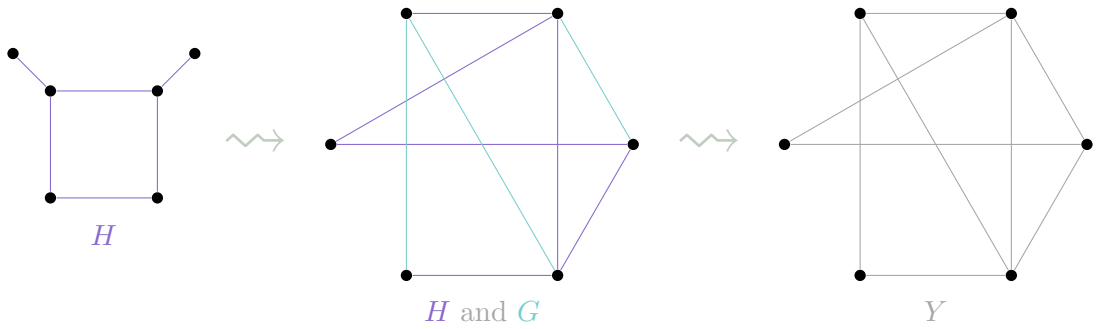
Sharp thresholds in inference of planted subgraphs

Talk by Byron Chin
Notes by Sanjana Das
April 5, 2024

This is based on work by Mossel, Niles-Weed, Sohn, Sun, and Zadik.

§1 Setup

The setup we'll consider is as follows — we imagine there's a 'signal' subgraph H, which is a fixed graph that has been planted randomly (among n vertices). And on top of this signal, there's some 'noise' — modelled by an Erdős–Rényi random graph $G = \mathcal{G}(n, p)$. We get to observe $Y = H \cup G$ (i.e., the signal overlaid with the noise), and our goal is to recover H (i.e., figure out where it was planted).



§1.1 Quantifying recovery

Question 1.1. How well can we recover H, as a function of p?

We need a way of defining what it means to recover H; the notion of recovery that we'll use is the *minimum mean squared error*.

Definition 1.2. We define $\mathsf{MMSE}(p)$ as $\mathbb{E}\,\|H - \mathbb{E}[H \mid Y]\|_2^2$ for H and Y as described above (so H is a uniform random copy of a fixed graph, $G = \mathcal{G}(n, p)$, and $Y = H \cup G$).

We'll often write G to refer to the indicator vector of the edges of G (i.e., the $\binom{n}{2}$-dimensional vector with 1's corresponding to the edges of G), and that's what we're doing here — so H and $\mathbb{E}[H \mid Y]$ are both vectors, and we're considering the 2-norm of their difference.

Example 1.3

We have that $\mathbb{E}[H]$ is the $\binom{n}{2}$-dimensional vector all of whose entries are $e(H)/\binom{n}{2}$ — we're taking the fixed graph H and randomly permuting it, so each of the $\binom{n}{2}$ pairs of vertices has this probability of being one of the edges of H.

So H is a random variable whose prior distribution is a uniform copy of the fixed graph H, and once we see Y, the best posterior estimate for H is $\mathbb{E}[H \mid Y]$. And we're trying to measure how far this posterior estimate is expected to be from the actual value of H, using the L^2 norm to measure distance.

> **Remark 1.4.** The notion of recovery we're considering here is purely information-theoretic, not algorithmic — we're looking at the best possible posterior estimate for H (once we've observed Y). By 'best possible' we mean that if we replaced $\mathbb{E}[H \mid Y]$ by any other function of Y, the error would only increase.

There are a few equivalent ways to rewrite $\mathsf{MMSE}(p)$. First, for fixed Y, we can think of the conditional expectation of $\|H - \mathbb{E}[H \mid Y]\|^2$ (given Y) as the 'variance' of H given Y, which we'll write as $\mathrm{Var}(H \mid Y)$ (we're really summing the variances of each coordinate); and then we have

$$\mathsf{MMSE}(p) = \mathbb{E}_Y[\mathrm{Var}(H \mid Y)].$$

We can also write

$$\mathsf{MMSE}(p) = \mathbb{E}[\|H\|_2^2 - \langle H, \mathbb{E}[H \mid Y]\rangle] \tag{1}$$

(here we're choosing both H and Y according to their distributions, but $\mathbb{E}[H \mid Y]$ is a function of only Y representing the conditional expectation of H after having seen just Y) — this is essentially because we can think of conditional expectation as a projection — or as

$$\mathsf{MMSE}(p) = \mathbb{E}[\|H\|_2^2 - \|\mathbb{E}[H \mid Y]\|_2^2]$$

(this makes sense from the variance formulation); these alternate formulations will be useful later.

Note that $\|H\|_2^2$ is fixed — it's always $e(H)$. This means $\mathsf{MMSE}(p)$ is between 0 and $e(H)$ for all p; we'll often divide by $e(H)$ so that it's between 0 and 1.

§1.2 All-or-nothing transitions

We're going to talk about when recoverability has a 'sharp threshold,' and to talk about thresholds we first want monotonicity.

> **Fact 1.5 —** The quantity $\mathsf{MMSE}(p)$ is monotone in p.

This makes sense — as p increases, we're adding more and more noise (as G has more edges), which means it should be harder to recover H (and the error should get bigger).

Proof. Suppose we're comparing p_1 and p_2, with $p_1 < p_2$. We can couple $G_1 \sim \mathcal{G}(n, p_1)$ and $G_2 \sim \mathcal{G}(n, p_2)$ such that $G_1 \subseteq G_2$ (in the standard way — we can construct G_2 by first taking G_1, and then making every non-edge an edge with some appropriately chosen probability). Then

$$\mathsf{MMSE}(p_1) = \mathbb{E}[\mathrm{Var}(H \mid Y_1)].$$

But if we already know Y_1, then being given Y_2 gives us no additional information about where H is (we know that H is contained in Y_1, so which additional edges got added in Y_2 is irrelevant), so

$$\mathsf{MMSE}(p_1) = \mathbb{E}[\mathrm{Var}(H \mid Y_1)] = \mathbb{E}[\mathrm{Var}(H \mid Y_1, Y_2)].$$

And dropping conditioning can only increase the expected variance (this can be shown by the law of total variance, which states that $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$; intuitively this makes sense because if we have less information then we'd expect our random variable to vary more). So

$$\mathsf{MMSE}(p_1) = \mathbb{E}[\mathrm{Var}(H \mid Y_1, Y_2)] \le \mathbb{E}[\mathrm{Var}(H \mid Y_2)] = \mathsf{MMSE}(p_2). \qquad \square$$

Remark 1.6. Here our random variables are *vectors* rather than numbers, so we actually use the law of total variance in each coordinate (and sum over all coordinates).

Then because of monotonicity, we can define a 'sharp threshold' for recoverability in the following way.

Definition 1.7. We say there is an all-or-nothing transition at p_{AN} if for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \frac{\mathsf{MMSE}(p)}{e(H)} = \begin{cases} 1 & \text{if } p > (1+\varepsilon)p_{\text{AN}} \\ 0 & \text{if } p < (1-\varepsilon)p_{\text{AN}}. \end{cases}$$

To be more precise, we're typically going to consider a *sequence* of graphs H, where H grows with n.

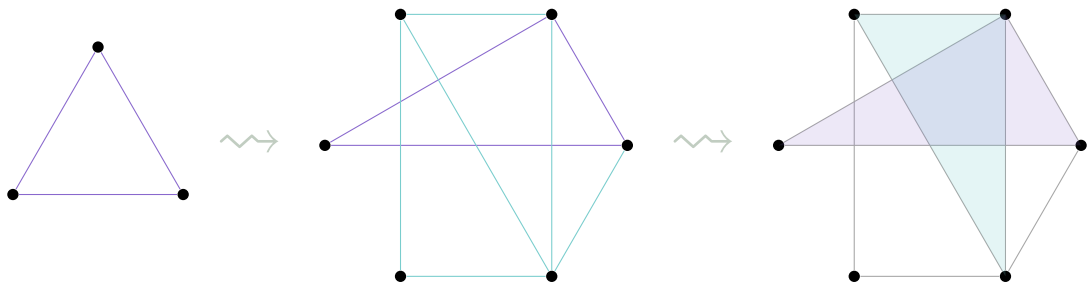
Remark 1.8. We're considering the question of whether a sharp threshold exists and where it is, but it's not obvious even that a weak threshold exists.

§1.3 Some related thresholds

Question 1.9. If such a threshold p_{AN} exists, then where is it?

To try to answer this, we'll define a few thresholds that we might expect to be related.

First, if the noise $G = \mathcal{G}(n,p)$ itself contains a copy of H, then that's bad for us — then we can make mistakes, because we're going to see two copies of H and we won't be able to figure out which one is the signal and which one comes from the noise.



So we can consider the first moment threshold, which is essentially the point at which in *expectation* you have a copy of H in the noise.

Definition 1.10. The first moment threshold, denoted p_{1M}, is the value of p for which

$$\mathbb{E}[\#\text{copies of } H \text{ in } \mathcal{G}(n,p)] = 1.$$

We can write p_{1M} explicitly as

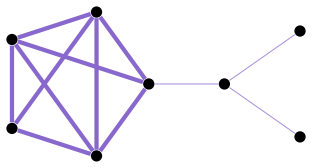
$$p_{\text{1M}} = \left(\frac{1}{\#(\text{copies of } H \text{ in } K_n)}\right)^{1/e(H)}$$

(since in $\mathcal{G}(n,p)$, each copy of H in K_n appears with probability $p^{e(H)}$).

However, the first moment threshold isn't actually what controls the point at which H appears in $\mathcal{G}(n,p)$. Instead, we need to look at the expectation threshold from the Kahn–Kalai conjecture — where we take the maximum first moment threshold over all subgraphs of H.

Definition 1.11. The expectation threshold, denoted p_{E}, is defined as $\max\{p_{1\mathrm{M}}(J) \mid J \subseteq H\}$.

In other words, the expectation threshold is the point at which in expectation, $\mathcal{G}(n,p)$ has at least one copy of every subgraph of H — certainly if $\mathcal{G}(n,p)$ contains H then it contains all its subgraphs. The point is that if H has a really dense subgraph J, then that subgraph is the last thing to appear in $\mathcal{G}(n,p)$. (At the first moment threshold for H itself, even though we see one copy of H in *expectation*, we don't expect to see a copy of J, which means we don't *typically* see a copy of H — instead, the expected number of copies of H is driven up by having a small probability of having lots of copies of H.) So we can get a better heuristic for when H appears in $\mathcal{G}(n,p)$ by looking at the first moment threshold for J instead.



And it turns out that the expectation threshold does roughly capture the point at which you should start seeing copies of H in the random noise.

Heuristically, we might expect this to kind of capture where an all-or-nothing transition occurs as well (at least, under some assumptions). There's one technicality — because in the definition of all-or-nothing transitions we're normalizing by $e(H)$, we only really care about linear-sized subgraphs (since whether or not we recover a subgraph with $o(e(H))$ edges doesn't affect the limit in Definition 1.7). So we define *generalized expectation thresholds* in the same way as the expectation threshold, but restricting J only to subgraphs of linear size.

Definition 1.12. For each $q \in (0,1)$, we define $\psi_q = \max\{p_{1\mathrm{M}}(J) \mid J \subseteq H,\, e(J) \geq qe(H)\}$.

§1.4 The main theorem

The main theorem of the paper essentially gives a characterization of when these generalized expectation thresholds ψ_q capture an all-or-nothing transition (as in the above heuristic).

Theorem 1.13

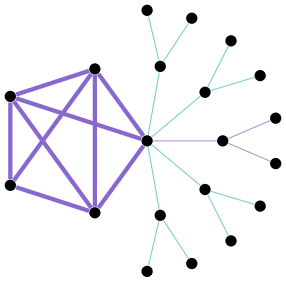
If H is sufficiently dense and delocalized, then it has an all-or-nothing transition at $p_{\mathrm{AN}} = \psi_q$ (for some $q \in (0,1)$) if and only if $\lim_{n\to\infty} \psi_q/\psi_{q'} = 1$ for all fixed $q, q' \in (0,1)$.

We're not going to define precisely what *sufficiently dense* and *delocalized* mean (since we're not going to prove this theorem). But roughly speaking, *sufficiently dense* means that H has enough edges — specifically, something like at least $|V(H)| \log |V(H)|$ edges. (This is important because it means automorphisms of H won't matter when trying to compute $p_{1\mathrm{M}}$ — if H had too few edges then automorphisms *would* matter, and $p_{1\mathrm{M}}$ and the expectation thresholds would be much harder to deal with.) And *delocalized* is a technical condition that roughly means that the edges of H don't concentrate on some tiny part.

Then in words, this theorem says that you get an all-or-nothing transition at the generalized expectation thresholds ψ_q if and only if they're all roughly equal to each other. (Note that p_{AN} is only defined up to $1 + o(1)$, so the statement makes sense — if the ψ_q's are all roughly equal to each other, then p_{AN} can be equal to any one of them.)

First we'll explain (heuristically) why it makes sense that this condition is necessary — let's think about what happens if there are $q < q'$ for which we *don't* have $\lim_{n\to\infty} \psi_q/\psi_{q'} = 1$. This means there's some

linear-sized subgraph J which is 'denser' than all other subgraphs (specifically, $e(J)$ is between $qe(H)$ and $q'e(H)$, and $p_{1M}(J)$ is significantly greater than the first moment thresholds of all the subgraphs considered for $\psi_{q'}$). Then let's consider what happens just below $p_{1M}(J)$ (the first moment threshold for our dense subgraph)? The picture suggested by random graph theory is that you won't see any copies of J, but you *will* kind of see everything else, and once you get a copy of J, it'll extend to *many* copies of H. (However, actually proving this is open in general — though it's been proven for *fixed*-size subgraphs.) So we get a kind of 'sunflower' picture where there's a dense part that's hard to make appear, but once it appears we get many copies of H.



And then when we're trying to recover the signal, we'll be able to recover this dense part but not the rest — there'll be a unique copy of J (which will correspond to the subgraph J from the signal), but there'll be many ways to extend it to a copy of H (and we won't be able to figure out which one is the actual signal — we'll have to choose randomly between them). So $\mathsf{MMSE}(p)/e(H)$ (which very roughly speaking corresponds to the 'fraction' of H that we can recover) isn't going to immediately jump from 0 to 1 — we'll have a region in between where we can recover this dense part J, but not the rest of the signal.

Here's an example where we *do* get an all-or-nothing transition (the condition of Theorem 1.13 is satisfied), but it's not at the expectation threshold.

Example 1.14

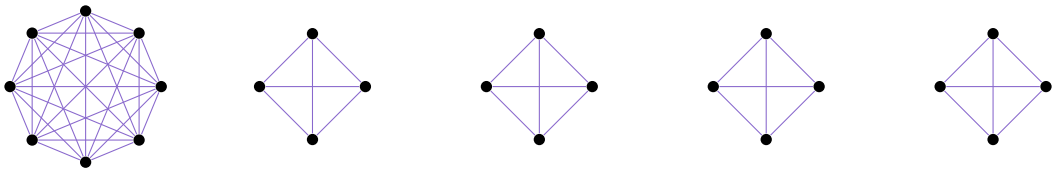
Let $k = \log n$, and let

$$H = K_{2k} \cup \underbrace{K_k \cup \cdots \cup K_k}_{k \text{ copies}}$$

(so H is a $(\log n)^2$-vertex graph consisting of a bunch of disjoint cliques, with one clique of size $2k$ and the rest of size k). Then we can compute that

$$p_{\mathrm{E}} = \frac{1}{e} \quad \text{and} \quad \psi_q = \frac{1 + o(1)}{e^2} \text{ for all } q \in (0, 1).$$

In particular, the condition of Theorem 1.13 is satisfied (we have $\lim_{n\to\infty} \psi_q/\psi_{q'} = 1$ for all $q, q' \in (0, 1)$), so there is an all-or-nothing transition at $\frac{1}{e^2}$.



(The difference between p_{E} and the ψ_q's comes from playing around with k vs. $2k$ — p_{E} comes from K_{2k}, while the ψ_q's come from taking graphs consisting of most of the K_k's — the K_{2k} doesn't affect them because it has sublinear size.)

Remark 1.15. We're assuming throughout that H grows with n, with $e(H) \to \infty$ as $n \to \infty$. It's possible to ask this question even in the case where H is a fixed graph. But in that case, there's older classical results about random graphs that essentially prove all the things we want; the issues come when H is growing.

In the rest of the talk, we're not going to prove this theorem, but we'll prove a related thing. To prove Theorem 1.13, the authors introduce the *Bernoulli model* (which we will define soon). They show there is an all-or-nothing transition in the Bernoulli model, and then deduce Theorem 1.13 by plugging things in and computing. So our plan for today is to discuss the Bernoulli model (and their proof in that setting).

This proof is quite nice in a sense — when you see a problem like this, the first thing you might think of trying is first and second moments. And the key point is that they take a more inference-based perspective on the first and second moments — they consider a *planted model* where you plant H and then add noise, and a *null model* where you don't plant H; and they find ways to pass between these two models that make the first and second moments easier to compute.

§2 The Bernoulli model

Now we'll define the Bernoulli model. Here we have a ground set $[n]$ and a collection of subsets $\mathcal{S} \subseteq \binom{[n]}{k}$ of size $|\mathcal{S}| = m$. We assume that $\mathcal{S}$ is 'symmetric' in the sense that

$$\mathbb{P}_{S\sim\mathcal{S}}[i \in S] = \frac{k}{n}$$

for all $i \in [n]$ (so if we choose a random set from $\mathcal{S}$, then each element of the ground set is equally likely to appear in it). In the graph setting, the ground set is the set of all possible edges (in particular, it has size $\binom{n}{2}$), k is the number of edges of H, and $\mathcal{S}$ is the collection of all possible copies of H; the 'symmetry' condition corresponds to the fact that a random copy of H is equally likely to contain each edge.

We then sample $S \in \mathcal{S}$ uniformly at random, which is our 'signal,' and a p-random subset $V \subseteq [n]$, which is our 'noise.' And we get to observe $Y = S \cup V$, and our goal is to recover S. (This generalizes the problem we defined in the graph setting.) We can define $\mathsf{MMSE}(p)$ in the same way as in the graph setting (where when we write expectations and 2-norms involving subsets, we're viewing subsets $S \subseteq [n]$ as their indicator vectors $\mathbf{1}_S \in \{0,1\}^n$).

Definition 2.1. We define $\mathsf{MMSE}(p)$ as $\mathbb{E}\,\|S - \mathbb{E}[S \mid Y]\|_2^2$ for S and Y described as above — i.e., $S \sim \mathcal{S}$, V is a p-random subset of $[n]$, and $Y = S \cup V$.

Definition 2.2. We say there is an all-or-nothing transition at p_{AN} if for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \frac{\mathsf{MMSE}(p)}{k} = \begin{cases} 1 & \text{if } p > (1+\varepsilon)p_{\mathrm{AN}} \\ 0 & \text{if } p < (1-\varepsilon)p_{\mathrm{AN}}. \end{cases}$$

(Note that here we're normalizing by k, which corresponds to $e(H)$.)

§2.1 A growth condition

We'll now describe a growth condition that the authors consider. First, we can define the first moment threshold in the same way as in the graph setting.

Definition 2.3. We define the first moment threshold, denoted $p_{1\text{M}}$, as the value of p for which

$$\mathbb{E}[\#\text{elements of } \mathcal{S} \text{ contained in } V] = 1.$$

As in the graph setting, $p_{1\text{M}}$ is easy to compute — this expected number is $m \cdot p^k$, so $p_{1\text{M}} = m^{-1/k}$.

Definition 2.4. We say $\mathcal{S}$ satisfies the growth condition if

$$\limsup_{n\to\infty} \sup_{1\le \ell \le k} \frac{1}{k} \log\left(\frac{\mathbb{P}_{S,S'\sim\mathcal{S}}[|S\cap S'| = \ell]}{p_{1\text{M}}^{\ell}}\right) \le 0.$$

Very vaguely speaking, the growth condition says that if we sample two copies S and S' of our signal, they don't tend to overlap too much. More precisely, if we're trying to compute a second moment for the number of possible signals $S \in \mathcal{S}$ that appear in V (with $p \approx p_{1\text{M}}$), then you're trying to look at all pairs S and S' and consider the probability that both appear. If S and S' are *disjoint*, then you'll get a probability of p^{2k}. If they have an overlap of ℓ, then you're going to lose a factor of p^{ℓ}; the growth condition says that even though we lost this factor of p^{ℓ}, there's fewer ways to get an intersection of size ℓ, and this lets us sort of gain that factor back.

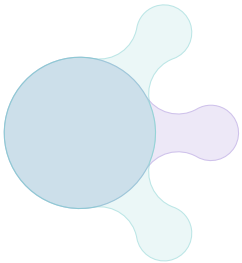
Remark 2.5. The growth condition is sort of like spreadness — it roughly means that $\mathcal{S}$ is $p_{1\text{M}}^{1-\varepsilon}$-spread for every ε or something like this. (And you should be able to prove statements similar to the ones we'll prove using spreadness instead of this growth condition; the authors use this condition because it relates to existing conditions in statistical inference for other problems.)

And the theorem the authors prove in the Bernoulli model is the following.

Theorem 2.6

In the Bernoulli model, assuming the growth condition and that $k \to \infty$, there is an all-or-nothing transition at $p_{\text{AN}} = p_{1\text{M}}(1 + o(1))$.

Right now, it's not clear how this relates to the condition $\lim_{n\to\infty} \psi_q/\psi_{q'} = 1$ from the graph setting (in Theorem 1.13). But the intuition is that both conditions are trying to capture the heuristic that the 'bad' situation (where we don't have an all-or-nothing transition) is when there's a region of p where we'll have our signal and the noise contains some copy of the signal that partially overlaps with it — because then we'll be able to recover the shared part of the signal, but not the rest (since there's more than one way to extend this shared part).



And the growth condition corresponds to controlling such overlaps (since it's saying there aren't too many copies of the signal with intersection ℓ for any ℓ).

§2.2 The 'all' regime

Now we're going to prove Theorem 2.6. We'll start with the 'all' regime, where $p \leq (1-\varepsilon)p_{1\text{M}}$ (for some $\varepsilon > 0$) and we want to show that then $\mathsf{MMSE}(p)$ is tiny — i.e., we can almost fully recover our signal.

For this, we'll use the following lemma.

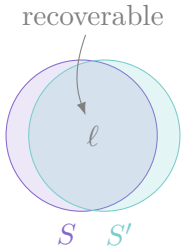
Lemma 2.7

For any $\delta > 0$, if we have

$$\sum_{\ell \leq (1-\delta)k} \mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell] m p^{k-\ell} = o(1),$$

then $\mathsf{MMSE}(p)/k \leq \delta + o(1)$.

First, why does this intuitively make sense? If we wave our hands a bit, to show that $\mathsf{MMSE}(p)/k \leq \delta + o(1)$, we want to show that we can recover all but a δ-fraction of our signal. And for this, what we *want* to happen is that if we look at the observation $Y = S \cup V$ (consisting of the signal together with the noise), then every reasonably large subset in $\mathcal{S}$ that appears in Y overlaps almost entirely (i.e., in all but a δ-fraction) with the signal. And this is essentially exactly what the condition in Lemma 2.7 is telling us — the left-hand side is roughly measuring the expected number of sets in $\mathcal{S}$ whose overlap with our signal is too small (i.e., of size less than $(1-\delta)k$).



Proof. For each $0 \leq \ell \leq k$, let $Z_\ell(S, Y)$ be the number of sets $S' \in \mathcal{S}$ such that $S' \subseteq Y$ and $|S \cap S'| = \ell$ (where S is our signal and $Y = S \cup V$ is our observation, consisting of the signal together with noise). So in words, $Z_\ell(S, Y)$ is a random variable representing the number of 'bad signals' that we get (i.e., elements of $\mathcal{S}$, which look like they *could* be the signal but aren't) which intersect the true signal S in a certain size.

We're first going to show that $\mathbb{E}[Z_\ell(S, Y)]$ is small. First, we have

$$\mathbb{E}[Z_\ell(S, Y)] = \mathbb{E}\,|S' \in \mathcal{S} \mid |S \cap S'| = \ell|\, p^{k-\ell} \tag{2}$$

(since once we've chosen S, for each S' with $|S \cap S'| = \ell$, to end up with $S' \subseteq Y$ we need the $k - \ell$ elements of $S' \setminus S$ to all be placed in V, which occurs with probability $p^{k-\ell}$). And we can rewrite this as

$$\mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell] m p^{k-\ell}$$

(we can think of (2) as involving an expectation over $S \sim \mathcal{S}$ and a sum over $S' \in \mathcal{S}$, and here we're replacing that sum with a probability over uniform $S' \sim \mathcal{S}$, which picks up a factor of m).

And this is exactly the summand in the given condition, so the given condition means that

$$\mathbb{E} \sum_{\ell \leq (1-\delta)k} Z_\ell(S, Y) = o(1),$$

and therefore by Markov's inequality we have $\sum_{\ell \leq (1-\delta)k} Z_\ell(S, Y) = 0$ with high probability.

Finally, how do we conclude that $\mathsf{MMSE}(p)$ is small? We have

$$\mathsf{MMSE}(p) = k - \mathbb{E}\langle S, \mathbb{E}[S \mid Y]\rangle$$

(as in (1)). And $\mathbb{E}[S \mid Y]$ is some weighted average over all $S' \in \mathcal{S}$ that are contained in Y (since once we see Y, the only possible signals are the ones contained in Y).

Remark 2.8. In fact, $\mathbb{E}[S \mid Y]$ is simply the average of all $S' \in \mathcal{S}$ that are contained in Y (each possible signal from $\mathcal{S}$ that's present in Y is equally likely to be the true signal when conditioned on Y, since the prior probability of choosing each $S' \in \mathcal{S}$ as the signal is the same, and the probability of obtaining this observation Y given that the signal was S' is also the same — here we're using the fact that each $S' \in \mathcal{S}$ has the same size). But this isn't necessary here.

And if $\sum_{\ell \le (1-\delta)k} Z_\ell(S, Y) = 0$, then every $S' \in \mathcal{S}$ contained in Y has overlap at least $(1-\delta)k$ with S, which means $\langle S, S'\rangle \ge (1-\delta)k$; and then averaging over S' gives $\mathbb{E}\langle S, \mathbb{E}[S \mid Y]\rangle \ge (1-\delta)k$ as well. And since this occurs with probability $1 - o(1)$ (over the choice of S and Y), this m eans

$$\mathbb{E}\langle S, \mathbb{E}[S \mid Y]\rangle \ge (1 - o(1))(1-\delta)k,$$

giving the desired conclusion. □

Finally, to conclude the 'all' case of Theorem 2.6, the point is that the growth condition implies that the condition of Lemma 2.7 holds for all $p < (1-\varepsilon)p_{1\mathrm{M}}$ and all δ — we want to show that

$$\sum_{\ell \le (1-\delta)k} \mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell] m p^{k-\ell} = o(1),$$

and we can do so by just taking the expression from the growth condition and plugging it in. Explicitly, let the left-hand side be $(*)$. Then we have $p < (1-\varepsilon)p_{1\mathrm{M}}$ and $m p_{1\mathrm{M}}^k = 1$ (by definition), so

$$(*) \le \sum_{\ell \le (1-\delta)k} \frac{\mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell]}{p_{1\mathrm{M}}^\ell} \cdot (1-\varepsilon)^{k-\ell}.$$

And the growth condition essentially states that we can find a sequence $\alpha_n \to 1$ such that

$$\frac{\mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell]}{p_{1\mathrm{M}}^\ell} \le \alpha_n^k$$

for all ℓ (the growth condition only considers $1 \le \ell \le k$, but the left-hand side is at most 1 for $\ell = 0$), so

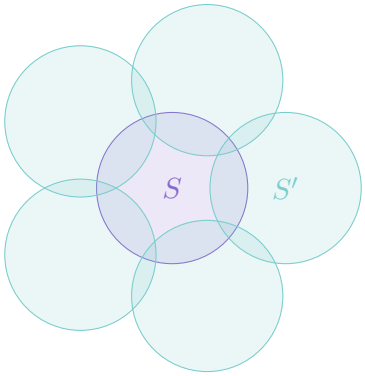
$$(*) \le \sum_{\ell \le (1-\delta)k} \alpha_n^k (1-\varepsilon)^{k-\ell} \le \sum_{\ell \le (1-\delta)k} (\alpha_n^{1/\delta}(1-\varepsilon))^{k-\ell}.$$

If we take n large enough that $\alpha_n \le (1 + \frac{\varepsilon}{2})^\delta$, then we get a geometric series whose common ratio is fixed (in terms of ε) starting with exponent δk; and since $k \to \infty$, this means its sum is $o(1)$.

So this finishes the 'all' regime.

§2.3 The 'nothing' regime

For the 'nothing' regime (where $p > (1+\varepsilon)p_{1\mathrm{M}}$ and we want to show $\mathsf{MMSE}(p)$ is really large — i.e., we can recover almost nothing of the signal), we again have our signal S and we cover it with noise V to get an observation Y. And we'd *like* to say that there's enough 'possible signals' $S' \in \mathcal{S}$ present in Y that we can't tell where our actual signal S is coming from. More precisely, in the 'all' regime we said that there were no possible signals $S' \subseteq Y$ with small overlap (with the actual signal), so here we'd like to say the opposite — that there's *many* possible signals $S' \subseteq Y$ with small overlap.



This is sort of a second moemnt-type statement, because we want to show that there's *many* of something. So the first thing you could try is computing second moments. But the fact that we're planting things makes this difficult — it's not so clear what to do.

So the authors find a nice trick — they consider the interplay between the *planted model* $Y = S \cup V$ (the model we actually have, where producing Y by planting a signal and then adding noise) and the *null model* $Y = V$ (where we just have noise, and no signal). And it turns out that we can transfer the desired statement in the planted model to a nicer statement in the null model.

§2.3.1 Transferring between models

We're going to use $\mathbb{P}$ (and $\mathbb{E}_{\mathbb{P}}$) for the probability measure corresponding to the planted model $Y = S \cup V$, and $\mathbb{Q}$ (and $\mathbb{E}_{\mathbb{Q}}$) for the probability measure corresponding to the null model $Y = V$. We're also going to write $\mathbb{P}[S, Y]$ to denote the probability of seeing a certain specific instance of S and Y.

First we have to relate $\mathbb{P}$ and $\mathbb{Q}$ in some way, so that we can transfer between the two distributions. First, for all S and Y, we have

$$\mathbb{P}[S, Y] = \frac{1}{m} \cdot \frac{\mathbb{Q}[Y]}{p^k} \cdot \mathbf{1}\{S \subseteq Y\}.$$

(Of course we must have $S \subseteq Y$. Then we have a $\frac{1}{m}$ probability of choosing S, and once we've chosen S, we need the noise V to match our observation Y on $[n] \setminus S$; the probability it matches Y on *all* of $[n]$ would be $\mathbb{Q}[Y]$, and we divide by p^k to account for the fact that we don't care about matching Y on the k elements in S — each of which contributes an extra factor of p to $\mathbb{Q}[Y]$, since $S \subseteq Y$.)

And mp^k is precisely the expected number of elements of $\mathcal{S}$ that are present in Y in the null model (i.e., the number of 'possible signals' — because each of the m elements of $\mathcal{S}$ appears with probability p^k). So letting $Z(Y)$ denote the number of elements of $\mathcal{S}$ that appear in Y, we get that

$$\mathbb{P}[S, Y] = \frac{\mathbf{1}\{S \subseteq Y\}}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]} \cdot \mathbb{Q}[Y]. \tag{3}$$

And finally, summing over all S gives that

$$\mathbb{P}[Y] = \frac{Z(Y)}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]} \cdot \mathbb{Q}[Y]. \tag{4}$$

This means the probability of seeing a certain observation Y in the planted model is related to the probability of seeing Y in the null model, up to a factor that looks like a Radon–Nikodym derivative.

This is already enough to get one immediate consequence.

Lemma 2.9

For all $\varepsilon > 0$, we have $\mathbb{P}[Z(Y) \le \varepsilon \mathbb{E}_{\mathbb{Q}}[Z(Y)]] \le \varepsilon$.

In words, this says that it's unlikely that the number of possible signals we see in Y in the planted model is much smaller than the *expected* number we'd see in the null model.

Proof. We can write the probability on the left-hand side as the expectation of an indicator, i.e.,

$$\mathbb{E}_{\mathbb{P}}[\mathbf{1}\{Z(Y) \le \varepsilon \mathbb{E}_{\mathbb{Q}}[Z(Y)]\}]$$

(where we think of this indicator as a function of Y). And now we can transfer this to an expectation over $\mathbb{Q}$ by multiplying by the same Radon–Nikodym derivative-like factor used to transfer between $\mathbb{P}[Y]$ and $\mathbb{Q}[Y]$ in (4) (by writing out the expectation as a sum over Y, and replacing each $\mathbb{P}[Y]$ with $\mathbb{Q}[Y]$ times that extra factor), so we get that this is equal to

$$\mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}\{Z(Y) \le \varepsilon \mathbb{E}_{\mathbb{Q}}[Z(Y)]\} \cdot \frac{Z(Y)}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]}\right].$$

And finally, whenever the indicator function is 1, this extra factor is at most ε; so we get that this expression is at most ε for *all* Y, and therefore its expectation is at most ε. □

This means if in the null model we'd *expect* there to be lots of possible signals in the observation, then in the planted model there's *typically* lots of possible signals.

In order to actually solve the 'nothing' regime, though, it's not enough to just say that there's lots of possible signals in Y; what we *actually* want is a version of this saying that there's lots of possible signals with *small overlap* with the actual signal. First, we need the following statement about the *expected* number of possible signals with given overlap.

Lemma 2.10

Let $Z_\ell(S, Y)$ denote the number of sets $S' \in \mathcal{S}$ with $|S \cap S'| = \ell$ (as before), and let $Z_2(\ell, Y)$ denote the number of pairs $S, S' \in \mathcal{S}$ such that $S, S' \subseteq Y$ and $|S \cap S'| = \ell$.

Then for all $0 \le \ell \le k$, we have

$$\mathbb{E}_{\mathbb{P}}[Z_\ell(S, Y)] = \frac{\mathbb{E}_{\mathbb{Q}}[Z_2(\ell, Y)]}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]}.$$

Intuitively, we can think of $Z_2(\ell, Y)$ as a second moment-type quantity (we're counting pairs of sets). So this lemma transfers a first moment computation in the planted model to a second moment computation in the null model; and this turns out to be much easier to compute (because everything is independent).

Proof. We can first use (3) to transfer the expectation over $\mathbb{P}$ to an expression involving $\mathbb{Q}$, by writing

$$\mathbb{E}_{\mathbb{P}}[Z_\ell(S, Y)] = \sum_{S,Y} \mathbb{P}[S, Y] Z_\ell(S, Y) = \sum_{S,Y} \frac{\mathbf{1}\{S \subseteq Y\}\mathbb{Q}[Y]}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]} \mathbb{Z}_\ell(S, Y).$$

And now we can expand out

$$Z_\ell(S, Y) = \sum_{S' \in \mathcal{S}} \mathbf{1}\{S' \subseteq Y, \, |S \cap S'| = \ell\},$$

and plugging this into the above equation gives

$$\mathbb{E}_{\mathbb{P}}[Z_\ell(S, Y)] = \sum_{Y} \frac{\mathbb{Q}[Y]}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]} \sum_{S \in \mathcal{S}} \mathbf{1}\{S \subseteq Y\} \sum_{S' \in \mathcal{S}} \mathbf{1}\{S' \subseteq Y, \, |S \cap S'| = \ell\}.$$

But the sum over S and S' is precisely $Z_2(\ell, Y)$ by definition (we're counting the number of pairs $S, S' \in \mathcal{S}$ with $S, S' \subseteq Y$ and $|S \cap S'| = \ell$), so we get that

$$\mathbb{E}_{\mathbb{P}}[Z_\ell(S,Y)] = \sum_Y \frac{\mathbb{Q}[Y] Z_2(\ell, Y)}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]} = \frac{\mathbb{E}_{\mathbb{Q}}[Z_2(\ell,Y)]}{\mathbb{E}_{\mathbb{Q}}[Z(Y)]}. \qquad \square$$

Next, we'll prove a sort of opposite of the 'all' statement in Lemma 2.7 — that if the probabilities of having *big* overlaps (i.e., overlaps of size at least δk) is small, then our error is large.

Lemma 2.11

For any $\delta > 0$, if we have

$$\sum_{\ell \geq \delta k} \mathbb{P}[|S \cap S'| = \ell] p^{-\ell} = o(1),$$

then $\mathsf{MMSE}(p)/k \geq 1 - \delta - o(1)$.

Remark 2.12. Lemma 2.7 had an extra factor of mp^k on the left-hand side; the reason for this was that we proved it by computing the *expected* number of sets with small overlap and showing it was small, and this expectation corresponded to that extra factor. Meanwhile, here we don't have such a term. This matters because when applying Lemma 2.7 to deduce the 'all' case of Theorem 2.6 (by showing that the growth condition implies the hypothesis of Lemma 2.7), we had $p < (1-\varepsilon)p_{1\mathrm{M}}$, so it was useful to have a positive power of p (note that $mp_{1\mathrm{M}}^k = 1$). Meanwhile, for the 'nothing' case of Theorem 2.6 we have $p > (1+\varepsilon)p_{1\mathrm{M}}$, so it's important that the exponent of p is negative.

Proof. First, as seen in Remark 2.8, we have that

$$\mathsf{MMSE}(p) = k - \mathbb{E}_{\mathbb{P}}\langle S, \mathbb{E}_{\mathbb{P}}[S \mid Y]\rangle = k - \mathbb{E}_{\mathbb{P}}\left[\frac{1}{Z(Y)} \sum_{S' \subseteq Y} \langle S \cap S' \rangle\right],$$

since once we've seen Y, the best posterior estimate for S is the uniform distribution on all 'possible signals' present in Y (i.e., all $S' \in \mathcal{S}$ with $S' \subseteq Y$). and $\langle S, S' \rangle = |S \cap S'|$. We'll call the latter term $(*)$, so our goal is to show that

$$(*) = \mathbb{E}_{\mathbb{P}}\left[\frac{1}{Z(Y)} \sum_{S' \subseteq Y} |S \cap S'|\right] \leq (\delta - o(1))k.$$

To do so, terms with $|S \cap S'| \leq \delta k$ contribute a total of at most δk to the sum (for any given S and Y), so we can pull them out. The number of remaining terms S; in the sum is $\sum_{\ell \geq \delta k} Z_\ell(S,Y)$, and each contributes $|S \cap S'| \leq k$ to the sum, so we get

$$(*) \leq \delta k + k \cdot \mathbb{E}_{\mathbb{P}}\left[\frac{1}{Z(Y)} \sum_{\ell \geq \delta k} Z_\ell(S,Y)\right].$$

So now our goal is to show that this expectation, which we'll call $(**)$, is $o(1)$. To do so, we'll split the expectation into two cases depending on whether the denominator $Z(Y)$ is very small. More specifically, fix any $\varepsilon > 0$. Then in Lemma 2.9 we saw that $Z(Y) \leq \varepsilon \mathbb{E}_{\mathbb{Q}}[Z(Y)]$ with probability at most ε; in this case, all we can say is that the expression we're taking the expectation of is at most 1 (because $\sum_{\ell \geq 0} Z_\ell(S,Y) = Z(Y)$ by definition). And otherwise we can lower-bound the denominator by $\varepsilon \mathbb{E}_{\mathbb{Q}}[Z(Y)]$ and pull it out; this gives

$$(**) \leq \varepsilon \cdot 1 + \frac{1}{\varepsilon \cdot \mathbb{E}_{\mathbb{Q}}[Z(Y)]} \cdot \mathbb{E}_{\mathbb{P}}\left[\sum_{\ell \geq \delta k} Z_\ell(S,Y)\right].$$

And now we can use Lemma 2.10 to transfer this expectation in the planted model to a second moment in the null model — so we have

$$(**) \le \varepsilon + \frac{\mathbb{E}_{\mathbb{Q}}[\sum_{\ell \ge \delta k} Z_2(\ell, Y)]}{\varepsilon \cdot \mathbb{E}_{\mathbb{Q}}[Z(Y)]^2}.$$

Finally, it remains to compute both of these expectations. And the nice thing is that they're both under the null model, so we can essentially just expand things out as sums of indicators and write down the exact probabilities. In more detail, in the numerator we're counting pairs $S, S' \in \mathcal{S}$ with intersection $\ell \ge \delta k$ that both are subsets of Y; for any pair with $|S \cap S'| = \ell$, the probability that both S and S' appear in Y is $p^{2k-\ell}$ (since all $2k - \ell$ elements of $S \cup S'$ need to be placed in Y), so the numerator is

$$\mathbb{E}_{\mathbb{Q}}\left[\sum_{\ell \ge \delta k} Z_2(\ell, Y)\right] = \sum_{\ell \ge \delta k} m^2 \mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell] \cdot p^{2k-\ell}$$

(where $m^2 \mathbb{P}_{S,S' \sim \mathcal{S}}[|S \cap S'| = \ell]$ is the number of pairs with intersection ℓ). Meanwhile, in the denominator, $\mathbb{E}_{\mathbb{Q}}[Z(Y)]$ is counting the expected number of sets $S \in \mathcal{S}$ which appear in Y; this is mp^k (since each S appears in Y with probability p^k), so we get

$$\mathbb{E}_{\mathbb{Q}}[Z(Y)]^2 = m^2 p^{2k}.$$

This means their ratio is precisely the quantity in the hypothesis of Lemma 2.11, which is $o(1)$; and taking $\varepsilon \to 0$ gives $(**) = o(1)$, as desired. □

And finally, we can again deduce the 'nothing' case of Theorem 2.6 by noting that the growth condition implies the hypothesis of Lemma 2.11 for $p > (1+\varepsilon)p_{1\text{M}}$ (for any $\varepsilon > 0$ and $\delta > 0$) — the growth condition means there's a sequence $\alpha_n \to 1$ with

$$\mathbb{P}_{S,S' \in \mathcal{S}}[|S \cap S'| = \ell] p_{1\text{M}}^{-\ell} \le \alpha_n^k$$

for all ℓ, so then the expression in Lemma 2.11 is at most $\sum_{\ell \ge \delta k} (1+\varepsilon)^{-\ell} \alpha_n^k$, which is $o(1)$ as $n, k \to \infty$.