Analysis of Deshpande, Abbe and Montanari's Asymptotic Mutual Information for the Two-Groups Stochastic Block Model [1]

Max Lovig

May 4, 2023

1 Results:

This paper analyzes the stochastic block model (SBM), specifically under two groups. It quantifies the asymptotic per-vertex mutual information and asymptotic estimation metrics. To define this model, consider a label vector X with the following properties:

1. each element of $X = \{X_i\}_{i\in 1:n}$, is ± 1 (indicating they are in the positive or negative group)
2. It is a bisection, i.e. $\sum_i X_i = 0$ (We also can analyze the case where each index is drawn from Bernoulli $(1/2)$ as this will concentrate about such X).

Now lets consider an undirected graph of n nodes with labels $1:n$. We want to connect edges between these node to represent a relationship between the nodes. We would like to do this in a way that represents if the nodes belong to the same group. We give two distributions P and Q in which edge weights are drawn between two nodes (say $E_{i,j}$ for the edge between nodes i and j) can be drawn from in the following way

$$E_{i,j} \sim \begin{cases} P & \text{if } X_i X_j = 1 \\ Q & \text{if } X_i X_j = -1 \end{cases}$$

We then form a graph G with vertices $1:n$ and edges E. Depending on our choice of P and Q we are able to elicit some well known models. For example the famous spiked wigner model can be created from

$$P = N\left(\sqrt{\frac{\mu}{n}}, 1\right) \quad Q = N\left(-\sqrt{\frac{\mu}{n}}, 1\right)$$

In our case we will study a different choice of P and Q, that of

$$P = \text{Bernoulli}\,(p_n) \quad Q = \text{Bernoulli}\,(q_n)$$

This elicits the stochastic block model.

As for the results, this paper makes a further assumption on the values of p_n and q_n as $n \to \infty$. With $\bar{p}_n = \frac{1}{2}(p_n + q_n)$ we require

$$n\bar{p}_n(1-\bar{p_n}) \to \infty \tag{A}$$

this is a difficult assumption to justify as there is a large mount of interest in a situation where edge generating distribution is

$$p_n = \frac{a}{n} \quad q_n = \frac{b}{n}$$

This is due to this model having a average edge count of order 1. We can see that this choice of P and Q barely does not comply with (A) as

$$\bar{p}_n = \frac{a+b}{n} \implies n\bar{p}_n\left(1-\bar{p}_n\right) = \Theta(1)$$

Notice, however, that if we had $n^{1-\varepsilon}$ on the denomentaor of p_n and q_n we would have:

$$\bar{p}_n = \frac{a+b}{n^{1-\epsilon}} \implies n\bar{p}_n\left(1-\bar{p}_n\right) \to \infty$$

Thus, if a and b are particularly large can still justify this approach as we can consider them as having an arbitrarily slow growth rate. This means that we can handle bounded yet large average degree cases (like above) with a vanishing error.

As we mentioned previously, the purpose of this paper was to show the limiting per-vertex mutual information $\frac{1}{n}I(X;G)$. This quantity is of its own interest and, perhaps even more useful, it intimately related with our ability to estimate X from G. The authors are able to reduce the SBM model into a "single-letterization". This means that we can consider the per-index mutual information as a function of a much simpler scalar model. The single-letter model used is the following Gaussian channel[1]:

$$Y_0 = \sqrt{\gamma}X_0 + Z_0$$

where we have $X_0 \sim \text{Uniform}\left(\{-1\},\{+1\}\right)$ and $Z_0 \sim \text{Normal}\left(0,1\right)$. We can then define

$$I\left(\gamma\right) = E\log\left(\frac{dp_{y|x}\left(Y_0\left(\gamma\right)x_0\right)}{dp_y\left(Y_0\left(\gamma\right)\right)}\right)$$
$$mmse\left(\gamma\right) = E\left[\left(X_0 - E\left[X_0|Y_0\left(\gamma\right)\right]\right)^2\right]$$

I can now introduce the first of two major results:

Theorem 1.1: For any $\lambda > 0$, let $\gamma_* = \gamma_*\left(\lambda\right)$ be the largest non-negative solution of the equation

$$\gamma = \lambda\left(1 - mmse\left(\gamma\right)\right) \tag{$\star_1$}$$

We name $\gamma_*\left(\lambda\right)$ the effective signal to noise ratio. Further we define $\Psi\left(\gamma,\lambda\right)$ by

$$\Psi\left(\gamma,\lambda\right) = \frac{\lambda}{4} + \frac{\gamma^2}{4\lambda} + I\left(\gamma\right)$$

Let G and X be distributed according to the stocastic block model with n vertices and parameters p_n and q_n. Define

$$\lambda_n = n\left(p_n - q_n\right)^2 / \left(4\bar{p}_n\left(1-\bar{p}_n\right)\right)$$

Then as $n \to \infty$, we assume that $\lambda_n \to \lambda$ and $n\bar{p}_n\left(1-\bar{p}_n\right) \to \infty$. Then,

$$\lim_{n\to\infty}\frac{1}{n}I\left(x;G\right) = \Psi\left(\gamma_*\left(\lambda\right),\lambda\right)$$

It turns out that this result on the per-vertex mutual information can also be worked into a statement about estimation metrics. The canonical metric here is the matrix minimum mean square error, for reasons we will see in the proofs below. We define the matrix minimum mean square error as:

$$MMSE_n\left(\lambda\right) = \frac{1}{n\left(n-1\right)}E\left[||XX^T - E\left[XX^T|G\right]||_F^2\right]$$

[1]This paper indicates this scalar model by a 0 subscript

We also note that the $MMSE \in (0,1)$ and where $MMSE = 0$ means recovery of all of the labels for X up to sign and $MMSE = 1$ means we are not better off then random guessing. Thus we can reformulate the information theoretic Theorem 1.1 into a theorem about large sample statistical estimation:

Theorem 1.4 Under the assumptions of Theorem 1.1, for any $\lambda > 0$, let $\gamma_* = \gamma_*(\lambda)$ be the largest non-negative solution[a] of the equation

$$\gamma = \lambda (1 - mmse(\gamma)) \tag{$\star_1$}$$

We have the following limit:

$$\lim_{n\to\infty} MMSE_n(\lambda_n) = 1 - \frac{\gamma_*(\lambda)^2}{\lambda^2}$$

When $\lambda \leq 1$, we have $\gamma_* = 0$ and $\lim MMSE = 1$
When $\lambda > 1$, we have $\gamma_* > 0$ and $\lim MMSE =< 1$

[a] Throughout this paper I may abbreviate γ_* if the parameters it relies on are obvious

So we recover the phase transition we have in class

$$\lambda = \lim \lambda_n = \lim n (p_n - q_n)^2 / (4\bar{p}_n (1 - \bar{p}_n)) \gtrless 1$$

Unfortunately, the $MMSE$ is not the most intuitive way to think about this estimation problem. Clearly a count of the correct versus incorrect labels (up to a sign change) is much more relvent to our problem. A metric which captures this information quite well is the overlap, defined as

$$\text{Overlap}(\lambda) = \frac{1}{n} \sup_{\hat{x}} E[|\langle X, \hat{x}\rangle|]$$

Where the supremum is taken over all estimators $\hat{x}$ where the value of each index of $\hat{x}$ is either plus of minus 1. Intuitively this measures (up to a sign change) what proportion the best estimator overlaps with the true labels. This paper establish a similar phase transition as above where of $\lambda \leq 1$ then in the large n limit we have

$$\text{Overlap}(\lambda) \to 0$$

which mean no estimator can achieve recovery better than random guessing. When $\lambda \geq 1$ we have in the large n limit that[2]

$$0 < \frac{\gamma_*(\lambda)^2}{\lambda^2} \leq \text{Overlap}(\lambda)$$

Meaning that we have establish a similar bound for the overlap and can conclude the threshold for correlated recovery agrees with the $MMSE$ metric above. We can only achieve correlated recovery when:

$$\lim n (p_n - q_n)^2 / (4\bar{p}_n (1 - \bar{p}_n)) > 1$$

2 Proofs

Below we have a flow chart which show the order of logic between different sections of this paper. The direction of the arrows show which Lemmas/Theorems are used to prove certain results in our papers.

[2] I am going to state these without proof since the real meat of the paper is the above two theorem.

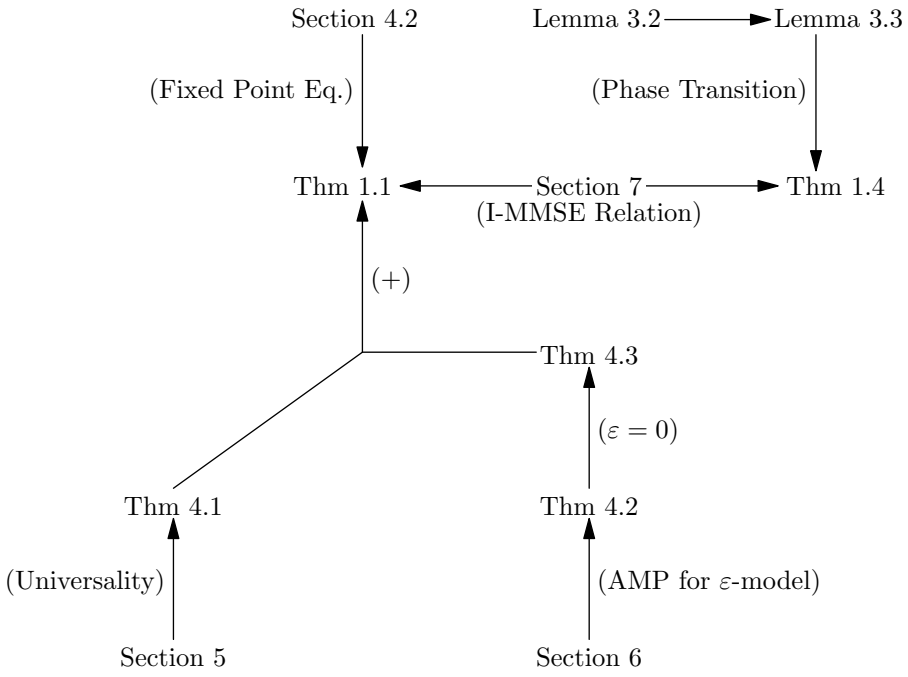

Figure 1: The order of logic in proving Theorems 1.1 and 1.4

2.1 (Fixed Point Eq.)

Here I will give a brief introduction to AMP theory and give some intuition on where the fixed point equation comes from.

First we embed our guassian channel into a more extensive model. We consider alongside our output Y_0 a secondary output $X_0(\varepsilon)$ which is the output from a binary erasure channel. Meaning we reveal the true label X_0 with probability $1-\varepsilon$ (erasing it with probability ε). This gives model

$$\begin{aligned} X_0(\varepsilon) &= B_0 X_0 && (B_0 \sim \text{Bernoulli}\,(\varepsilon)) \\ Y_0 &= \sqrt{\gamma} X_0 + Z_0 \end{aligned}$$

The AMP algorithm for the vectorization of this more general scalar problem is of the form

$$\begin{aligned} x^{t+1} &= \frac{Y}{\sqrt{n}} f_t\left(x^t, X\left(\varepsilon\right)\right) - b_t f_{t-1}\left(x^{t-1}, X\left(\varepsilon\right)\right) \\ b_t &= \frac{1}{n}\sum_{i=1}^{n} f_t'\left(x_i^t, X\left(\varepsilon\right)_i\right) \end{aligned}$$

Where f_t is a sequence of lipshitz continuous functions which we apply component-wise to x^t and b_t is the so-called Onsager correction term. This term removes a bias that is induced in each step of the algorithm and allows for analysis through the state evolution equations. We denote these set of recursive equations as:

$$\begin{aligned} \mu_{t+1} &= \sqrt{\lambda} E\left[X_0 f_t\left(\mu_t X_0 + \sigma_t Z_0, X_0\left(\varepsilon\right)\right)\right] \\ \sigma_{t+1}^2 &= E\left[f_t\left(\mu_t X_0 + \sigma_t Z_0, X_0\left(\varepsilon\right)\right)^2\right] \end{aligned}$$

Where the expectation is taken over $X_0 \sim U(\{-1,1\})$, $Z_0 \sim N(0,1)$ and $B_0 \sim \text{Bernoulli}\,(1-\varepsilon)$.

A common choice of f_t is the *Bayes Optimal* function, letting $D = \mu_t X_0 + \sigma_t Z_0 = y, X_0 = s$, we define it as

$$f_t = E\left[X_0|D\right]$$

The reason we make this choice is that it elicits a fixed point equation involving the *mmse*. This can be seen by

$$\begin{aligned}
\mu_{t+1} &= \sqrt{\lambda} E\left[X_0 E\left[X_0|D\right]\right] \\
&= \sqrt{\lambda} E\left[E\left[X_0 E\left[X_0|D\right]|D\right]\right] \\
&= \sqrt{\lambda} E\left[E\left[X_0|D\right] E\left[X_0|D\right]\right] \\
&= \sqrt{\lambda} E\left[E\left[X_0|D\right]^2\right] \\
&= \sqrt{\lambda}\sigma_{t+1}^2
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{t+1}^2 &= 2\sigma_{t+1}^2 - \sigma_{t+1}^2 \\
&= 2\frac{\mu_{t+1}}{\sqrt{\lambda}} - \sigma_{t+1}^2 \\
&= E\left[X_0\right] - E\left[X_0\right] + 2E\left[X_0 E\left[X_0|D\right]\right] - E\left[E\left[X_0|D\right]^2\right] \\
&= E\left[X_0^2\right] - E\left[\left(X_0 - E\left[X_0|D\right]\right)^2\right]
\end{aligned}$$

Due to the distribution of X_0 the first term is 1. For the second term we use the fact that observing $Y = \mu_t X_0 + \sigma_t Z_0$ is equivalent to observing $Y = \sqrt{\lambda}\sigma_t^2 X_0 + \sigma_t Z_0$. This is equivalent to observing $\tilde{Y} = \frac{Y}{\sigma_t} = \sqrt{\lambda\sigma_t^2} X_0 + Z_0$. Thus,

$$\sigma_{t+1}^2 = 1 - (1-\varepsilon)\, mmse\left(\lambda\sigma_t^2\right)$$

Where the $(1-\varepsilon)$ is due to the proportion of times we simply observe the true label. We can unify these two state evolution parameters into one by defining

$$\gamma_t = \lambda\sigma_t^2 = \sqrt{\lambda}\mu_t$$

Giving us the recursion

$$\begin{aligned}
\frac{\gamma_{t+1}}{\lambda} &= 1 - (1-\varepsilon)\, mmse\left(\gamma_t\right) \\
\implies \gamma_{t+1} &= \lambda\left(1 - (1-\varepsilon)\, mmse\left(\gamma_t\right)\right)
\end{aligned}$$

Which converges to fixed point γ_* which satisfies the all to familiar

$$\gamma_* = \lambda\left(1 - (1-\varepsilon)\, mmse\left(\gamma_*\right)\right)$$

When using AMP as a proof technique we will show that the MSE_{AMP} of the Bayes-AMP estimate is a upper bound for the $MMSE$ with a vanishing gap as n grows larger. We can formalize the limits of such a test function of an AMP estimate in the following result due to Javanmard and Montanari,

Lemma 4.4 [2]: Given f_t defined above and psuedo-lipshitz test function ψ we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi\left(x_i^t, X_i, X\left(\varepsilon\right)_i\right) = E\left[\psi\left(\mu_t X_0 + \sigma_t Z_0, X_0, X_0\left(\varepsilon\right)\right)\right]$$

We can then see the limit as $t \to \infty$, i.e. our algorithm converges to its fixed point. One may be interested as to why we care about the largest non-negative solution to the fixed point equation $(\star_1)$. We will see in (phase transition) subsection that this is the correct fixed point to choose. The existence of multiple fixed points actually hints at a more general theory for AMP models which I will briefly discuss in the discussion section.

2.2 (Universality)

This section reveals why we chose $\lambda_n = \frac{(p_n - q_n)^2}{4\bar{p}_n(1-\bar{p}_n)}$ as the SBM's signal to noise ration. Some intuition for the choice is as follows. Let $\Delta_n = \frac{p_n - q_n}{2}$ we can then see that conditioned on X:

$$G_{ij} \sim \text{Bernoulli}\left(\bar{p}_n + \Delta_n X_i X_j\right)$$

Unconditional on X we have

$$E[G_{ij}] = \bar{p}_n \quad V[G_{ij}] = \bar{p_n}(1-\bar{p}_n)$$

So we can consider a shifted and scaled version of $\breve{G}_{ij} = \frac{G_{ij} - \bar{p}_n}{\sqrt{\bar{p}_n(1-\bar{p}_n)}}$. Under this new $\breve{G}_{ij}$ we see that conitioned on X

$$E\left[\breve{G}_{ij}|X\right] = \frac{\Delta_n}{\sqrt{\bar{p}_n\left(1-\bar{p}_n\right)}} X_i X_j$$

Perhaps (since the expectation and variance is 0 and 1), we can relate this to a more familiar model, what is more familiar then Normality!

Recall the well-known spiked wigner model as,

$$Y\left(\lambda\right) = \sqrt{\frac{\lambda}{n}} X X^T + Z$$

Where X is the label vector as before and $Z \sim$ GOE. We see that the expectation of Y_{ij} conditioned on X is

$$E[Y_{ij}|X] = \sqrt{\frac{\lambda}{n}} X_i X_j$$

So to give a connection between the two we solve

$$\begin{aligned}
\sqrt{\frac{\lambda}{n}} X_i X_j &= \frac{\Delta_n}{\sqrt{\bar{p}_n\left(1-\bar{p}_n\right)}} X_i X_j \\
\lambda &= \frac{n\Delta_n^2}{\bar{p}_n\left(1-\bar{p}_n\right)} \\
&= \frac{n(p_n - q_n)^2}{4\bar{p}_n\left(1-\bar{p}_n\right)}
\end{aligned}$$

Of course this is just matching the conditional expectation. To solidify the connection to the Mutual information more analysis is needed. We bound the per-vertex mutual information $\frac{1}{n}|I\left(X;G\right) - I\left(X;Y\right)|$ by a value which vanishes as $n \to \infty$.

We can start with a obvious reduction to the mutual information as

$$I(X;Y) = E\left[\log \frac{\partial p_{Y|X}\left(Y|X\right)}{\partial p_Y\left(Y\left(\lambda\right)\right)}\right]$$

We see that for both the SBM and wigner models we can write

$$I(X;Y) = E\left[\log \frac{\partial p_{Y|X}\left(Y|X\right)}{\sum_x 2^{-n}\ \partial p_{Y|x}\left(Y|x\right)}\right] = n\log 2 + E\left[\log \frac{\partial p_{Y|X}\left(Y|X\right)}{\sum_x\ \partial p_{Y|x}\left(Y|x\right)}\right]$$

$$I(X;G) = E\left[\log \frac{\partial p_{G|X}\left(G|X\right)}{\sum_x 2^{-n}\ \partial p_{G|x}\left(G|x\right)}\right] = n\log 2 + E\left[\log \frac{\partial p_{G|X}\left(G|X\right)}{\sum_x\ \partial p_{G|x}\left(G|x\right)}\right]$$

We then aim to pull out a second term in both models. Define the Hamiltonian $\mathcal{H}(V,\lambda) = \sum_{i<j} V_{ij}(x_i x_j - X_i X_j) + \frac{\lambda}{n} x_i x_j X_i X_j$ and the function $\phi(V) = \log \sum_x \exp\left(\mathcal{H}(V,\lambda)\right)$. For the Spiked Wigner model we have:

Lemma 5.1

$$E\left[\log \frac{\partial p_{Y|X}\left(Y|X\right)}{\sum_x \partial p_{Y|x}\left(Y|x\right)}\right] = \frac{\left(n-1\right)\lambda}{2} - E_{X,Z}\left[\phi\left(X; Z\sqrt{\lambda/n}, \lambda, n\right)\right]$$

Sketch: This is due to

$$\begin{aligned}
\log \frac{\partial p_{Y|X}\left(Y|X\right)}{\sum_x \partial p_{Y|x}\left(Y|x\right)} &= \log \frac{\exp\left(-||Y - \sqrt{\frac{\lambda}{n}} XX^T||_F^2/4\right)}{\sum_x \exp\left(-||Y - \sqrt{\frac{\lambda}{n}} xx^T||_F^2/4\right)} \\
&= \log \frac{\exp\left(-||Z||_F^2/4\right)}{\sum_x \exp\left(-||Z + \sqrt{\frac{\lambda}{n}} xx^T||_F^2/4\right)} \\
&= -\log \sum_x \exp\left(\sum_{i<j} Z_{ij}\sqrt{\frac{\lambda}{n}}\left(x_i x_j - X_i X_j\right) - \frac{\lambda}{2n}\left(x_i x_j - X_i X_j\right)^2\right) \\
&= -\log \sum_x \exp\left(\sum_{i<j} Z_{ij}\sqrt{\frac{\lambda}{n}}\left(x_i x_j - X_i X_j\right) - \frac{\lambda}{2n}\left(\left(x_i x_j\right)^2 - 2x_i x_j X_i X_j + \left(X_i X_j\right)^2\right)^2\right) \\
&= -\log \sum_x \exp\left(\sum_{i<j} Z_{ij}\sqrt{\frac{\lambda}{n}}\left(x_i x_j - X_i X_j\right)\right) \exp\left(-\frac{\lambda}{2n}\sum_{i<j} 2 - 2x_i x_j X_i X_j\right) \\
&= -\log \sum_x \exp\left(\sum_{i<j} Z_{ij}\sqrt{\frac{\lambda}{n}}\left(x_i x_j - X_i X_j\right)\right) \exp\left(-\frac{\lambda}{2n}\left(n\left(n-1\right) - \sum_{i<j} 2x_i x_j X_i X_j\right)\right) \\
&= \frac{\lambda\left(n-1\right)}{2} - \log \sum_x \exp\left(\sum_{i<j} Z_{ij}\sqrt{\frac{\lambda}{n}}\left(x_i x_j - X_i X_j\right) + \sum_{i<j} 2x_i x_j X_i X_j\right) \\
&= \frac{\lambda\left(n-1\right)}{2} - \log \sum_x \exp\left(\mathcal{H}(Z, \sqrt{\lambda/n})\right) \\
&= \frac{\lambda\left(n-1\right)}{2} - \phi(Z, \sqrt{\lambda/n})
\end{aligned}$$

We take the expectation to reach our conclusion. ∎

Now to analyze the SBM model we will introduce some extra notation (which will be revealed throughout the sketch):

Lemma 5.3

$$E\left[\log \frac{\partial p_{G|X}\left(G|X\right)}{\sum_x \partial p_{G|x}\left(G|x\right)}\right] = \frac{\left(n-1\right)\lambda_n}{2} - E_{X,\tilde{G}}\left[\phi\left(X, \tilde{G}, \lambda_n, n\right)\right] + O\left(\frac{n\lambda_{3/2}}{\sqrt{n\bar{p}_n\left(1-\bar{p}_n\right)}}\right)$$

Sketch: We Have

$$\begin{aligned}
\log \frac{\partial p_{G|X}\left(G|X\right)}{\sum_x \partial p_{G|x}\left(G|x\right)} &= \log \frac{\prod_{i<j}(\bar{p}_n + \Delta_n X_i X_j)^{G_{ij}}(1-\bar{p}_n - \Delta_n X_i X_j)^{1-G_{ij}}}{\sum_x \prod_{i<j}(\bar{p}_n + \Delta_n x_i x_j)^{G_{ij}}(1-\bar{p}_n - \Delta_n x_i x_j)^{1-G_{ij}}} \\
&= -\log \sum_x \exp\left(\sum_{i<j} G_{ij} \log\left(\frac{\bar{p}_n + \Delta_n x_i x_j}{\bar{p}_n + \Delta_n X_i X_j}\right) + (1-G_{ij})\log\left(\frac{1-\bar{p}_n - \Delta_n x_i x_j}{1-\bar{p}_n - \Delta_n X_i X_j}\right)\right) \\
&= -\log \sum_x \exp\left(\mathcal{H}\right)
\end{aligned}$$

When $x = \pm 1$ we have the following identity

$$\log\left(a + bx\right) = \frac{1}{2}\log\left(a^2 + b^2\right) + \frac{x}{2}\log\left(\left(a+b\right)/\left(a-b\right)\right)$$

We can extend this to get for x and $y = \pm 1$

$$\log\left(\frac{a+bx}{a+by}\right) = \left(\frac{x}{2} - \frac{y}{2}\right)\log\left(\frac{a+b}{a-b}\right)$$

With $a = \bar{p}_n$, $b = \Delta_n$, $x = x_i x_j$ and $y = X_i X_j$ for the first term and similarly for the second term, applying this to our work on the log ratio above we see

$$\begin{aligned}
\mathcal{H} &= \sum_{i<j}\left(x_i x_j + X_i X_j\right)\left(\frac{G_{ij}}{2}\log\left(\frac{p_n + \Delta_n}{p_n - \Delta_n}\right) + \frac{1-G_{ij}}{2}\log\left(\frac{1-p_n-\Delta_n}{1-p_n+\Delta_n}\right)\right) \\
&= \sum_{i<j}\left(x_i x_j + X_i X_j\right)\left(\frac{G_{ij}}{2}\log\left(\frac{1+\Delta_n/p_n}{1-\Delta_n/p_n}\right) + \frac{1-G_{ij}}{2}\log\left(\frac{1-\Delta_n/(1-p_n)}{1+\Delta_n/(1-p_n)}\right)\right)
\end{aligned}$$

Here we can invoke a bound for small enough z that

$$\left|\frac{1}{2}\log\left(\frac{1+z}{1-z}\right) - z\right| \leq z^3$$

with $z = \Delta_n/\bar{p}_n$ for the first term and simularly for the second term, this leads to

$$= \sum_{i<j}\left(x_i x_j + X_i X_j\right)\left(\frac{\Delta_n G_{ij}}{\bar{p}_n} - \frac{\Delta_n(1-G_{ij})}{1-\bar{p}_n}\right) + \mathrm{err}_n$$

Here err_n is the difference z^3 terms, we have $\text{err}_n \lessapprox \sqrt{n}$. Letting $\tilde{G}_{ij} = \frac{\Delta_n G_{ij} - \Delta_n \bar{p}_n - \Delta_n^2 X_i X_j)}{\bar{p}_n(1-\bar{p}_n)}$ we can rewrite as

$$
\begin{aligned}
&= \sum_{i<j} (x_i x_j + X_i X_j) \left(\frac{\Delta_n G_{ij}(1-\bar{p}_n) - \Delta_n (1-G_{ij})\bar{p}_n}{\bar{p}_n(1-\bar{p}_n)} \right) + \text{err}_n \\
&= \sum_{i<j} (x_i x_j + X_i X_j) \left(\frac{\Delta_n G_{ij} - \Delta \bar{p}_n}{\bar{p}_n(1-\bar{p}_n)} \right) + \text{err}_n \\
&= \sum_{i<j} (x_i x_j + X_i X_j) \left(\frac{\Delta_n G_{ij} - \Delta_n \bar{p}_n - \Delta_n^2 X_i X_j + \Delta_n^2 X_i X_j}{\bar{p}_n(1-\bar{p}_n)} \right) + \text{err}_n \\
&= \sum_{i<j} (x_i x_j + X_i X_j) \left(\tilde{G}_{ij} + \frac{\Delta_n^2 X_i X_j}{\bar{p}_n(1-\bar{p}_n)} \right) + \text{err}_n \\
&= \sum_{i<j} (x_i x_j + X_i X_j) \left(\tilde{G}_{ij} + \frac{\lambda_n X_i X_j}{n} \right) + \text{err}_n \\
&= \sum_{i<j} -\frac{\lambda_n}{n} (X_i X_j)^2 + \tilde{G}_{ij} \left(x_i x_j + X_i X_j \right) + \frac{\lambda_n}{n} x_i x_j X_i X_j + \text{err}_n \\
&= -\frac{\lambda_n (n-1)}{2} + \sum_{i<j} \mathcal{H}(\tilde{G}, \lambda) + \text{err}_n
\end{aligned}
$$

So we can then see

$$
\begin{aligned}
\log \frac{\partial p_{G|X}(G|X)}{\sum_x \partial p_{G|x}(G|x)} &= -\log \sum_x \exp \left(-\frac{\lambda_n (n-1)}{2} + \sum_{i<j} \mathcal{H}(\tilde{G}, \lambda_n) + \text{err}_n \right) \\
&= \frac{(n-1)\lambda_n}{2} - \log \sum_x \exp \left(\sum_{i<j} \mathcal{H}(\tilde{G}, \lambda_n) + \text{err}_n \right) \\
&= \frac{(n-1)\lambda_n}{2} + \phi(\tilde{G}_{ij}, \lambda_n)
\end{aligned}
$$

By characterizing the err_n term and taking the expectation we get our result. ☭

We are left with the expectations of ϕ and $err \lessapprox \sqrt{n}$, we can bound the difference of the former by the following:

Lemma 5.5

$$
E_{X,\tilde{G}} \left[\phi \left(X, \tilde{G}, \lambda_n, n \right) \right] = E_{X,Z} \left[\phi \left(X; Z\sqrt{\lambda/n}, \lambda, n \right) \right] + O \left(\frac{n\lambda^{3/2}}{\sqrt{n\bar{p}_n(1-\bar{p}_n)}} + n|\lambda_n - \lambda| \right)
$$

Putting this all together we can see that

Proposition 4.1: Assume that as $n \to \infty$, $\lambda_n \to \lambda$ and $n\bar{p}_n(1-\bar{p}_n) \to \infty$. Then there is a constant C independent of n such that

$$
\frac{1}{n} |I(X;G) - I(X;Y)| \leq C \left(\frac{\lambda^{3/2}}{\sqrt{n\bar{p}_n(1-\bar{p}_n)}} + |\lambda_n - \lambda| \right)
$$

2.3 (AMP for ε-model)

Recall our general channel we defined in the (Fixed Point Eq.) subsection

$$\begin{aligned} X_0(\varepsilon) &= B_0 X_0 && (B_0 \sim \text{Bernoulli}(\varepsilon)) \\ Y_0 &= \sqrt{\gamma} X_0 + Z_0 \end{aligned}$$

Under this model we can generalize our previous results to

Proposition 4.2: For any $\lambda > 0$, $\epsilon \in (0,1)$ let $\gamma_* = \gamma_*(\lambda, \epsilon)$ be the largest non-negative solution of the equation

$$\gamma = \lambda(1 - (1-\epsilon) mmse(\gamma)) \qquad (\star_1)$$

We define $\Psi(\gamma, \lambda, \epsilon)$ by

$$\Psi(\gamma, \lambda, \epsilon) = \frac{\lambda}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \epsilon \log 2 + (1-\epsilon) I(\gamma)$$

Then,

$$\lim_{n\to\infty} \frac{1}{n} I(X; X(\epsilon), Y) = \Psi(\gamma_*(\lambda, \epsilon), \lambda, \epsilon)$$

Initially alot of my confusion was based the form of Ψ, why is it that way? Intuitively can we construct Ψ from some simple observations about our fixed point equation and the AMP. First it is very illuminating to see the derivatives of Ψ. Before that, however, we invoke a well-known result.

I-MMSE Relation [3]: We have

$$\frac{\partial I(\gamma)}{\partial \gamma} = \frac{1}{2} mmse(\gamma)$$

We take the differential of Ψ with respect to λ and γ to give us

$$\begin{aligned} \frac{\partial \Psi(\gamma, \lambda, \varepsilon)}{\partial \gamma} &= \frac{\gamma}{2\lambda} - \frac{1}{2} + \frac{1}{2}(1-\varepsilon) mmse(\gamma) \\ \frac{\partial \Psi(\gamma, \lambda, \varepsilon)}{\partial \lambda}|_{\gamma=\gamma_*} &= \frac{1}{4}\left(1 - \frac{\gamma_*^2}{\lambda^2}\right) \end{aligned}$$

First we notice that the fixed point equation

$$\gamma = \lambda(1 - (1-\varepsilon) mmse(\gamma))$$

can be derived by setting the derivative with respect to γ and rearranging.

Second we notice the derivative with respect to λ is this term connected with theorem 1.4, this is no coincidence. We will see that this derivative is equal to the limiting matrix mean squared error of the AMP algorithm (which we will eventually show is the $MMSE$) when both the number of iterations and the size of the graph approaches infinity. This is defined rigorously in:

Lemma 6.3

$$MSE_{AMP}(t; \lambda, \varepsilon) = \lim_{n\to\infty} MSE_{AMP}(t; \lambda, \varepsilon, n) = 1 - \frac{\gamma_t^2}{\lambda^2}$$

$$MSE_{AMP}(\lambda, \varepsilon) = \lim_{t\to\infty} MSE_{AMP}(t; \lambda, \varepsilon) = 1 - \frac{\gamma_*^2}{\lambda^2}$$

Sketch: We have that

$$\begin{aligned} MSE_{AMP}(t;\lambda,\varepsilon,n) &= \frac{1}{n^2}E\left[||XX^T - \hat{x}^t\hat{x}^{t^T}||_F^2\right] \\ &= \frac{1}{n^2}E\left[||XX^T||_F^2 + ||\hat{x}^t\hat{x}^{t^T}||_F^2 - 2||X\hat{x}^{t^T}||_F^2\right] \\ &= E\left[\frac{||X||^4}{n^2} + \frac{||\hat{x}^t||^4}{n^2} - 2\frac{\langle X,\hat{x}^t\rangle}{n^2}\right] \end{aligned} \tag{$\star_2$}$$

Where $\star_2$ holds by the following (demonstrated on the first term)

$$\begin{aligned} ||XX^T||_F^2 &= \sum_{i,j}(X_iX_j)^2 \\ &= \left(\sum_i x_i\right)^2 \\ &= ||X||_2^2 \end{aligned}$$

We have $||X||^2 = n^2$ so the first term in $(\star_2)$ evaluates to 1. Now we consider the last term. We know denote $\phi(\hat{x}_i^{t-1}, X_i, X(\varepsilon)_i) = X_0E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right]$. By the State-evolution of AMP we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_0E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right] = E\left[X_0E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right]\right] = \frac{\mu_t}{\sqrt{\lambda}} = \frac{\gamma_t}{\lambda}$$

As $\langle X,\hat{x}^t\rangle/n$ is bounded we then have

$$\lim_{n\to\infty}E\left[\frac{\langle X,\hat{x}^t\rangle}{n^2}\right] = \frac{\gamma_t^2}{\lambda^2}$$

We then find for the middle term in $(\star_2)$ using $\phi(\hat{x}_i^{t-1}, X_i, X(\varepsilon)_i) = E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right]^2$ that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right]^2 = E\left[E\left[X_0|\mu_{t-1}X_0 + \sigma_{t-1}Z_0, X\left(\varepsilon\right)_0\right]^2\right] = \sigma_t^2 = \frac{\gamma_t}{\lambda}$$

And simularly

$$\lim_{n\to\infty}E\left[\frac{||\hat{x}^t||^4}{n^2}\right] = \frac{\gamma_t^2}{\lambda^2}$$

Thus the limit value of $(\star_2)$ is

$$1 + \frac{\gamma_t^2}{\lambda^2} - 2\frac{\gamma_t^2}{\lambda^2} = 1 - \frac{\gamma_t^2}{\lambda^2}$$

Now letting $t\to\infty$ we then get the second limit as

$$1 - \frac{\gamma_*^2}{\lambda^2}$$

∎

Finally to recover the function Ψ we need to know its value at a given point. We can derive such a value through the methods of:

Lemma 6.2

$$\lim_{\lambda\to 0}\Psi\left(\gamma_*\left(\lambda,\varepsilon\right),\lambda,\varepsilon\right) = \varepsilon\log 2$$
$$\lim_{\lambda\to\infty}\Psi\left(\gamma_*\left(\lambda,\varepsilon\right),\lambda,\varepsilon\right) = \log 2$$

Sketch: Construct the minimal linear mean squared error estimator. This estimator will have MSE of

$$(1-\rho)^2\sigma_X^2 = \left(1 - \left(\frac{\sqrt{\gamma}}{1+\gamma}\right)^2\right) = \frac{1}{1+\gamma}$$

So we know that $mmse(\gamma) \leq \frac{1}{1+\gamma}$. Now plugging into $(\star_1)$ we have

$$\begin{aligned}
\gamma &= \lambda\left(1 - (1-\varepsilon)\frac{1}{1+\gamma}\right) \\
(1+\gamma)\gamma &= \lambda(1+\gamma) - \lambda(1-\varepsilon) \\
\gamma^2 + \gamma(1-\lambda) - \lambda\varepsilon &= 0
\end{aligned}$$

We then see that the largest root is of form

$$\gamma_{LB} = \frac{1}{2}\left((\lambda-1) + \sqrt{(\lambda-1)^2 + 4\lambda\varepsilon}\right) \overset{\lambda\to\infty}{\to} = \lambda - (1-\varepsilon) + o \qquad (o \lesssim 1)$$

We also observe that γ_* is bounded above by λ by plugging λ into $(\star_1)$ and seeing that the left side is larger than the right trivially. Thus

$$\max(0, \gamma_{LB}) \leq \gamma_* \leq \lambda$$

So as $\lambda \to \infty$ we have $\gamma^* = \lambda$ and plugging into Ψ gives

$$\Psi = \frac{\lambda}{4} + \frac{\lambda}{4} - \frac{\lambda}{2} + (1-\varepsilon)I(\lambda) + \varepsilon\log 2 = (1-\varepsilon)I(\lambda) + \varepsilon\log 2$$

We have that $\lim_{\lambda\to\infty} I(\lambda) = H(X) - H(Y|X) = \log 2$. This is due the signal in the Gaussian channel begin so high that knowing Y determines X. So,

$$\lim_{\lambda\to\infty} \Psi = \log 2$$

For the result as $\lambda \to 0$ we have the inequality $\gamma^* \leq \lambda$. Thus,

$$0 < \Psi = \frac{\lambda}{4} + \frac{{\gamma^*}^2}{4\lambda} - \frac{\gamma^*}{2} + (1-\varepsilon)I(\gamma^*) + \varepsilon\log 2 \leq \frac{\lambda}{4} + \frac{\lambda}{4} - \frac{\lambda}{2} + (1-\varepsilon)I(\gamma) + \varepsilon\log 2 = (1-\varepsilon)I(\lambda^*) + \varepsilon\log 2$$

We see that as the signal approaches 0 that X and Y are independent to $I(\lambda) \to 0$. Thus,

$$\lim_{\lambda\to 0} \Psi = \varepsilon\log 2$$

∎

Putting this all together we have shown:

Lemma 6.4

$$\Psi(\gamma_*(\lambda,\varepsilon)\lambda,\varepsilon) = \varepsilon\log 2 + \frac{1}{4}\int_0^\lambda MSE_{AMP}\left(\breve{\lambda},\varepsilon\right)\,\partial\breve{\lambda}$$

Now we can relate how $\Psi(\gamma_*(\lambda,\varepsilon)\lambda,\varepsilon)$ is equivalent to the mutual information. Before we approach this argument we will need the following:

Remark 6.5 We have the following facts about the mutual information

1. $|I(X;Y(\lambda),X(\varepsilon)) - I\left(XX^T;Y(\lambda),X(\varepsilon)\right)| \leq \log 2$
2. $\lim_{n\to\infty}\frac{1}{n}I\left(XX^T;X(\varepsilon),Y(0)\right) = \varepsilon\log 2$
3. $\lim_{\lambda\to\infty}\liminf_{n\to\infty}\frac{1}{n}I\left(XX^T;X(\varepsilon),Y(\lambda)\right) = \log 2$

The first fact is due to the following: For any r.v. R we have

$$\begin{aligned} H\left(X|R\right) - H\left(XX^T|R\right) &= H\left(X, XX^T|R\right) - H\left(XX^T|R\right) - \left(H\left(X, XX^T|R\right) - H\left(X|R\right)\right) \\ &= H\left(X|XX^T, R\right) - H\left(XX^T|X, R\right) \\ &= H\left(X|XX^T, R\right) && \text{(As } X \text{ deterines } XX^T\text{)} \end{aligned}$$

We then see that

$$\begin{aligned} H\left(X|XX^T, R\right) &\leq H\left(X|XX^T\right) && \text{(Conditioning reduced entropy)} \\ &\leq \log 2 && \text{(only two atoms } X \text{ and } -X \text{ can make } XX^T\text{)} \end{aligned}$$

Leading us to conclude

$$0 \leq H\left(X|R\right) - H\left(XX^T|R\right) \leq \log 2 \qquad (\dagger)$$

We then see that

$$\begin{aligned} I\left(X; Y\left(\lambda\right), X\left(\varepsilon\right)\right) - I\left(XX^T; Y\left(\lambda\right), X\left(\varepsilon\right)\right) &= H\left(X\right) - H\left(X|Y, X\left(\varepsilon\right)\right) - \left(H\left(XX^T\right) - H\left(XX^T|Y, X\left(\varepsilon\right)\right)\right) \\ &= H\left(X\right) - H\left(XX^T\right) - \left(H\left(X|Y, X\left(\varepsilon\right)\right) + H\left(XX^T|Y, X\left(\varepsilon\right)\right)\right) \end{aligned}$$

Applying $(\dagger)$ for $R = \varnothing$ and $R = (Y, X(\varepsilon))$ gives us the first result.

The second and third result can be logic-ed through is a similar way to the mutual information in the proof of lemma 6.2. If there is no signal $\lambda = 0$ then we can only reliable message $\varepsilon \log 2$ nats of information (through the Bernoulli realization of $X(\varepsilon)$). Similarly if we have a infinite signal to noise ratio we can realize the mutual information as the entropy of XX^T minus a vanishing conditional entropy. ☭

Now we can sketch an argument for Proposition 4.2:

Sketch: We then see in the large n limit that we have the correct values for $\lambda = 0$ and $\lambda = \infty$

Now if we can match the derivative of Ψ and the derivative of $\frac{1}{n} I\left(XX^T; Y\left(\lambda\right), X\left(\varepsilon\right)\right)$ we could formalize this argument and show proposition 4.2

First we apply the conditional I-MMSE (GSV05) to give use

$$\begin{aligned} \frac{1}{n} \frac{\partial I\left(XX^T; Y\left(\lambda\right), X\left(\varepsilon\right)\right)}{\partial \lambda} &= \frac{1}{n^2} \sum_{i<j} E\left[\left(X_i X_j - E\left[X_i X_j | Y\left(\lambda\right), X\left(\varepsilon\right)\right]\right)^2\right] \\ &= \frac{1}{4} MMSE\left(\lambda, \varepsilon, n\right) \\ &\leq \frac{1}{4} MSE_{AMP}\left(t; \lambda, \varepsilon, n\right) \\ &\quad \text{(As the AMP estimator can at best achieve the minimum error)} \end{aligned}$$

We can not set up the following chain of inequalities (this is why we above showed the limit of Ψ in λ)

$$
\begin{aligned}
(1-\varepsilon)\log 2 &= \lim_{\lambda\to\infty}\liminf_{n\to\infty}\frac{1}{n}\left(I\left(X;Y\left(\lambda\right),X\left(\varepsilon\right)\right)-I\left(X;Y\left(0\right),X\left(\varepsilon\right)\right)\right)\\
&= \lim_{\lambda\to\infty}\liminf_{n\to\infty}\frac{1}{4}\int_0^{\lambda} MMSE\left(\breve{\lambda},\varepsilon,n\right)\,\partial\breve{\lambda}\\
&\leq \lim_{\lambda\to\infty}\lim_{t\to\infty}\limsup_{n\to\infty}\frac{1}{4}\int_0^{\lambda} MSE_{AMP}\left(t;\breve{\lambda},\varepsilon,n\right)\,\partial\breve{\lambda} \qquad (\star_3)\\
&= \lim_{\lambda\to\infty}\lim_{t\to\infty}\limsup_{n\to\infty}\frac{1}{4}\int_0^{\lambda} MSE_{AMP}\left(t;\breve{\lambda},\varepsilon,n\right)\,\partial\breve{\lambda}\\
&= \lim_{\lambda\to\infty}\lim_{t\to\infty}\frac{1}{4}\int_0^{\lambda} MSE_{AMP}\left(t;\breve{\lambda},\varepsilon\right)\,\partial\breve{\lambda}\\
&= \lim_{\lambda\to\infty}\frac{1}{4}\int_0^{\lambda} MSE_{AMP}\left(\breve{\lambda},\varepsilon\right)\,\partial\breve{\lambda}\\
&= \lim_{\lambda\to\infty}\left(\Psi\left(\gamma_*,\lambda,\varepsilon\right)-\Psi\left(\gamma_*,0,\varepsilon\right)\right)\\
&= (1-\varepsilon)\log 2
\end{aligned}
$$

So we can then see that the inequality in $(\star_3)$ is an equality and as $MMSE(\lambda,\varepsilon,n)\leq MSE_{AMP}(t;\lambda,\varepsilon,n)$ for all λ we have

$$MSE_{AMP}(\lambda,\varepsilon)=\lim_{n\to\infty}MMSE(\lambda,\varepsilon,n)$$

For almost all λ, we can argue through montonicity and continuity to extend this to all λ. So, we can have (ignoring total formality)

$$\lim_{n\to\infty}\frac{1}{n}\frac{\partial I\left(X;Y\left(\lambda\right),X\left(\varepsilon\right)\right)}{\partial\lambda}=\lim_{n\to\infty}\frac{1}{n}\frac{\partial I\left(XX^T;Y\left(\lambda\right),X\left(\varepsilon\right)\right)}{\partial\lambda}=\lim_{n\to\infty}\frac{1}{4}MMSE\left(\lambda,\varepsilon,n\right)=MSE_{AMP}(\lambda,\varepsilon)$$

So Ψ also matches the derivative of $\lim_{n\to\infty}\frac{1}{n}I\left(X;Y\left(\lambda\right),X\left(\varepsilon\right)\right)$ and its value at $\lambda=0$ so we can conclude:

$$\lim_{n\to\infty}\frac{1}{n}I\left(X;Y\left(\lambda\right),X\left(\varepsilon\right)\right)=\Psi(\gamma_*(\lambda,\varepsilon),\lambda,\epsilon)$$

∎

2.4 $(\varepsilon=0,+)$

Now that we have analyzed the more general channel we can use these results to specify to the case that $\epsilon=0$ we can easily see that setting $\varepsilon=0$ gives us our original Ψ in Theorem 1.1. To set the stage for thiis limit we must show two things

1. $I(X;X(\varepsilon),Y)/n$ has a well defined limit. This was proved in proposition 4.2
2. We can show the difference between $I(X;X(\varepsilon),Y)/n$ and $I(X;Y)$ is vanishing when $\varepsilon\to 0$, we can see this as
$$\left|\frac{1}{n}I(X;X(\varepsilon),Y)-\frac{1}{n}I(X;Y)\right|\leq\frac{1}{n}I(X;X(\varepsilon),Y)\leq\epsilon\log 2$$

Accounting for this we have

$$\lim_{n\to\infty}\frac{1}{n}I\left(X;Y\right)=\lim_{\varepsilon\to 0}\Psi\left(\gamma_*,\lambda,\varepsilon\right)$$

We have that $\Psi(\gamma, \lambda, \varepsilon)$ is continuous in ε and γ and that the unique solution to the ε fixed point equation $\gamma_*(\lambda, \varepsilon)$ converges to $\gamma_*(\lambda)$. As the function $1 - mmse(\gamma)$ is smooth and concave, we know γ_* satisfies the equation:

$$\gamma = \lambda(1 - mmse(\gamma))$$

We then have

$$\lim_{n\to\infty} \frac{1}{n} I\left(X;Y\right) = \Psi(\gamma_*\left(\lambda\right), \lambda)$$

We then must show that

$$\lim_{n\to\infty} \frac{1}{n} |I(X;G) - \frac{1}{n}\Psi(\gamma_*(\lambda), \lambda)| \to 0$$

This was shown in Proposition 4.1

2.5 (Phase Transition)

We consider, again, our single-letter model. We preciously defined

$$mmse\left(\gamma\right) = E\left[\left(X_0 - E\left[X_0|Y_0\left(\gamma\right)\right]\right)^2\right]$$

Under scalar model, we can simplify the $mmse$ term above. First we have

$$\begin{aligned}
E\left[X|Y_0\right] &= P\left(X_0 = 1|Y_0\right) - P\left(X_0 = -1|Y_0\right) \\
&= \frac{f\left(Y_0 - \sqrt{\gamma}\right) - f\left(Y_0 + \sqrt{\gamma}\right)}{f\left(Y_0 - \sqrt{\gamma}\right) + f\left(Y_0 + \sqrt{\gamma}\right)} && (f \text{ is the stanard gaussian PDF}) \\
&= \frac{e^{Y_0\sqrt{\gamma}} - e^{-Y_0\sqrt{\gamma}}}{e^{Y_0\sqrt{\gamma}} + e^{-Y_0\sqrt{\gamma}}} \\
&= \tanh Y_0\sqrt{\gamma} \\
&= \tanh X_0\gamma + Z_0\sqrt{\gamma}
\end{aligned}$$

And then we can reduce the $mmse$ to

$$\begin{aligned}
mmse(\gamma) &= E\left[\left(X_0 - E\left[X_0|Y_0\left(\gamma\right)\right]\right)^2\right] \\
&= E\left[X_0^2 - 2X_0E\left[X_0|Y_0\left(\gamma\right)\right] + E\left[X_0|Y_0\left(\gamma\right)\right]^2\right] \\
&= 1 - 2E\left[X_0E\left[X_0|Y_0\left(\gamma\right)\right]\right] + E\left[E\left[X_0|Y_0\left(\gamma\right)\right]^2\right] \\
&= 1 - 2E\left[E\left[X_0|Y_0\left(\gamma\right)\right]^2\right] + E\left[E\left[X_0|Y_0\left(\gamma\right)\right]^2\right] && \text{(By The Tower PRoprty of Cond. Exp)} \\
&= 1 - E\left[E\left[X_0|Y_0\left(\gamma\right)\right]^2\right] \\
&= 1 - E\left[\tanh\left(\gamma + \sqrt{\gamma}Z_0\right)^2\right] \\
&\coloneqq 1 - G\left(\gamma\right)
\end{aligned}$$

Notice that $(\star_1)$ can be rewritten as the equivalent fixed point equation

$$\gamma = \lambda G\left(\gamma\right)$$

Thus an analysis of $G\left(\gamma\right)$ will reveal the possible solutions to equation $(\star_1)$. We note the following without proof

1. Utilizing the continuity of tanh we have that G is a continious function

2. We see that $G(0) = E[\tanh(0)^2] = E[0] = 0$
3. As $\lim_{x\to\infty} \tanh x = 1$ we also have that $\lim_{\gamma\to\infty} G(\gamma) = 1$
4. we also have that G is monotone increasing and strictly concave on $[0, \infty)$ so it will have at most 2 fixed points (one of which is trivially 0)

We also need to analyze the derivative of G with respect to γ we do this as

$$\begin{aligned}
E\left[\frac{\partial}{\partial\gamma}\tanh\left(\gamma+\sqrt{\gamma}Z\right)^2\right] &= E\left[2\tanh\left(\gamma+\sqrt{\gamma}Z\right)\left(1-\left(\tanh\left(\gamma+\sqrt{\gamma}Z\right)\right)^2\right)\left(1+\frac{1}{2\sqrt{\gamma}}Z\right)\right] \\
&= E\left[2\left(\gamma-\sqrt{\gamma}Z\right)\left(1-\left(\gamma+\sqrt{\gamma}Z\right)^2\right)\left(1+\frac{1}{2\sqrt{\gamma}}Z\right)\right] \qquad \text{(Taylor Expand)} \\
&= E\left[Z^2\right] \qquad \text{(only term free of } \gamma \text{ and has } Z \text{ of even power)} \\
&= 1
\end{aligned}$$

Thus we have $\lambda G'(0) = \lambda$, leading us to conclude that a non-zero fixed point will only occur for $\lambda > 1$. This fixed point is trivially larger than the other known fixed point 0 so we define the value of γ_* as the largest solution to the fixed point equation.

2.6 (I-MMSE Relation)

Section 7 presents an argument which bounds the difference of Ψ and the $MMSE$.

First we will find it convenient two rewrite how p_n and q_n are defined. Once could motivate this form as a version of Noise $\pm$ Signal. We parameterize p_n and q_n with parameter θ as

$$p_n = \bar{p}_n + \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}\theta \quad q_n = \bar{p}_n - \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}\theta$$

We can then formalize the difference between the derivative of the mutual information and the MMSE through the following lemma (which is proved from results by Measson, Montanari, Richardson, and Urbanke [4]).

Lemma 7.2 Let $I(X;G)$ be the mutual information of the two-group stochastic block models with parameters $p_n = p_n(\theta)$ and $q = q_n(\theta)$. Then

$$\left|\frac{1}{n}\frac{\partial I(X;G)}{\partial\theta} - \frac{1}{4}MMSE_n(\theta)\right| \lessapprox \max\left(\sqrt{\frac{\theta}{n\bar{p}_n(1-\bar{p}_n)}}, \frac{1}{n}\right)$$

We have now formalized the ability to take the derivative of $\lim_{n\to\infty}\frac{\partial}{\partial\theta}I(X;G) = \frac{1}{4}MMSE$. We can then

finish the relationship from Theorem 1.1 to Theorem 1.4 By:

$$\begin{aligned}
\lim_{n\to\infty}\int_{\lambda_1}^{\lambda_2}\frac{1}{4}MMSE\,\partial\theta &= \lim_{n\to\infty}\int_{\lambda_1}^{\lambda_2}\frac{1}{n}\frac{\partial I\left(X;G\right)}{\partial\theta}\,\partial\theta && \text{(Lemma 7.2)}\\
&= \Psi(\gamma_*\left(\lambda_2\right),\lambda_2)-\Psi\left(\gamma_*\left(\lambda_1\right),\lambda_1\right) && \text{(Theorem 1.1)}\\
&= \int_{\lambda_1}^{\lambda_2}\frac{\partial}{\partial\lambda}\Psi\left(\gamma_*\left(\lambda\right),\lambda\right)\,\partial\lambda && \text{(Lemma 6.4)}\\
&= \int_{\lambda_1}^{\lambda_2}\left(1-\frac{\gamma_*\left(\lambda^2\right)}{\lambda^2}\right)\,\partial\lambda && \text{(Lemma 6.3)}\\
&= \int_{\lambda_1}^{\lambda_2}\left(1-\frac{\gamma_*\left(\theta^2\right)}{\theta^2}\right)\,\partial\theta && \text{(Re-Assign Dummy Variable)}
\end{aligned}$$

Thus we have (through similar arguments as the end of Subsection 2.3)

$$\lim\frac{1}{4}MMSE=\left(1-\frac{\gamma_*\left(\theta^2\right)}{\theta^2}\right)$$

3 Discussion

Over all I really enjoyed this paper, I think answers a important question for a very relevant model in applications. Th SBM is used in many areas (social media political networks always come to mind) and knowing the fundamental limits of achievability is important. I would like to mention what I believe is the most innovative aspect of this paper, the impact AMP it leaves on theory. Of course this is hard to judge as the question answered here is very directly related with one specific model. We can, however, extract an argument from this paper and compare with a similar argument from around this time.

This paper gives a good introduction to using AMP as a proof mechanism. One of the largest success in AMP theory is that one can utilize a well thought-out choice for f_t and, in many cases, this choice will provide an easy to analyze algorithm which can achieve the minimum of a estimation metric. This cannot be understated, I will mention an application whose question seems unconnected but their analysis is almost identical.[3]

M-Estimation with large p [5]:

Consider a Model

$$Y=X\theta_0+W$$

Where Y is a response vector, X is a design matrix and W is a per index iid noise vector. We can define a M-estimator through a non-negative Convex Function $\rho:\mathbb{R}\to\mathbb{R}^{\geq 0}$

$$\hat{\theta}_M=\arg\min_{\theta}\sum_{i=1}^{n}\rho(Y_i-\langle X_i,\theta\rangle)$$

We see that choosing $\rho(\cdot)=\frac{\partial}{\partial\theta}\log(f_W(\cdot)))$ leads to the MLE estimate. We have from classical statistical theory that the asymptotic distribution of $\hat{\theta}-\theta$ is $N(0,V)$, where the asymptotic variance V is

$$V=U(\phi,F_W)(X^TX)^{-1}$$

[3]Thank you to Zhou Fan, I found out this result through his course

With $U(\phi, F) = \frac{\int \phi^2 \, \partial F}{(\int \phi' \, \partial F)^2}$ and $\phi = \rho'$. We see that, as expected, using $\rho(\cdot) = \frac{\partial}{\partial \theta} \log(f_W(\cdot)))$ elicits $U = \frac{1}{nI(\theta)}$. Fascinatingly, if wee consider a large number of both observations and predictions under the paradigm $n, p \to \infty$, $n/p \to \delta \in (1, \infty)$, we see that asymptotic variance above is wrong. We instead have the following result:

Theorem
Assuming that ρ is strongly convex ans smooth with X being stanrads guassian design with $\delta > 1$ and F_W has finite second moment. Let (τ_*, b_*) be the fixed point equation to the set of equations

$$\tau^2 = \delta E[\Psi(W + \tau Z; b)^2]$$
$$\frac{1}{\delta} = E[\Psi'(W + \tau Z; B)]$$

Where $\Psi(z; b) = \rho_b'(x)$, $\rho_b(z) = \min_x(b\rho(x) + \frac{1}{2}(x - z)^2)$ and $Z \sim N(0, 1)$

$$\lim_{n,p \to \infty} \frac{1}{p} \sum_{i=1}^{p} V[\hat{\theta}_i] \overset{a.s.}{=} U(\Psi(\cdot, b_*), F_W \star N(0, \tau_*^2))$$

Giving the asymptotic covariance matrix as

$$V = U(\Psi(\cdot, b_*), F_W \star N(0, \tau_*^2)) E[(X^T X)^{-1}]$$

An application of this theorem shows that high-dimensional confidence regions for M-estimators have incorrect coverage when using the classical formula for the covariance matrix. One can begin to see the familiarizes between the SMB result of theorem 1.1 and the above result. Even more striking, the methods for proving both results have many similarities. They both roughly follow the same outline:

1. Introduce an AMP algorithm for the application
2. Find a fixed point for the state evolution equations
3. Analyze a test function (For example MSE) at the limiting iterate
4. Show the test function is the same for AMP and a given estimator
5. (optional) Use the test function to derive other quantities

Overall, I really enjoyed seeing this argument in action for both papers. In the SBM paper, it was delivered in a interesting way and It is very clear how using AMP has broader applications.

As for issues that the paper has, I mainly found two parts of the paper to be sub-optimal.

First, I have an issue with the assumptions made in the paper. Even though dense graphs like this paper assumes do have applications, the most important regime of a fixed average number of vertices is left out. We argued in the Results section how we can force though a large but bounded degree, often our data may just not fit this criterion. I turns out that to analyze this case requires much more complicated machinery. Coja-Oghlan, Krzakala, Perkins, and Zdeborova [6] formalized the use of the cavity method, a well known statistical-physics method, in the context of the SBM. Then, they are able to elicit a form for the mutual information. The connection between these papers is tenuous, [6] did not utilize AMP. There is one connection, however, between these two papers, mainly that the cavity method is commonly associated through belief propagation and the AMP can be derived from belief propagation. Current research investigates this connection though the relation between AMP and the Cavity method, of which there are some interesting results.

Second, I didn't really enjoy the section on showing the mutual information bounds between $X; G$ and $X; Y$, I found it to be overly challenging to understand. This is mainly due to the number of inequalities and approximations which are needed to formalize the result[4]. I have tried to present the major connections that I saw throughout the paper but I would have appreciated if the authors gave a more high-level explanation of why the result holds, then leave all the gritty details in the appendix.

References

[1] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari, *Asymptotic mutual information for the two-groups stochastic block model*, arXiv:1507.08685, 2015.

[2] Adel Javanmard and Andrea Montanari, *State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling*, arXiv:1211.5164, 2012

[3] Dongning Guo, Shlomo Shamai (Shitz), Sergio Verdu, *Mutual Information and Minimum Mean-square Error in Gaussian Channels*, arXiv:cs/0412108, 2004

[4] C. Measson, A. Montanari, T. J. Richardson and R. Urbanke, *The Generalized Area Theorem and Some of its Consequences*, in IEEE Transactions on Information Theory, vol. 55, no. 11, pp. 4793-4821, Nov. 2009, doi: 10.1109/TIT.2009.2030457.

[5] David Donoho, Andrea Montanari, *High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing*, arXiv:1310.7320, 2013

[6] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, Lenka Zdeborova, *Information-theoretic thresholds from the cavity methodg*, arXiv:1611.00814, 2018

[4]The all too well known death by 1000 inequalities