Lognormal limit for the symmetric perceptron

Talk by Ashwin Sah

Notes by Sanjana Das

April 14, 2023

§1 Introduction

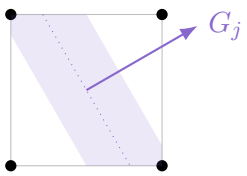
This is based on the paper *Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron* by Emmanuel Abbe, Shuangping Li, and Allan Sly. The problem is motivated by statistical physics, but we'll look at it combinatorially; the proof uses the second moment method, but modified in a way that has interesting applications.

§1.1 The symmetric perceptron

Fix some $\kappa > 0$. For each $j \geq 1$, we sample a Gaussian row $G_j \sim \mathcal{N}(0,1)^{\otimes n}$ (i.e., a vector of n independent Gaussians), and consider the set

$$S_j = \{x \in \{\pm 1\}^n \mid |\langle G_j, x\rangle| \leq \kappa\sqrt{n}\}.$$

In words, S_j is the part of the unit cube whose dot product with our chosen Gaussian row is small. This (i.e., defining 'small' by comparison to $\sqrt{n}$) is the right parameter regime because G_j is Gaussian and the entries of x are ± 1, so $\langle G_j, x\rangle \sim \sqrt{n} \cdot \mathcal{N}(0,1)$.



We then define $S^m(G) = \bigcap_{j=1}^m S_j(G)$; this is the symmetric perceptron.

> **Remark 1.1.** The fact that the entries of G_j are Gaussian is not very important — you could also use $\textsc{Unif}(\{\pm 1\})$. But Gaussians are nice because they're continuous.

§1.2 Capacity

> **Definition 1.2.** The capacity of the symmetric perceptron is
>
> $$m^*_\kappa(n) = \max\{m \geq 1 \mid S^m(G) \neq \emptyset\}.$$

In words, the capacity is the last time at which there's a point on the hypercube that survives. Every step is likely going to kill some point, so with high probability the capacity is finite; we won't worry about the case where it's infinite (meaning the process goes on forever).

Conjecture 1.3 — For some explicit constant α_κ^*, we have

$$\frac{m_\kappa^*(n)}{n} \to \alpha_\kappa^*$$

in distribution as $n \to \infty$.

This makes intuitive sense — if $\kappa > 0$, then at every step, we'd expect a positive fraction of points to get killed. We have 2^n points to start with, so if we do this $\Theta(n)$ times, then all points should die.

First, we can get a nice combinatorial bound using this; in some asymmetric physics problems this naive bound doesn't give the correct prediction, but it turns out that in this case it does. Let's fix some vector $x \in \{\pm 1\}^n$ and consider the probability that some Gaussian row G_j keeps x, meaning that $|\langle G_j, x\rangle| \le \kappa\sqrt{n}$. When we compute $\langle G_j, x\rangle$, we're just adding a bunch of standard Gaussians (with signs given by x), so

$$\langle G_j, x\rangle \sim \sqrt{n} \cdot \mathcal{N}(0,1).$$

Definition 1.4. We define $p_\kappa = \mathbb{P}_{Z\sim\mathcal{N}(0,1)}[|Z| \le \kappa]$.

Then we have $\mathbb{P}[|\langle G_j, x\rangle| \le \kappa\sqrt{n}] = p_\kappa$ for each j. And if we consider the perceptron after m steps, we need this to occur for m independent rows G_j, so we have

$$\mathbb{P}[x \in S^m(G)] = p_\kappa^m.$$

By linearity of expectation, this means

$$\mathbb{E}[|S^m(G)|] = 2^m p_\kappa^m.$$

So if we define

$$\alpha_c(\kappa) = \frac{\log 2}{\log(1/p_\kappa)},$$

then we have $\mathbb{E}[|S^m(G)|] = o(1)$ for all $m \ge (\alpha_c(\kappa) + \varepsilon)n$, which by Markov means that $\mathbb{P}[\text{capacity} \ge m] = o(1)$. So this shows that $\alpha_\kappa^* \le \alpha_c(\kappa)$ (if the limit exists).

For any problem like this, you can take an expectation and it'll give you *some* information (e.g., an expectation threshold); the hard part is the reverse direction. And this paper shows that the reverse direction holds as well (so Conjecture 1.3 is true with $\alpha_\kappa^* = \alpha_c(\kappa)$).

For this, we fix $0 < \alpha < \alpha_c(\kappa)$, and let $m = \alpha n$ and $Z(G) = |S^m(G)|$. Our goal is to show that for any such α, we have $Z(G) > 0$ with high probability. (This would mean $S^m(G)$ is nonempty, so the capacity is at least m.)

Remark 1.5. The capacity m_κ^* is a random variable defined based on the entire matrix G, But here we're only looking at a specific time m slightly before the critical one, and seeing whether it's good enough; so we don't really need to deal with the precise definition of capacity.

§2 The second moment method

Let $\mu = \mathbb{E}[Z(G)]$; as calculated earlier, we have $\mu = 2^n p_\kappa^{\alpha n} = \omega(1)$ (in fact, μ is exponentially large). So in *expectation*, we have lots of solutions (i.e., elements of $S^m(G)$, which we think of as solutions to our system of inequalities). But this doesn't imply what we want, namely that there exists a solution with high probability — for example, there could be 0 solutions almost all of the time but 2^n with probability $2^{-n/2}$, and we'd still have exponentially many solutions in expectation.

One standard way to prove such a statement is with the second moment method. We'd *like* to show that

$$\mathbb{E}[Z(G)^2] = (1 + o(1))\mu^2$$

(for comparison, we always have $\mathbb{E}[|Z(G)|^2] \geq \mu^2$ by Cauchy–Schwarz). If we could show this, we'd have

$$\mathbb{E}[(Z(G) - \mu)^2] = o(\mu^2),$$

and then by Markov we would have $Z(G) = \mu + o(\mu)$ with high probability.

§2.1 A simpler example

Before we try to do this, we'll see a simpler second moment argument in order to highlight a perspective that'll help us understand heuristically what's happening.

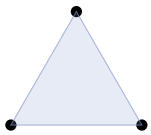
Example 2.1

Let H be a random graph $H \sim \mathcal{G}(n, p)$ with $p = \omega(1/n)$, and let $T(H)$ be the number of triangles in H. Then $T(H) > 0$ with high probability.

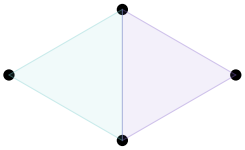
We'll do this moment computation in detail (we won't for the others).

Proof. First, we have $\mu = \mathbb{E}[T(H)] = p^3\binom{n}{3}$. Meanwhile, $\mathbb{E}[T(H)^2]$ counts *pairs* of triangles — so we consider pairs of triangles in K_n, and for each pair, we consider the probability it's in our random graph. There are only so many ways that two triangles can exist, so we can classify all the cases and compute the expectation contributed by each case.

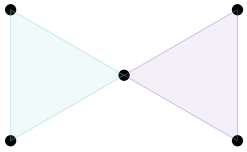
- If the two triangles overlap completely, then we get a contribution of $p^3\binom{n}{3}$.

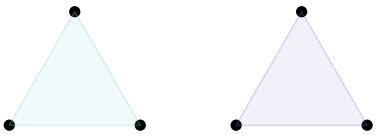


- If the triangles partially overlap (so they share one edge), then we have 5 edges and 4 vertices; so we get a contribution of $p^5 \cdot 12\binom{n}{4}$.



- If the triangles share one vertex, then we get $p^6 \cdot 30\binom{n}{5}$; if they don't share any, we get $p^6 \cdot 20\binom{n}{6}$.





When we add these together, we get

$$\mathbb{E}[T(H)^2] = (1 + o(1))\left(\frac{p^3n^3}{6}\right)^2 = (1 + o(1))\mu^2,$$

which is what we wanted. Here the final case is the main term and contributes roughly μ^2, and all the remaining terms are small.

(Sometimes this is phrased in terms of covariance; if we compute $\mathrm{Var}[T(H)]$ using bilinearity of covariance, then the p^6 terms — where the triangles aren't correlated at all — will cancel out, and the remaining terms will be small.) □

With this in mind, we'll now see a slightly different approach to this computation for $\mathbb{E}[T(H)^2]$, using a *planted model*. We're again trying to count pairs of triangles; but all triangles are symmetric, so let's fix a *specific* triangle $\mathcal{T}$. There are $\binom{n}{3}$ choices for this specific triangle, and each appears with probability p^3.

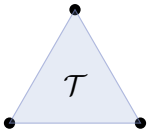
Then to compute $\mathbb{E}[T(H)^2]$, we want the expected number of further triangles which are H, given that this specific triangle $\mathcal{T}$ is in H — we have

$$\mathbb{E}[T(H)^2] = p^3\binom{n}{3}\sum_{\mathcal{T}'}\mathbb{P}[\mathcal{T}' \in H \mid \mathcal{T} \in H] = \mu \cdot \mathbb{E}[T(H) \mid \mathcal{T} \subseteq H].$$

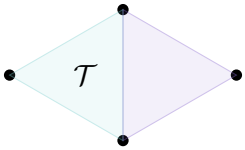
And our goal is to show that $\mathbb{E}[T(H) \mid \mathcal{T} \subseteq H] = (1 + o(1))\mu$.

In this new expectation, we've essentially planted the triangle $\mathcal{T}$. Then $T(H)$ will follow a different distribution than the original one — for one thing, it'll always be counting $\mathcal{T}$. But what we're claiming is that above the threshold (i.e., for $p = \omega(1/n)$), the mean of this distribution is basically the same as that of the original. (In fact, here we can even say that the *distributions* are basically the same.)

Here, this is pretty easy to see. In the new distribution, when we're counting triangles, we have a 1 for free coming from our planted triangle $\mathcal{T}$.



We could also have a triangle which shares one edge with $\mathcal{T}$; then there are roughly ways to choose the new vertex, and such a triangle appears with probability p^2 (corresponding to the two new edges — we're thinking about what is further needed for the new triangle, given that we already have $\mathcal{T}$).



Finally, we have all the triangles which are independent of $\mathcal{T}$; there are $\binom{n}{3} - O(n^2)$ of them, and each appears with probability p^3.

From here it's easy to see that all the overlapping correlated terms are small, so we do get

$$\mathbb{E}[T(H) \mid \mathcal{T} \subseteq H] = (1 + o(1))\mu.$$

§3 A first attempt at second moments

Now let's try calculating the second moment of $Z(G)$. This isn't going to work — instead of getting that $\mathbb{E}[Z(G)^2]$ is $(1 + o(1))\mu^2$, we'll get that it's a constant multiple of μ^2. (So in this case, planting one solution is enough to distort the expectation by a little bit; we'll maybe see why.) Then we'll see how to fix it (this

technique only works when the second moment is a constant factor off; if it's exponentially off then you need to bring in other things).

We refer to elements of $S^m(G)$ (which $Z(G)$ is counting) as *solutions* (since they're points $x \in \{\pm 1\}^n$ which satisfy our inequality system). Earlier, we saw that

$$\mu = \mathbb{E}[Z(G)] = 2^n p_\kappa^{\alpha n},$$

since there are 2^n possible solutions and each occurs with probability p_κ.

Using the planted perspective on second moments, all values of x are symmetric, so we can just imagine planting $v = (1, 1, \dots, 1)$; then we get

$$\mathbb{E}[Z(G)^2] = \mu \cdot \mathbb{E}[Z(G) \mid v \in S^m(G)].$$

The hope would be that planting v doesn't affect things at all, but unfortunately this turns out to be false.

Let's quantify what happens when we condition on v being a solution. If we write out our matrix, we have m rows $G_i = (G_{i1}, \dots, G_{in})$; and this condition tells us that

$$|G_{i1} + \cdots + G_{in}| \le \kappa\sqrt{n}$$

for each i. And the distribution of $G_i = (G_{i1}, \dots, G_{in})$ once we reveal the sum $G_{i1} + \cdots + G_{in}$ is still nice — it's a tuple of Gaussians correlated in some way.

Now when we're considering the probability that each $x \in \{\pm 1\}^n$ is a solution, they're no longer all symmetric. But we can consider splitting them up by how many 1's and -1's there are — all vectors with the same number of 1's and -1's are symmetric. So if we let v_t be a vector with $\frac{n+t}{2}$ and $\frac{n-t}{2}$ 1's and -1's, respectively (for each $t \equiv n \pmod 2$), then we get

$$\mathbb{E}[Z(G) \mid v \in S^m(G)] = \sum_t \binom{t}{(n+t)/2} \mathbb{P}[v_t \in S^m(G) \mid v \in S^m(G)].$$

And the rows G_i are still independent, so we can write this as

$$\mathbb{E}[Z(G) \mid v \in S^m(G)] = \sum_t \binom{t}{(n+t)/2} \mathbb{P}[v_t \in S^1(G) \mid v \in S^1(G)]^m. \tag{3.1}$$

Now we need to find this probability. As seen earlier, the thing we're conditioning on is that

$$|G_{11} + \cdots + G_{1n}| \le \kappa\sqrt{n},$$

which we can equivalently think of as

$$\left|\left\langle \frac{v}{\sqrt{n}}, G_1 \right\rangle\right| \le \kappa.$$

(This dot product is some unit-variance Gaussian.) And we want to find the conditional probability that

$$\left|\left\langle \frac{v_t}{\sqrt{n}}, G_1 \right\rangle\right| \le \kappa,$$

i.e., that some other unit-variance Gaussian is also at most κ (in magnitude).

If $t = 0$, then we have $\langle v, v_0 \rangle = 0$, which means that these two projected Gaussians

$$\left\langle \frac{v}{\sqrt{n}}, G_1 \right\rangle \quad \text{and} \quad \left\langle \frac{v_0}{\sqrt{n}}, G_1 \right\rangle$$

are actually independent. So the probability remains the same as the unconditional one (i.e., p_κ).

When t is large (e.g., a constant fraction of n), $\langle v, v_t \rangle$ will be very large, and the probability will change by a constant factor. But this doesn't affect (3.1) much, because the binomial coefficients in this case are very small.

The issue is when t is roughly $\sqrt{n}$ (where the binomial coefficients are still reasonable). Here, v and v_t are still *approximately* orthogonal, but they're a bit off, and this is enough to cause a nonnegligible contribution. If $t = b\sqrt{n}$, then we'll have

$$\mathbb{P}[v_t \in S^1(G) \mid v \in S^1(G)] \approx p_\kappa \cdot \left(1 + \frac{f(b)}{n}\right),$$

where $f(b)$ is some function only depending on b. To see why, this is really a two-dimensional problem — $\langle v, G_1 \rangle$ and $\langle v_t, G_1 \rangle$ are jointly Gaussian, so their relationship only depends on their covariance, which is a function of the angle between v and v_t. If v and v_t were orthogonal, then these two Gaussians would be independent, and we'd have a probability of p_κ. Now we're shifting the angle by $\frac{1}{\sqrt{n}}$; and if you think about areas, you can convince yourself that the probability changes quadratically based on this angle (so we get a deviation on the scale of $\frac{1}{n}$; and $f(b)$ should be quadratic in b).

So now our sum (3.1) basically becomes

$$\mathbb{E}[Z(G) \mid v \in S^m(G)] \approx \sum_t \binom{n}{(n+t)/2} p_\kappa^m \left(1 + \frac{f(b)}{n}\right)^{\alpha n},$$

which can be approximated as

$$\sum_t \frac{2^n e^{-b^2/2}}{\sqrt{2\pi n}} \cdot p_\kappa^m e^{\alpha f(b)}.$$

So we get back the original heuristic for $\mu = \mathbb{E}[Z(G)]$ (which was $2^n \cdot p_\kappa^m$), but it's modified by something — there's an extra factor of roughly

$$\int_b \frac{e^{-b^2/2}}{\sqrt{2\pi}} \cdot e^{\alpha f(b)}\, db.$$

It turns out that this integral always converges, so it'll come out to some constant C (depending on κ and α), and we'll get

$$\mathbb{E}[Z(G)^2] \approx C\mu^2.$$

This in particular tells us that the distribution of $Z(G)$ is *not* concentrated at μ — in fact, we're going to see that when we divide by μ, the distribution will be lognormal.

§4 Small graph conditioning

The method we'll use is called *small graph conditioning*. One reference paper is Robinson–Wormald (1994), on random regular graphs having Hamilton cycles — they were in a similar situation where if you try to do a second moment computation, you get a constant C that you need to get rid of (i.e., you need to explain away the extra variance). A follow-up paper by Janson (1995) then introduced a more general framework for thinking about these things.

The phenomenon they discovered is that $Z(G)/\mu$ is going to converge to some distribution whose square mean is C instead of 1; and in fact, not only can we say what this distribution is, but we can come up with a simple statistic whose distribution it is. In this case, that statistic will be $\exp(Y)$ for some random variable Y, and we're going to see that

$$\frac{Z(G)}{\mu \exp(Y)} \to 1$$

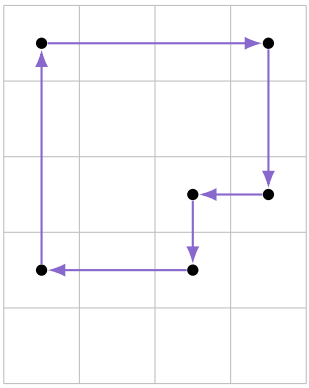
in distribution, which will show that $Z(G) = (1 \pm o(1)) \cdot \mu \exp(Y)$ with high probability. Then in order to understand the distribution of $Z(G)$, it's enough to understand the distribution of Y; and Y will turn out to be Gaussian, so $Z(G)$ will have a lognormal distribution.

> **Remark 4.1.** A lognormal distribution is a distribution whose logarithm is normal, not the logarithm of a normal distribution.

And the idea is that you'd expect Y to be related to *small graph statistics*. In our setting, we want to look at certain cross-correlations in our matrix G, and this amounts to looking at cycles inside the matrix — so for each k, we define

$$C_k^* = \sum_{i_1,\dots,i_k} \sum_{j_1,\dots,j_k} \prod_{\ell=1}^{k} G_{j_\ell i_\ell} G_{j_\ell i_{\ell+1}}.$$

For example, when $k = 1$ we have $C_1^* = \sum_{i,j} G_{ij}^2 = \sum_j |G_j|^2$.



Then we define C_k as a normalized version of C_k^* — specifically, we define

$$C_k = \frac{C_k^* - \mathbb{E}[C_k^*]}{(mn)^{k/2}}$$

(since the standard deviation is on the order of $(mn)^{k/2}$; the term $\mathbb{E}[C_k^*]$ only matters when $k = 1$, because otherwise C_k^* has mean 0). These variables C_k are expected to be of constant order. In fact, each C_k is a polynomial of independent Gaussians, so there is some machinery to deal with them — it can be shown that in the limit, their distribution is that of independent Gaussians.

Remark 4.2. What motivates the definition of C_k^*? One way to get to this is that if you try computing things more explicitly in the first attempt, then you get some partition function, expand it as a series, and look at what terms you need to care about; and you start getting things like this.

Now we consider some real parameters γ_0, γ_1, γ_2, ..., and we let

$$Y = \gamma_0 + \sum_{k \geq 1} \gamma_k C_k.$$

(In practice, you don't want to use the actual infinite sum, so you cap it at some slowly growing function such as $\log \log \log n$.) The best way to deal with these parameters is to keep them as variables for now, and set them later.

Remark 4.3. It maybe seems weird that you'll be able to force the values of γ_k to be unique. But to see why you can expect this, you can think of these random variables Y and C_k as vectors, and covariances as dot products. Then these coefficients γ_k should correspond to what happens if we project Y onto each C_k (since the C_k are independent in the limit). And we want Y to be a specific thing — we want it to be roughly $\log(Z(G)/\mu)$ — so this should give you unique values for γ_k.

Now instead of doing second moments on $Z(G)$ directly, we'll do second moments on the random variable $Z(G)e^{-Y}$ (which we're trying to show is concentrated at μ — i.e., that dividing out by e^Y explains away our extra variance).

The goal is to convince us that there's a reasonable way to do second moments on this random variable; but there are a lot of computational details we won't go into.

§4.1 An overview

First, let's try to compute the *first* moment of $Z(G)e^{-Y}$, since even that isn't obvious. It's still true that all the solutions $x \in \{\pm 1\}^n$ are symmetric, so it's enough to look at $v = (1, 1, \ldots, 1)$, and we have

$$\mathbb{E}[Z(G)e^{-Y}] = 2^n \cdot \mathbb{E}[\mathbf{1}_{v \in S^m(G)} e^{-Y}].$$

Motivated by the planted model, we can write this as

$$\mathbb{E}[Z(G)e^{-Y}] = 2^n p_\kappa^m \cdot \mathbb{E}[e^{-Y} \mid v \in S^m(G)]. \tag{4.1}$$

First, if we didn't plant anything at all and just wanted to find $\mathbb{E}[e^{-Y}]$, we could do so — the constituents of Y are basically independent Gaussians (in the limit), so we'd just be working in an independent Gaussian model. In order to compute (3.1) (where we *have* planted something), it's enough to prove a distributional statement about Y — given that $v \in S^m(G)$, what's the distribution of Y?

Similarly, for the *second* moment, we'll have

$$\mathbb{E}[Z(G)^2 e^{-2Y}] = 2^n \sum_t \binom{n}{(n+t)/2} \mathbb{E}[\mathbf{1}_{v \in S^m(G)} \mathbf{1}_{v_t \in S^m(G)} e^{-2Y}],$$

and if we define $p_\kappa(t) = \mathbb{P}[v_t \in S^1(G) \mid v \in S^1(G)]$ and

$$\widetilde{Y}(t) = \mathbb{E}[e^{-2Y} \mid v, v^t \in S^m(G)],$$

then we can rewrite this as

$$\mathbb{E}[Z(G)^2 e^{-2Y}] = 2^n p_\kappa^m \sum_t \binom{n}{(n+t)/2} p_\kappa(t)^m \cdot \widetilde{Y}(t).$$

So the point is that for the second moment, we can sort of do the same thing, but we pull out two conditionals (instead of one). Then we get the first term of $2^n p_\kappa^m$ (as before), and a slightly modified probability $p_\kappa(t)^m$ and a correction term $\widetilde{Y}(t)$. The goal is essentially to make these two things kill each other, so that the second moment ends up being within $1 + o(1)$ of the first moment squared.

Computing $\widetilde{Y}(t)$ is again a distributional statement about Y, except that now we're conditioning on two planted things instead of one (and these two planted things might be correlated). If we didn't plant anything, then we'd still have an independent Gaussian model, and we could compute $\mathbb{E}[e^{-2Y}]$ using standard machinery. And we've reduced our problem to just taking two plants and then trying to compute the same statistic; so at this point, it seems quite plausible that you can do this.

Now we'll see some bonus calculations regarding what Y looks like when we plant things.

§4.2 Cycle counts in the planted models

First, we'll write down what Y looks like in the non-planted model. Let

$$V = \left(\frac{C_1}{\sqrt{2}}, \frac{C_2}{\sqrt{4}}, \dots, \frac{C_\ell}{\sqrt{2\ell}} \right).$$

Proposition 4.4

In the non-planted model (where the rows G_j are independent Gaussians), we have $V \to \mathcal{N}(0,1)^{\otimes \ell}$.

So in the non-planted model, the cycle counts C_k behave as independent Gaussians, as mentioned earlier; what's actually interesting is the planted models.

Now we'll look at the 1-planted model, where we've planted one solution v. To get this distribution, we sample the rows G_j independently and then take our one vector $v = (1, 1, \dots, 1)$ and enforce that all the dot products with v are small, i.e., $|\langle G_j, v \rangle| \le \kappa\sqrt{n}$. In particular, the rows G_j are still independent.

Let $H_j = \langle G_j, v \rangle = G_{j1} + \cdots + G_{jn}$. Then the distribution of $H_j/\sqrt{n}$ is that of a standard Gaussian $Z \sim \mathcal{N}(0,1)$ conditioned on $|Z| \le \kappa$.

Meanwhile, if we take our row G_j and subtract H_j/n (which is its average) from each entry, then the resulting entries are independent of H_j and are jointly Gaussian.

Proposition 4.5

In the 1-planted model, we have $V - \boldsymbol{\mu} \to \mathcal{N}(0,1)^{\otimes \ell}$, for some explicit vector $\boldsymbol{\mu}$.

So the means of our cycle counts shift, but once we subtract out these new means, we still get independent Gaussians — and the conditioning also doesn't really affect the variances. In fact, there is some constant β (which can be written in terms of α and κ) such that

$$\mu_k = \frac{(2\beta)^k}{\sqrt{2k}}$$

for all k.

Now we'll consider the 2-planted model, where we're planting both v_0 and v_t. If $t \approx n$, then this is the same as the 1-planted model. But large values of t don't really matter (because the corresponding binomial coefficients will be tiny) — we only really need to consider t up to roughly $\sqrt{n}$, or more precisely $\sqrt{n} \cdot \log n$.

In this regime, the value of t doesn't matter too much, so we can just pretend that $t = 0$; and in this case, the solutions we're planting are independent, so the mean shift happens twice.

Proposition 4.6

In the 2-planted model with $t = 0$, we have $V - 2\boldsymbol{\mu} \to \mathcal{N}(0,1)^{\otimes \ell}$.

So these distributions all have the same variances, and it's possible to understand what their means are.

§4.3 The moment calculations

Now let's return to our first moment calculation. First, we saw in (4.1) that

$$\mathbb{E}[Z(G)e^{-Y}] = \mu \cdot \mathbb{E}[e^{-Y} \mid v \in S^m(G)].$$

From Proposition 4.5 we know what the cycle counts C_k look like distributionally in the 1-planted model, which tells us what Y looks like — so if we let $Z_1, \dots, Z_\ell$ be the independent Gaussians in Proposition 4.5, then we get

$$Y = \gamma_0 + \sum_k \gamma_k C_k \approx \gamma_0 + \sum_k \gamma_k \cdot \sqrt{2k} \cdot (\mu_k + Z_k),$$

which means we have

$$\mathbb{E}[Z(G)e^{-Y}] \approx \mu \cdot \mathbb{E}_{Z_k \sim \mathcal{N}(0,1)}[e^{-\gamma_0 - \sum_k \gamma_k \sqrt{2k}(\mu_k + Z_k)}].$$

We can pull out the constant terms to rewrite this as

$$\mathbb{E}[Z(G)e^{-Y}] \approx \mu \cdot e^{-\gamma_0 - \sum_k \gamma_k \sqrt{2k}\mu_k} \mathbb{E}_{Z_k \sim \mathcal{N}(0,1)}[e^{-\sum_k \gamma_k \sqrt{2k} Z_k}].$$

And a sum of independent Gaussians is still a Gaussian, so if we let $\Delta^2 = \sum_k \gamma_k^2 \cdot 2k$, then we have

$$\mathbb{E}_{Z_k \sim \mathcal{N}(0,1)}[e^{-\sum_k \gamma_k \sqrt{2k} Z_k}] = \mathbb{E}_{Z \sim \mathcal{N}(0,1)}[e^{-\Delta Z}].$$

This thing is the moment generating function of a standard Gaussian, so it's equal to $e^{-\Delta^2/2}$. So we get

$$\mathbb{E}[Z(G)e^{-Y}] \approx \mu \cdot e^{-\gamma_0 - \sum_k \gamma_k \sqrt{2k}\mu_k + \sum_k k\gamma_k^2}.$$

This concludes the first moment computation. For the secomd moment, we need to basically do the same thing, but there's an extra t parameter floating around, and we have

$$\mathbb{E}[Z(G)^2 e^{-2Y}] = \mu^2 \cdot \sum_t \binom{n}{(n+t)/2} 2^{-n} \cdot \left(\frac{p_\kappa(t)}{p_\kappa}\right)^m \cdot \widetilde{Y}(t).$$

We've already studied the ratio $p_\kappa(t)/p_\kappa$ before — it corresponds to the term with $f(b)$ from earlier (our first attempt at second moments). So the only new thing we need to study is $\widetilde{Y}(t)$, which is our doubly planted expectation. And as mentioned earlier, it suffices to consider $|t| \le \sqrt{n} \log n$ — when t is much larger than $\sqrt{n}$, the binomial coefficient is tiny, so such terms don't matter.

And we saw earlier that in this regime $\widetilde{Y}(t)$ doesn't really depend on t, and we get two mean shifts instead of one; so again letting $Z_1, \dots, Z_\ell$ be the independent normals in Proposition 4.6, we get

$$\widetilde{Y}(t) \approx \mathbb{E}_{Z_k \sim \mathcal{N}(0,1)}[e^{-2\gamma_0 - 2\sum_{k \ge 1} \gamma_k \sqrt{2k}(2\mu_k + Z_k)}].$$

And by the same computation as before, this ends up giving

$$\widetilde{Y}(t) \approx e^{-2\gamma_0 - \sum_{k \ge 1} 4\gamma_k \sqrt{2k}\mu_k + \sum_{k \ge 1} 4k\gamma_k^2}.$$

Now, why do we expect something nice to happen? The intuition is that the terms $\gamma_k C_k$ are supposed to explain all the variance in $Z(G)$. In our original second moment computation, we got that $\mathbb{E}[Z(G)^2] \approx C\mu^2$;

the idea is that these terms $\gamma_k C_k$ are sort of a proxy for C, and if we condition on them then the value of C will drop.

This maybe looks a bit like nonsense, but it'll become clear what's happening if we try dividing the moments — suppose we consider

$$\frac{\mathbb{E}[Z(G)^2 e^{-2Y}]}{\mathbb{E}[Z(G)e^{-Y}]^2}.$$

Then the terms with γ_0 will cancel, but the other ones won't (it's important that they don't cancel — if they did, then they'd be unrelated to what we care about), and we'll end up with a factor of

$$\exp\left(-\sum_k 2\gamma_k\sqrt{2k}\mu_k + \sum_k 2k\gamma_k^2\right) = \exp\left(-\sum_k 2\gamma_k(2\beta)^k + \sum_k 2k\gamma_k^2\right).$$

We'll also have a factor coming from the ratios $p_\kappa(t)/p_\kappa$ — this will end up being a simple integral, similarly to in our original second moment computation (where this integral corresponded to the extra variance that we needed to explain away). If all the γ_k were zero, then this would correspond to C; but the idea is that as we add in more of these terms (by setting γ_k appropriately), we can get corrections that account for this.

To see how to set the γ_k, we can complete the square to rewrite the corresponding factor as

$$\exp\left(2k\left(\gamma_k - \frac{(2\beta)^k}{2k}\right)^2 - \frac{(2\beta)^{2k}}{2k}\right).$$

The square is minimized when $\gamma_k = (2\beta)^k/2k$, so that's how we should set γ_k; and then the total amount of variance we get to subtract off is a factor of

$$\exp\left(-\sum_{k\geq 1}\frac{(2\beta)^{2k}}{2k}\right) = \exp\left(-\frac{1}{2}\log(1-4\beta^2)\right) = \frac{1}{\sqrt{1-4\beta^2}}.$$

And this ends up fully accounting for the extra variance, so that we end up with

$$\frac{\mathbb{E}[Z(G)^2 e^{-2Y}]}{\mathbb{E}[Z(G)e^{-Y}]^2} = 1 + o(1),$$

and we're done.