, 1\}^n$ with at least $n/2$ coordinates assigned ± 1 such that

$$\max_{j \in [m]} \left| \sum_{i \in S_j} x_i \right| \lesssim \sqrt{n \log\left(\frac{m}{n} + 2\right)}.$$

This statement about partial colorings suffices to prove Theorem 1.1 (on *full* colorings) because we can iterate — suppose we've got a partial signing $x \in \{-1, 0, 1\}^n$ as in Theorem 1.2. Then we can fix the coordinates assigned ± 1 and iterate on the coordinates assigned 0 — so we color the elements that are colored by x and leave the rest temporarily uncolored, and we pass down to the set system on the at most $n/2$ uncolored elements. And then we apply Theorem 1.2 to partially color those and pass down again, and so on. Then the total discrepancy we get will be at most roughly $(\sqrt{n} + \sqrt{n/2} + \sqrt{n/4} + \cdots) \log(m/n + 2)$. And this is a geometric series, so when we sum it, we'll still get a bound of $O(\sqrt{n \log(m/n + 2)})$.

Today we're going to discuss how you can prove partial coloring results such as Theorem 1.2 using volume lower bounds, and we'll see some nice proofs.

§2 The geometric partial coloring lemma

Spencer proved this result on partial coloring using entropy; but there was another independent proof by Gluskin, who took a more geometric point of view. Specifically, he proved the following statement, known as the geometric partial coloring lemma.

Notation 2.1. For $K \subseteq \mathbb{R}^n$, we use $\gamma(K)$ to denote the Gaussian volume of K — i.e., the probability that a standard n-dimensional Gaussian lands in K.

Theorem 2.2 (Gluskin 1980s, Giannapoulos 1990s)
For all sufficiently small $\varepsilon > 0$, there exists δ such that if $K \subseteq \mathbb{R}^n$ is a symmetric convex body with $\gamma(K) \geq e^{-\varepsilon n}$, then there exists $x \in [-1,1]^n \cap K$ such that at least δn coordinates of x are $\{\pm 1\}$.

Remark 2.3. The quantifiers are a bit strange here in that making ε smaller makes the hypothesis stronger, so it allows you to make δ *bigger*. (We could have also quantified this as 'for all sufficiently small ε and δ.')

(By *symmetric* we mean symmetric with respect to the origin — so if $x \in K$, then $-x \in K$ as well.)

§2.1 The connection to partial coloring

First, what does Theorem 2.2 have to do with partial coloring? For convenience, we'll think of m as being on the order of n (so that we want a partial coloring with discrepancy on the order of $\sqrt{n}$ — otherwise we need a log factor). Then we want to try to get a partial coloring with maximum discrepancy $C\sqrt{n}$, so we define the body

$$K_C = \left\{ x \in \mathbb{R}^n \;\middle|\; \left| \sum_{i \in S_j} x_i \right| \leq C\sqrt{n} \text{ for all } j \in [m] \right\}.$$

This set is clearly convex (as it's an intersection of strips) and symmetric (because we're taking absolute values). And what's the Gaussian volume of K_C? To bound $\gamma(K_C)$, we'll write K_C as an intersection

$$K_C = \bigcap_{j \in [m]} \left\{ x \in \mathbb{R}^n \;\middle|\; \left| \sum_{i \in S_j} x_i \right| \leq C\sqrt{n} \right\}.$$

And there's a fact, the *Gaussian correlation inequality*, that allows us to lower-bound the volume of an intersection of symmetric convex bodies.

Theorem 2.4 (Gaussian correlation inequality; Royen)
If K and L are symmetric convex bodies in $\mathbb{R}^n$, then $\gamma(K \cap L) \geq \gamma(K)\gamma(L)$.

Remark 2.5. This was a conjecture for a while (called the *Gaussian correlation conjecture*). Here, we only really need the theorem in the case where L is a strip; in that setting it's much easier to prove (and the statement is called *Sidak's lemma*). (We use the full version here just because even if Sidak's lemma is simpler to prove, they're the same complexity to cite.)

And for each j, the volume of the jth strip is the probability that $|S_j| \leq n$ independent standard Gaussians sum to at most $C\sqrt{n}$. And if we take C to be large, then this volume becomes very close to 1 (because the

sum of these standard Gaussians has standard deviation at most $\sqrt{n}$). This means we get a bound

$$\gamma(K_C) \geq e^{-\varepsilon_C m}$$

where $\varepsilon_C \to 0$ as $C \to \infty$ (since we've got m of these strips, each of which we can say has volume at least $e^{-\varepsilon_C}$). And now if we think of m as roughly n, then this is exactly the condition we need to be able to apply Theorem 2.2. This gives us a point $x \in [-1,1]^n \cap K$ with a decent number of ± 1's (we think of the coordinates with ± 1's as colored, and iterate on the remaining ones, which we think of as uncolored).

There's a slight cheat here — unlike in Theorem 1.2 (which gave us a partial coloring where the colored coordinates got ± 1's and the uncolored ones all got 0's), here the output of Theorem 2.2 is going to have uncolored coordinates which could be fractions. This makes the iteration messier; to fix this, you need a slightly more general statement. Essentially, if you've got an uncolored assignment that got assigned 0.7, then to color it you sort of rescale the coordinate (so you only go up to 0.3), and things are still fine.

§2.2 A constructive version of the partial coloring lemma

Now we're going to try to understand why the geometric partial coloring lemma (Theorem 2.2) is true. Theorem 2.2 is a purely existential statement — it just says there *exists* a point $x \in K \cap [-1,1]^n$ with a substantial number of ± 1-coordinates. But it turns out that we can actually get an *algorithmic* version (i.e., a statement that tells us how to actually construct such an x), and this is what we'll prove instead.

Theorem 2.6 (Constructive partial coloring; Rothvoss 2014, Eldan–Singh 2015)
In the same setup as Theorem 2.2, if we take $g \sim \mathcal{N}(0, \mathrm{Id}_n)$, then with probability at least $\frac{1}{2}$, the point

$$x^* = \operatorname*{argmax}_{x \in K \cap [-1,1]^n} \langle g, x \rangle$$

satisfies the conclusion of Theorem 2.2 (meaning that it has at least δn ± 1's).

So not only are we guaranteed that there *exists* a point $x \in K \cap [-1,1]^n$ with lots of 1's, but that point can even be obtained by solving a linear program (specifically, the linear program where we've got constraints $-1 \leq x_i \leq 1$ (and constraints corresponding to K) and we're trying to maximize $\langle g, x \rangle$, for fixed g).

First, as a sanity check, let's make sure that this is true when $K = \mathbb{R}^n$ is the entire space (so that $\gamma(K) = 1$). Then the x^* obtained from Theorem 2.6 corresponds to maximizing a linear program on the cube $[-1,1]^n$. A linear program always has a maximizer which is a vertex, and a vertex of the cube has *all* coordinates ± 1. This doesn't immediately imply the statement we want (with $\delta = 1$) — a linear program might not have a *unique* optimizer (so for example, the set of optimizers could be an entire face). But over the random choice of g, the probability this happens is 0 (as it requires a random Gaussian to be constant on a face of the cube). So the x^* from Theorem 2.6 has *all* coordinates ± 1 with probability 1.

§2.3 Some intuition

First, here's some intuition on why we'd expect Theorem 2.6 to be true. Let's assume it's not true, and let B be the set of 'bad' realizations of the Gaussian — i.e., B is the set of $g \in \mathbb{R}^n$ for which the corresponding x^* does *not* have at least δn coordinates of ± 1. We'll let $S(g) \subseteq [n]$ be the set of ± 1-coordinates — so for each $g \in B$, we have $|S(g)| \leq \delta n$, and $x^* = \pm 1$ only for $i \in S(g)$.

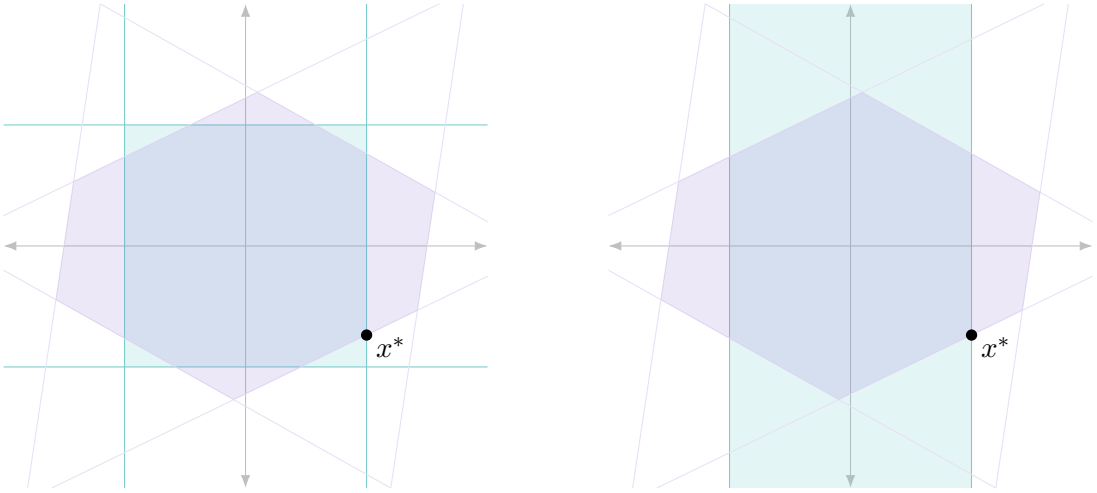
Claim 2.7 — For any $g \in B$, we have

$$\max_{x \in K \cap [-1,1]^n} \langle g, x \rangle = \max_{x \in K \cap [-1,1]^{S(g)}} \langle g, x \rangle.$$

(In the linear program on the left-hand side we're constraining *every* coordinate to be between -1 and 1, while on the right-hand side we're only constraining the coordinates in $S(g)$ — which is a very small fraction of all the coordinates).

There are a couple of ways of seeing why Claim 2.7 (the relevant keyword is 'complementary slackness'). One way to think of this is that we've got a linear program, and the optimum of any linear program is attained at a vertex; and any vertex of a polytope is obtained by making some set of n inequalities tight. By throwing away the inequalities that we did, we don't lose x^* (i.e., it's still a vertex), and we haven't created any new vertices. So the polytope corresponding to the new linear program can only have fewer vertices, and x^* is still a vertex; this means x^* is the optimum the new linear program over all *vertices*, so since the optimum is at some vertex, it has to actually be the optimum.

Another way to think about this argument is that we know none of the constraints that we deleted (when going from the linear program on the left-hand side to the one on the right) is that none of the constraints we deleted are tight for x^*, so when we delete them, we don't change the local picture around x^*. And x^* was a *local* optimum before, so it's still a local optimum. But we've got a convex program, so a local optimum is the same as a global optimum; and this means x^* is also still a global optimum.



Remark 2.8. For Claim 2.7, we can assume that x^* is a vertex, since there's always some vertex achieving the optimum (even if there are other points achieving the optimum too).

On the other hand, how does $K \cap [-1,1]^n$ compare with $K \cap [-1,1]^{S(g)}$? We'll specifically compare their *volumes* using Royen's theorem (Theorem 2.4 — here we only have strips, so we could actually just use Sidak's lemma). Let $\Gamma_1 = K \cap [-1,1]^n$ and $\Gamma_2 = K \cap [-1,1]^{S(g)}$. Then on one hand, we have

$$\gamma(\Gamma_1) \leq \gamma([-1,1]^n) \leq \left(\frac{7}{8}\right)^n$$

(here we're even forgetting about K and just bounding the Gaussian volume of the cube; this is the probability that n independent Gaussians all lie in $[-1,1]$, and this probability for *each* Gaussian is roughly 68%, which is certainly less than $\frac{7}{8}$).

Meanwhile, on the other hand we have

$$\mathrm{Vol}(\Gamma_2) \geq \mathrm{Vol}(K) \cdot \mathrm{Vol}([-1,1])^{\delta n}$$

by Royen's theorem. And $\mathrm{Vol}(K) \geq e^{-\varepsilon n}$, while $\mathrm{Vol}([-1,1])^{\delta n} \geq e^{-\delta n}$ (for example — since we've only got δn Gaussians that need to lie in $[-1,1]$, so we get an exponential with δn in the exponent). This means

$$\mathrm{Vol}(\Gamma_2) \geq e^{-(\varepsilon+\delta)n}.$$

And this means Γ_2 is way bigger than Γ_1 (in terms of Gaussian volume) — $\mathrm{Vol}(\Gamma_1)$ is an actual exponential (witn fixed coefficient of n in the exponent), while $\mathrm{Vol}(\Gamma_2)$ is an exponential with a tiny exponent of n.

And so since Γ_1 is much smaller than Γ_2, we might expect that it should *not* be the case that $\max_{x\in\Gamma_1}\langle g,x\rangle$ and $\max_{x\in\Gamma_2}\langle g,x\rangle$ end up being the same with high probability (over the random choice of g), since Γ_2 is just way bigger than Γ_1.(This would contradict the fact that x^* is the maximizer for both expressions, which means they're equal; to be more precise, we can get a statement like this for each *fixed* $S\subseteq[n]$ of size δn, and then union-bound over all such sets (to cover all possible $S(g)$).

This is pretty fuzzy, but we're now going to formalize it.

§2.4 Gaussian width

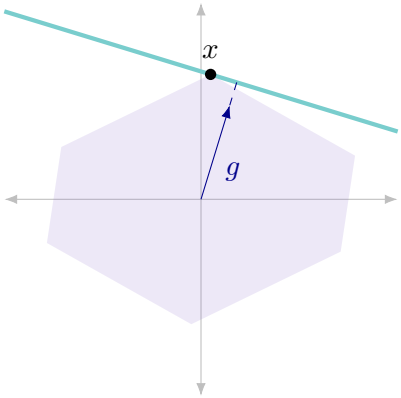
In order to formalize the above argument (that we expect $\max_{x\in\Gamma_1}\langle g,x\rangle$ and $\max_{x\in\Gamma_2}\langle g,x\rangle$ to be different), we'll introduce the concept of *Gaussian width*.

Definition 2.9. For a symmetric convex $K\subseteq\mathbb{R}^n$, its mean width (or Gaussian width) is defined as

$$w(K)=\mathbb{E}_{g\sim\mathcal{N}(0,\mathrm{Id}_n)}\max_{x\in K}\langle g,x\rangle.$$

So we're sampling a random Gaussian, and then essentially looking at the width of K in that direction, which is the reason for the name.

Remark 2.10. This isn't exactly true, because we have to scale based on $\|g\|$. But you could imagine choosing g to be uniform on the sphere instead (in which case this description would be true), in which case the definition would just scale by $\mathbb{E}\|g\|$ — what's important is just that the standard Gaussian is rotationally invariant.



Then what we're actually going to show is that $w(\Gamma_1)$ and $w(\Gamma_2)$ are far apart. For Γ_1, when we described the intuition (that the Gaussian volume of Γ_1 is much smaller than that of Γ_2) we simply bounded Γ_1 by $[-1,1]^n$, and we'll do the same thing here — we have

$$w(\Gamma_1)\le w([-1,1]^n)=\mathbb{E}_g\max_{x\in[-1,1]^n}\langle g,x\rangle.$$

And $\max_{x\in[-1,1]^n}\langle g,x\rangle$ is just $\|g\|_1$ (i.e., the L^1 norm of g), so we get

$$w(\Gamma_1)\le\mathbb{E}_g\|g\|_1=n\cdot\sqrt{\frac{2}{\pi}}$$

(it's a computation that the expected absolute value of a one-dimensional Gaussian is $\sqrt{2/\pi}$).

Now let's try to understand $w(\Gamma_2)$. For Γ_1, we got an upper bound on $\gamma(\Gamma_1)$ (in Subsection 2.3) by simply replacing it with $[-1,1]^n$, so it made sense to do the same thing to upper-bound $w(\Gamma_1)$. Meanwhile, in Subsection 2.3 we got our lower bound on $\gamma(\Gamma_1)$ by using some correlation inequality for Gaussian volume. So here we'd like to have some sort of a correlation inequality for Gaussian width.

We don't exactly have a correlation inequality for Gaussian width. But what we *can* do is first use the correlation inequality to lower-bound the *volume*, and then use the following theorem to go from large volume to large width.

Theorem 2.11 (Urysohn)

Among all symmetric convex bodies in $\mathbb{R}^n$ with a given Gaussian volume V, the minimum Gaussian width is attained by the ball centered at the origin (with the appropriate volume).

(This is true in the Euclidean case — the minimizer of width for a given *Euclidean* volume is also a ball — and it turns out to also be true in the Gaussian case.)

We'll see the proof soon, because it's a nice argument. But first we'll see how to use it to lower-bound $w(\Gamma_2)$. First we saw in Subsection 2.3 that $\mathrm{Vol}(\Gamma_2) = e^{-o(n)}$ (if we think of ε and δ as going to 0). And the origin-centered ball with volume $e^{-o(n)}$ has radius close to $\sqrt{n}$ — this is because the expectation of $\|g\|$ (for a standard Gaussian g) is $\sqrt{n}$, and it concentrates very tightly around its expectation. And the Gaussian width of the radius-$\sqrt{n}$ ball is roughly n (since $\|g\| \approx \sqrt{n}$ and we're trying to maximize $\langle g, x\rangle$ for x on this ball, and the best thing to do is take x in the same direction as g). So we get

$$w(\Gamma_2) \geq (1 - c_\varepsilon)n$$

(where $c_\varepsilon \to 0$ as $\varepsilon \to 0$). And $\sqrt{2/\pi} < 1$, so we get a separation between $w(\Gamma_1)$ and $w(\Gamma_2)$.

Remark 2.12. It looks like a coincidence that $\sqrt{2/\pi} < 1$, but it's not — the cube $[-1,1]^n$ is contained in (and smaller than) the ball of radius $\sqrt{n}$, so since the ball has a Gaussian width of n, the cube should have Gaussian width smaller than this, and the leading constant ends up being $\sqrt{2/\pi}$.

And once we've gotten a separation between $w(\Gamma_1)$ and $w(\Gamma_2)$, to deduce Theorem 2.6, we can show that $\max_x \langle g, x\rangle$ (as a function of the random Gaussian g) concentrates well enough around its expectation to get the result we want. (Specifically, this concentration means that for any $S(g)$ we'll have $\max_{x\in\Gamma_1}\langle g, x\rangle < \max_{x\in\Gamma_2}\langle g, x\rangle$ with very high probability (since their expectations are separated), and then we can union-bound over all possible $S(g)$ of size δn.)

§2.5 Proof of Urysohn's theorem

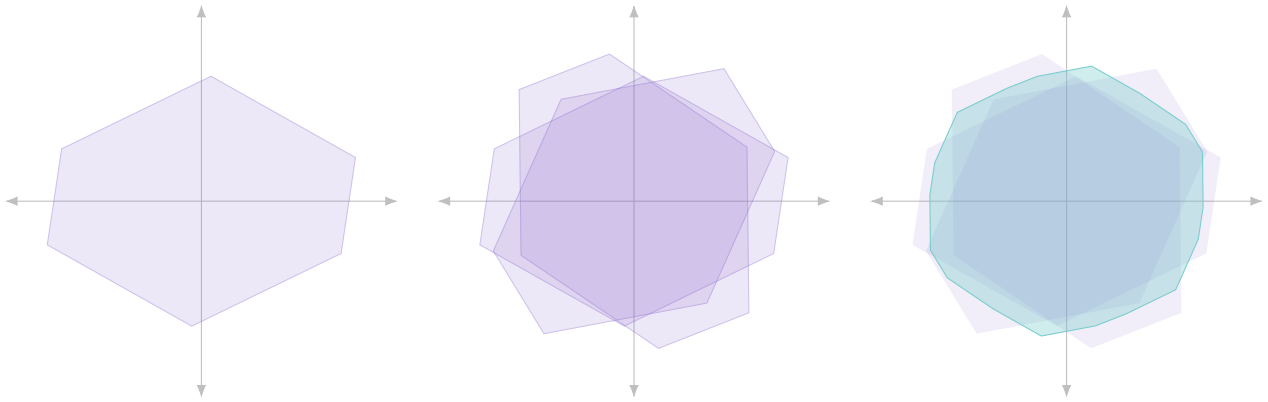
Now we'll prove Urysohn's theorem (Theorem 2.11). It suffices to show that if $K \subseteq \mathbb{R}^n$ is a symmetric convex body and B is an origin-centered ball with $w(K) = w(B)$, then $\gamma(B) \geq \gamma(K)$.

Often when you're trying to prove an iso-something inequality (such as this one) and the extremizer is a ball, you can try to run a symmetrization argument. And that's what we'll do here — we want to somehow symmetrize K to get to a ball.

To symmetrize K, we're going to consider the body

$$K_m = \frac{1}{m}\sum_{i=1}^{m} U_i K$$

where $U_1, \ldots, U_m$ are independent random elements of $\mathrm{SO}(n)$ (i.e., they're random rotation matrices), and the sum denotes the Minkowski sum. So what we're doing is we start with some convex body, and then we randomly rotate it a bunch of times, and we take the average of these rotations (using a Minkowski sum). Intuitively, you'd expect this to make it more and more round.



First, as a sanity check, we need to make sure that this symmetrization doesn't decrease the Gaussian width. But in fact, the width is *invariant*, due to the following fact.

Fact 2.13 — For any symmetric convex bodies K_1 and K_2, we have $w(K_1 + K_2) = w(K_1) + w(K_2)$.

Proof. Points in $K_1 + K_2$ are precisely those of the form $x_1 + x_2$ for $x_1 \in K_1$ and $x_2 \in K_2$, so

$$w(K_1 + K_2) = \max_{x_1 \in K_1, x_2 \in K_2} \langle g, x_1 + x_2 \rangle = \max_{x_1 \in K_1} \langle g, x_1 \rangle + \max_{x_2 \in K_2} \langle g, x_2 \rangle = w(K_1) + w(K_2). \qquad \square$$

This means we always have

$$w(K_m) = w(K) = w(B)$$

for all m (since we assumed $w(K) = w(B)$). So this process preserves width, and the next question is what it does to (Gaussian) volume — we're going to show that $\gamma(K_m) \geq \gamma(K)$ for all m, and moreover that $\gamma(K_m) \to \gamma(B)$ as $m \to \infty$; together, these imply $\gamma(K) \leq \gamma(B)$.

Claim 2.14 — We have $\gamma(K_m) \geq \gamma(K)$ for all m.

Proof sketch. Because Gaussian volume is log-concave in the Minkowski sum, we get

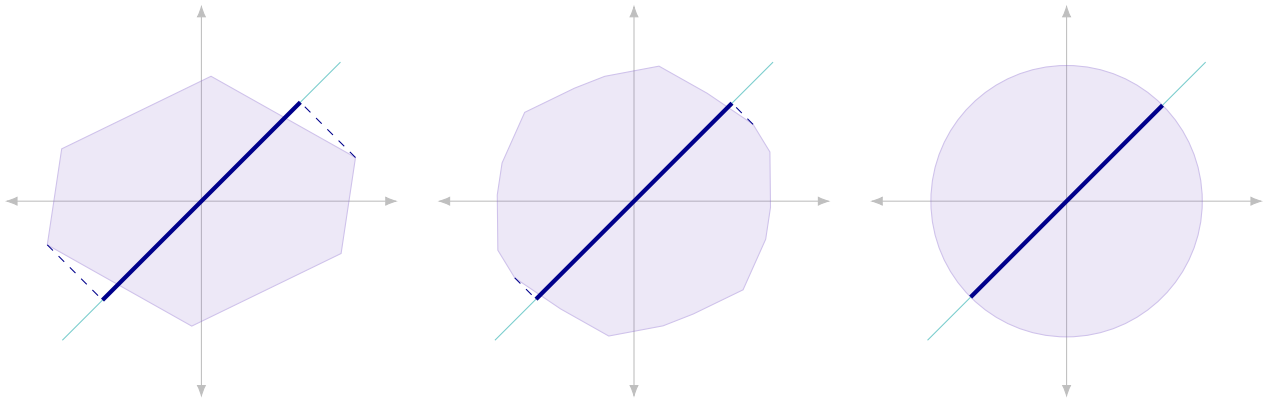
$$\mathrm{Vol}(K_m) = \mathrm{Vol}\left(\frac{1}{m}\sum_{i=1}^{m} U_i K\right) \geq \prod_{i=1}^{m} \mathrm{Vol}(U_i K)^{1/m} = \mathrm{Vol}(K).$$

(The last inequality is because Gaussians are rotationally invariant, so the Gaussian volume of a rotation of K is the same as that of K itself.) $\square$

So we've now seen that our symmetrization process preserves width and doesn't decrease volume. And intuitively, it's making our shape more and more like a ball (in a sense that means $\gamma(K_m) \to \gamma(B)$); now it remains to prove this.

We'd like to say that $K_m \to B$ in some appropriate topology (particularly, a topology for which $\gamma(\bullet)$ is continuous) — we'd intuitively expect this to be true (since symmetrization is supposed to make K look more and more like a ball), and it'd be enough to imply $\gamma(K_m) \to \gamma(B)$.

The idea behind choosing the topology is that we know K_m and B are both convex, so it suffices to check pointwise convergence in each direction — this means we imagine projecting the two shapes onto each direction θ and showing that the support of K_m in the direction θ converges to that of B. (The reason this works is that Gaussian volume is continuous with respect to these support functions.)



So more precisely, we want to prove the following statement.

> **Claim 2.15 —** For each direction $\theta \in \mathbb{S}^{n-1}$, we have $\max_{x \in K_m} \langle x, \theta \rangle \to \max_{x \in B} \langle x, \theta \rangle$.

(Note that the right-hand side is independent of θ, but the left-hand side could depend on θ.)

Proof. We can rewrite the left-hand side as

$$\max_{x \in K_m} \langle \theta, x \rangle = \frac{1}{m} \max_{x_1, \dots, x_m \in K} \langle \theta, U_1 x_1 + \cdots + U_m x_m \rangle = \frac{1}{m} \sum_{i=1}^{m} \max_{x_i \in K} \langle \theta, U_i x_i \rangle,$$

and we can move U_i to the other side to rewrite this as

$$\max_{x \in K_m} \langle \theta, x \rangle = \frac{1}{m} \sum_{i=1}^{m} \max_{x_i \in K} \langle U_i^\intercal \theta, x_i \rangle.$$

And the distribution of $U_i^\intercal \theta$ is uniform on the sphere (because θ is some fixed vector on the sphere, and we're hitting it with a random rotation matrix). So we can rewrite this as

$$\frac{1}{m} \sum_{i=1}^{m} \max_{x_i \in K} \langle \theta_i, x_i \rangle$$

(where $\theta_1, \dots, \theta_m$ are independent and uniform on the sphere). And now we can use the law of large numbers — this is an average of m i.i.d. terms $\max_{x \in K} \langle \theta_i, x \rangle$ (for $\theta \sim \mathbb{S}^{n-1}$), so by the law of large numbers it converges to the expectation of each of these terms, i.e.,

$$\mathbb{E}_{\theta \sim \mathbb{S}^{n-1}} \max_{x \in K} \langle \theta, x \rangle.$$

But this is precisely the mean width of K (except for a factor of $\mathbb{E}\,\|g\|$), which is the same as the mean width of the ball B (by definition); so we can rewrite this as

$$\mathbb{E}_{\theta \in \mathbb{S}^{n-1}} \max_{x \in B} \langle \theta, x \rangle = \max_{x \in B} \langle \theta, x \rangle$$

(we can remove the expectation over θ because this quantity doesn't depend on θ). □

§3 The matrix Spencer conjecture

There's a conjecture which is a generalization of Spencer's theorem to a noncommutative setting.

Conjecture 3.1 — Suppose $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ are symmetric matrices with $\|A_i\|_{\text{op}} \le 1$ for all i. Then there exists $x \in \{\pm 1\}^n$ such that

$$\left\|\sum_{i=1}^n x_i A_i\right\|_{\text{op}} \lesssim \sqrt{n \log\left(\frac{d}{n} + 2\right)}.$$

(By $\|A\|_{\text{op}}$ we're referring to the operator norm, where $\mathbb{R}^d$ has the L^2 norm.)

Why is this a generalization of Spencer's theorem? To prove Spencer's theorem (Theorem 1.1) from this, suppose we've got sets $S_1, \dots, S_d \subseteq [n]$. Then for each element i, we can associate to it a diagonal matrix $A_i \in \mathbb{R}^{d\times d}$ whose jth diagonal entry is 1 if $i \in S_j$ and 0 otherwise. Then it's clear that $\|A_i\|_{\text{op}} \le 1$ for all i (since these matrices are all diagonal matrices with entries 1 and 0). And $\sum_{i=1}^n x_i A_i$ will also be a diagonal matrix, whose diagonals are the discrepancies; so getting a bound on the operator norm of this matrix gets the same bound on the discrepancies.

Just as with Spencer's theorem, a random signing $x \in \{\pm 1\}^n$ gets a bound of $\sqrt{n \log d}$, so we'd like to go from $\log d$ to $\log d/n$. And just as with Spencer's theorem, in order to prove this, it'd be enough to prove a partial coloring version (because then iterating the partial coloring version would again give a geometric series).

What would happen if we tried to prove a partial coloring result using volume lower bounds? We'll again focus on the case $d = n$; then the body we want to define is

$$K_C = \left\{ x \in \mathbb{R}^n \;\middle|\; \left\|\sum_{i=1}^n A_i x_i\right\|_{\text{op}} \le C\sqrt{n} \right\}.$$

We'd *like* to show that

$$\gamma(K_C) \ge e^{-\varepsilon_C n} \quad \text{where } \varepsilon_C \to 0 \text{ as } C \to \infty. \tag{1}$$

If we could do this, then we could apply the geometric partial coloring lemma (Theorem 2.2) to get some $x \in K_C \cap [-1, 1]^n$ with a decent fraction of ± 1-coordinates, which would give us a partial coloring result (that we could iterate to prove the conjecture).

But we can't prove this in the same way as before — when we proved the analogous volume lower bound for Spencer's theorem, K_C was an intersection of n strips, which let us use Sidak's lemma (more generally, if K_C were an intersection of n large convex sets, then you could use the Gaussian correlation inequality). But here we've got too many things being intersected, so we can't just use Gaussian correlation.

Remark 3.2. Do we expect this statement (i.e., that $\gamma(K_C)$ is large) to be true? It's not clear — it's not clear even whether we should expect the matrix Spencer conjecture to be true, and this is stronger. (It would be very interesting if matrix Spencer is true and this statement is false.)

§3.1 The matrix Khintchine inequality

If we're trying to prove (1), the first thing to ask is what we can say about $\mathbb{E}\,\|\sum_{i=1}^n g_i A_i\|_{\text{op}}$.

Theorem 3.3 (Matrix Khintchine inequality)
We have $\mathbb{E}\,\|\sum_{i=1}^n g_i A_i\|_{\text{op}} \le \sqrt{\log d} \cdot \|\sum_{i=1}^n A_i^2\|_{\text{op}}^{1/2}$.

The way to think about this is that if we let $X = \sum g_i A_i$ (so X is some matrix with jointly Gaussian entries), then the term on the right-hand side involving $\sum A_i^2$ is some measure of the 'standard deviation' of X, and we have to pay an extra dimension-dependent factor of $\sqrt{\log d}$.

Remark 3.4. Given a matrix X with jointly Gaussian entries, we may be able to write X as $\sum g_i A_i$ in multiple different ways, and it might look like Theorem 3.3 would give different bounds for those different representations. But it actually doesn't — we have $\sum A_i^2 = \mathbb{E}X^2$, so the bound on the right-hand side doesn't actually depend on how we choose the A_i's in our representation of X.

Combined with Markov's inequality, Theorem 3.3 means that we can prove a statement like (1) if we replace $C\sqrt{n}$ with $2\sqrt{\log d}\,\|\sum A_i^2\|_{\mathrm{op}}^{1/2}$ — i.e., if we define $K = \{x \in \mathbb{R}^n \mid \|\sum A_i x_i\|_{\mathrm{op}} \le 2\sqrt{\log d}\,\|\sum A_i^2\|_{\mathrm{op}}^{1/2}\}$, then by Markov's inequality and Theorem 3.3 we have $\gamma(K) \ge \frac{1}{2}$ (since for a standard Gaussian g, for g to *not* be in K, this quantity would have to be at least twice its expectation), which is certainly $e^{-o(n)}$.

How good of a bound does this get? We have $\|\sum A_i^2\|_{\mathrm{op}} \le \sum \|A_i^2\|_{\mathrm{op}} \le \sum \|A_i\|_{\mathrm{op}}^2 \le n$ (the last inequality is because we assumed $\|A_i\|_{\mathrm{op}} \le 1$ for all i), so we get that the modified version of (1) — and therefore matrix Spencer — is true if we want a bound of $2\sqrt{n \log d}$ (in place of the right-hand side of Conjecture 3.1). But a random signing would give the same bound, so we'd really like to do better than this.

§3.2 A proof for certain matrices

We'll first state the result, and then say a bit about how it's proved.

Definition 3.5. We define the Frobenius norm of a matrix A, which we denote by $\|A\|_{\mathrm{F}}$, as the sum of squares of its eigenvalues.

Theorem 3.6 (Bansal–Jiang–Meka 2022)
The matrix Spencer conjecture (in the $d = n$ case) is true if $\|A_i\|_{\mathrm{F}}^2 \lesssim n/(\log n)^{10}$ for all i.

If $\|A\|_{\mathrm{op}} \le 1$, then each eigenvalue of A is at most 1, so we have $\|A\|_{\mathrm{F}} \le n$. And so Theorem 3.6 says that if we're bounded away from this by a polylog, then matrix Spencer is true.

This is an extremely nice result — for a while people were trying even to prove matrix Spencer for rank 1 or rank 4 matrices, and after a lot of work they got up to $n^{1/4}$ or $n^{1/2}$. So this result is a big jump (it in particular gets matrix Spencer for rank up to $n/\mathrm{polylog}(n)$).

Remark 3.7. Even for rank-1 matrices, matrix Spencer is nontrivial, and we can't get a bound better than $\Omega(\sqrt{n})$ in the $d \approx n$ case (which is what matrix Spencer gives). As a construction, take A_i to be the matrix with 1's in the coordinates $(1,1)$, $(1,i)$, $(i,1)$, and (i,i), and 0's elsewhere.

Remark 3.8. It's not clear that this should be considered good evidence for the full version of matrix Spencer being true. The interesting thing about Spencer's theorem and the partial coloring result in Theorem 1.2 is that they're about large or moderate deviations, rather than expectations — in particular, the region of partial colorings with discrepancy $C\sqrt{n}$ had subexponential Gaussian volume. On the other hand, Theorem 3.6 can be proven by sticking only to expectations — in particular, there's a way to do the proof such that the region we get has Gaussian volume $\frac{1}{2}$, which means this proof isn't really capturing the same thing as the proof of Spencer's theorem. In particular, we don't yet have an argument for a special cases of matrix Spencer which would also work to prove Spencer's theorem.

§3.2.1 A refinement of matrix Khintchine

The proof of Theorem 3.6 uses a recent refinement of the matrix Khintchine inequality. The matrix Khintchine inequality (Theorem 3.3) bounds the expected operator norm of $X = \sum g_i A_i$ by the square root of its variance times a dimension-dependent factor of $\sqrt{\log d}$. We can't hope to improve this in general, as seen by the following example.

Example 3.9

Suppose that each A_i is a diagonal matrix with a 1 on the ith entry and a 0 everywhere else. Then X has diagonal $(g_1, \dots, g_d)$, so $\|X\|_{\text{op}} = \max\{|g_1|, \dots, |g_d|\} = \Theta(\sqrt{\log d})$.

So we can't improve the bound that matrix Khintchine gives in general. But we can ask if we can do better in cases that don't look like this one in some sense, and the answer is yes!

Theorem 3.10 (Bandeira–Boedihardjo–van Handel 2021)

In the same setup as matrix Khintchine, letting $X = \sum_{i=1}^n g_i A_i$, we have

$$\mathbb{E}\,\|X\|_{\text{op}} \lesssim \mathbb{E}[X^2]^{1/2} + (\log d)^{3/2}\,\|\text{Cov}(X)\|_{\text{op}}^{1/2},$$

where $\text{Cov}(X)$ is the $d^2 \times d^2$ covariance matrix of X.

We can view X as a d^2-dimensional Gaussian vector (since its entries are jointly Gaussian), so it'll have some $d^2 \times d^2$ covariance matrix, which we call $\text{Cov}(X)$. Note that everything scales linearly with X, so these exponents make sense.

In comparison to Theorem 3.3, here we don't have the extra factor of $\sqrt{\log d}$ attached to $\mathbb{E}[X^2]^{1/2}$. Instead, we have an extra correction term depending on $\text{Cov}(X)$.

Remark 3.11. The intuition is that in Example 3.9, we're essentially embedding a commutative example in the non-commutative setting, and here we have to pick up the $\sqrt{\log d}$. But in a truly non-commutative example, you would expect the covariance to be much smaller than the variance, so the additional factors of $\log d$ shouldn't matter much.

This isn't exactly true though — there do exist non-commutative matrices with large covariance. You can imagine a spectrum between commutative matrices (which have *all* eigenvectors in common) and matrices with *no* eigenvectors in common. If we take matrices with a *smaller number* of eigenvalues in common instead, then they'll still be noncommutative (in some sense), but $\text{Cov}(X)$ will be large.

Then to prove Theorem 3.6, the authors show that the condition on Frobenius norms lets you arrange for $\text{Cov}(X)$ to be small, and then you can use Theorem 3.10 to get a volume bound.