CONNECTING AMP AND LOW DEGREE ESTIMATION: A TUTORIAL

MAX LOVIG

We Will Learn: In many statistical problems a common definition for hardness is when all low-degree polynomial estimators fail. The threshold for which this phenomena occurs is coupled with the success of AMP, a well known algorithm. This means that there is a unification of hardness as AMP has been related to many other algorithms performances (such as first order methods). We present a introduction[a] to AMP / Low-degree estimation and then give a walk-though of how this equivalence is shown under the spiked Wigner model. This result is due to [Montanari and Wein, 2022].

[a]Sorry about the length of this project, feel free to skip to section 2 if you are familiar with AMP. I wanted to also write this as a reference I can hand out if people ask me about how AMP works and some applications of it.

1. Introduction to Low degree estimation and Approximate Message Passing

1.1. Introduction to the Model and Bounded Degree Estimation.

We Will Learn: In many statistical problem there is a question of possibility of estimation and feasibility of estimation. We look towards the latter question, where we say a problem is feasible if a bounded degree estimator can achieve a non trivial MSE. The analysis take place under a simplified "PCA" model, also know as spiked Wigner. We touch on low degree estimation and why we care about it.

Consider the following model, know as the rank 1 estimation problem, we generate Y in the following way:

$$Y = \frac{1}{\sqrt{n}}\theta\theta^T + Z \tag{$\star$}$$

Where $\theta \in \mathbb{R}^n$ and $Z \sim GOE(n)$, where we define the a Gaussian orthogonal ensemble (GOE) as a symmetric matrix Z with $Z_{ij} \sim N(0,1)$ and $Z_{ii} \sim N(0,2)$. This model analyzes a simplifies "PCA" type model where there is a hidden rank one spike amongst noise. Many different results has converged on a characterization of the following type given some restriction to the possible class of θ, $\mathcal{S}$:

If $\theta \in \mathcal{S}$ then the estimation problem is (Impossible / Hard / Easy)

Date: November 2023.

Where impossible refers to an information theoretic threshold in which no estimator succeeds, hard refers to the success of some estimator but the failure within a (usually polynomial time) restricted class, and easy refers to success within a restricted class. Here we define success in the following manner: Let $\theta_i \sim \pi_\theta$ we want to find an estimator $\hat{\theta}$ which has

$$E^{\theta,Z}[\frac{1}{n}||\theta - \hat{\theta}||^2] < V[\theta]$$

and as such will give a more favorable MSE than the trivial estimator $E[\theta]$.

A common choice of restricted class is bounded degree polynomials. There motivation is to represent algorithms which can be evaluated in polynomial time. In this class we consider a maximum degree D and every estimator is a maximum degree D polynomial with respect to the matrix Y, for example some degree 3 polynomials are

$$Y_{1,2}^3 \qquad Y_{2,3}^2 Y_{5,6} \qquad Y_{1,2}Y_{1,3}Y_{1,4} + Y_{1,3}$$

We denote this class as $L_{\leq D}$. Clearly, when $\hat{\theta} \in L_{\leq D}$ each of these estimators is computable in polynomial time. When all estimators in this class fails then, for this papers purposes, we consider a problem hard. Notice that this is somewhat of a misnomer since the analysis for this paper is for bounded but large D, which is not enough to consider what is a very canonical estimator, the top eigenvector. Never the less in many cases top eigenvector is sub-optimal so it is possible some estimator in $L_{\leq D}$ can beat the spectral estimator. So $L_{\leq D}$ is a relatively reasonable class to being studying the hardness of $(\star)$.

1.2. **Approximate Message Passing.** Now we will describe a class of algorithms known as Approximate Message Passing (AMP). Lets recall our model $(\star)$:

$$Y = \frac{1}{\sqrt{n}}\theta\theta^T + Z$$

Consider an estimator x_t generated by the following iterative procedure:

$$x^{t+1} = \frac{1}{\sqrt{n}} Y f_t(x^t) - b_t f_{t-1}(x^{t-1})$$

Where f_t is a function which acts coordinate-wise on the vector x^t, i.e.

$$f_t(x^t) = (f_t(x_1^t), \cdots, f_t(x_n^t))$$

where f_t may also contain randomness independent from Y. We also define the "Onsager" correction term

$$b_t = \frac{1}{n}\sum_{i=1}^{n} f_t'(x_i^t)$$

at a high level this term exists to cancel out non-vanishing dependencies between the iterates.[1]

Notation

Before we dive into this analysis we define some common notation to help us along the way. Given a set of vectors $u, v \in \mathbb{R}^n$ we define there join empirical distribution as $\nu((u,v))$. We also note $d_2(X,Y)$ as the Wasserstein 2 distance from X to Y. We also define the

[1]Make sure to see later for the more generally defined AMP iterates for matrices.

independent product measure between μ and ν as $\mu \otimes \nu$. We also have 3 random variables which will be used repeatedly in our analysis:

$$\begin{aligned} \Theta &\sim \pi(\theta) \\ G &\sim N(0,1) \\ U &\sim P_U \qquad \text{(The distribution of the initialization)} \end{aligned}$$

1.2.1. *State Evolution.*

We Will Learn: Many high dimensional problems suffer from overt complexity when an algorithm operates on objects with high dimension. Approximate Message Passing is an iterative procedure that admits a simple "single letter" analysis in the limit as the size of the problem grows. Even more conveniently, once we have this single letter analysis, for a given error metric, we can tune the parameters (f_t) in this algorithm to optimize it amongst other AMP estimators. To be more precise, the empirical distribution of each coordinate of x_i^t converges in distribution to a transformation of a Gaussian with a mean and variance defined inductively by the state evolution parameters. This exposition is based off of [Feng et al., 2021]

Under model $(\star)$, we consider prior distribution $\pi(\theta)$, and consider a random variable $\Theta \sim \pi_\theta$. We also have random variable G distributed according to $N(0,1)$ and a distribution of the initialization $x^0 \sim U$. Under suitable assumptions on our AMP algorithm (see [Feng et al., 2021]) we can define the following two parameters of our recursion

$$\begin{aligned} \mu_1 &= E[\Theta f_0(\mu_0\Theta + \sigma_0 U)] \\ \sigma_1^2 &= E[f_0(\mu_0\Theta + \sigma_0 U)^2] \\ \mu_{t+1} &= E[\Theta f_t(\mu_t\Theta + \sigma_t G)] \\ \sigma_{t+1} &= E[f_t(\mu_t\Theta + \sigma_t G)^2] \end{aligned}$$

In a vacuum these may seem like somewhat meaningless results but these equations essentially encode all of the information of our algorithm in the high dimensional limit, before we present the state evolution theorem we make one more generalization of σ_t^2. We define a matrix Σ inductively with $\Sigma_{11} = \sigma_1^2$, and then we define

$$\Sigma_{k,l} = \begin{cases} E[f_0(\mu_0\Theta + \sigma_0 U) f_{k-1}(\mu_{k-1}\Theta + \sigma_{k-1}G_{k-1})] & \text{for } l = 1 \\ E[f_{l-1}(\mu_{l-1}\Theta + \sigma_{l-1}U) f_{k-1}(\mu_{k-1}\Theta + \sigma_{k-1}G_{k-1})] & \text{for } l \in 2, \cdots, k \end{cases}$$

Letting $(\sigma_1 G_1, \cdots, \sigma_{k-1}G_{k-1}) \sim N(0, \Sigma_{[k-1],[k-1]})$. We are now in a position to give the state evolution theorem.

Theorem, State Evolution

In words this theorem tells us that the limiting empirical distribution of the t-th iterate of AMP is $N(\mu_t, \sigma_t^2)$. Codified into notation we have under suitable assumptions[2]

$$d_2(\nu(x^0, x^1, \cdots, x^t, \theta), (N(\mu, \Sigma_{[t],[t]}) \otimes \theta) \overset{a.s.}{\to} 0$$

[2]Note that these statements may look slightly more complicated that those in [Feng et al., 2021], I am converting all of these statements to the single step recursion we designed in section 1.2, not the more standard 2 step recursion.

This is also seen in the literature as for any pseudo-Lipshitz function ψ

$$\lim_{n\to\infty} |\frac{1}{n}\sum_{i=1}^{n} \psi(x_i^0, \cdots, x_i^t, \theta) - E^{\Theta,N}[\psi(N(\mu_{[t]}\Theta, \Sigma_{[t],[t]}), \Theta)]| \to 0$$

Perhaps a few remarks are in order. First, we can see that can be easily interpreted as a algorithmic limit theorem for any AMP algorithm, in asymtopia the coordinate of each iterate have a simple limiting distribution defined by the state evolution. Second, with $\hat{\theta}_i^t = f_t(x_i^t)$, consider $\psi = (\hat{\theta}_i^t - \theta)^2$. Then we get for free the limiting statement,

$$MSE(\hat{\theta}^t, \theta) = \frac{1}{n}||x^t - \theta||^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i^t - \theta)^2 \to E[(\Theta - f_t(\mu_t\Theta + \sigma_t G))^2]$$

Analysing this limiting MSE we see that

$$\begin{aligned} E[(\Theta - f_t(\mu_t\Theta + \sigma_t G_t))^2] &= E[\Theta^2] - 2E[\Theta f_t(\mu_t\Theta + \sigma_t G_t)] + E[f_t(\mu_t\Theta + \sigma_t G_t)^2] \\ &= E[\Theta]^2 - 2\mu_{t+1} + \sigma_{t+1}^2 \end{aligned}$$

Meaning, due to the simple limit form of the MSE from an AMP algorithm we should choose function f such that μ_{t+1} is large and σ_{t+1} is small. In classical statistical analysis it is known that the expectation of the posterior minimizes MSE, AMP beautifully replicates this result, allowing us to have a known best choice of f_t.

Theorem, Optimally of Bayes AMP

We define the Bayes AMP as an AMP algorithm with iterates $x_B^1, \cdots, x_B^t$ with the following choices for $f_1, \cdots, f_t$:

$$f_t(x) = E[\Theta | \mu_k\Theta + \sigma_{k+1}G_{k+1} = x]$$

Now consider any AMP algorithm $x^1, \cdots, x^t$ with any functions $g_1, \cdots, g_t$. We have

$$MSE(x^t, \theta) \geq MSE(x_B^t, \theta)$$

Even more convenient to our cause is that it is known when, for model $(\star)$, the exact characterization of when x_B^t converges to the Bayes estimator. The exact formulae are not of direct importance to this analysis but for applications this threshold can be calculates relatively easily.

Besides optimally the Bayes AMP also has a rather simple recursion, we can see that

$$\begin{aligned} \mu_t &= E[\Theta E[\Theta | q_t\theta + \sqrt{q_t}G]] \\ &= E[\Theta E[\Theta|D]] && \text{(For ease of notation we let } D \equiv q_t\theta + \sqrt{q_t}G) \\ &= E^D E[\theta E[\Theta|D]|D] \\ &= E^D[E[\Theta|D]^2] \\ &= \sigma_{t+1}^2 \end{aligned}$$

Meaning that we can define $q_t(\pi_\Theta) = \mu_t = \sigma_t^2$, to unify our two state evolution parameters. We also define the limit of this algorithm with $q_{AMP}(\pi_\Theta) = \lim_{t\to\infty} q_t(\pi_\Theta)$. Thus, combining this with our earlier MSE calculation

$$MSE(x^t, \theta) \geq MSE(x_B^t, \theta) = E[\Theta^2] - q_{AMP}(\pi_\Theta)$$

1.2.2. *Combinatorial Analysis of AMP.*

We Will Learn: Common assumptions of AMP is that the non-linearity function f_t is Lipshitz, for later proof techniques one needs to consider f_t which are polynomials. It turns out that this extension is not only natural but also invokes many of the choices used in the equivalence proof. The Analysis shown below is from [Bayati et al., 2015] and involves unrolling the AMP algorithm, showing that the Onsager correction corresponds to removing a set of "non-backtracking" configurations, then showing that under the remaining configurations the limiting evaluation of the AMP algorithm is equivalent if the matrix is replaces with a new independent GOE matrix at each step of the algorithm. Under this asymptotically equivalent model the state evolution equation is immediate.

The state evolution results of the previous section are very interesting and useful in a wide variety of estimation tasks. To add another wrinkle to this already rich story, it also turns out there are multiple distinct approaches to show the state evolution theorem in some way. Classically the proof for such a theorem used a conditional distribution lemma dating back to [Bolthausen, 2012]. There have been information theoretic and statistical physics methods applied to proving/heuristically deriving such a statement. Below I will present a slightly different approach to proving the State Evolution theorem. This method was utilized in [Bayati et al., 2015] and still has common applications for analysing AMP algorithms related to polynomials.

First we need just analyze this model in lieu of $\star$.

$$Y = \frac{1}{\sqrt{N}} W$$

Where W is a GOE matrix In this case our algorithm becomes

$$x^{t+1} = \frac{1}{\sqrt{n}} W f_t(x^t) - b_t f_{t-1}(x^{t-1})$$

This reduction is accomplished by showing the state evolution results for the above iteration and then reducing the results from $(\star)$ to this model. This is accomplished by defining $u^k = x^k - \mu_k \theta$ and showing the iterates $\nu(u^k, \theta)$ converge to the distribution $(\sigma_t G_t, \Theta)$ in the large sample limit. We can then simply add back in the distribution of $\mu_k \Theta$ to this result. We also can relax the standard Lipshitz condition on f_t to any function which has finite moments. This reduction will be started here as we will show the state evolution for an arbitrary polynomial f_t and then in the equivalence proof we show a method to approximate the Bayes AMP algorithm. Polynomial approximations for other choices of f_t work similarly. First we approximate f_t with a polynomial up to finite degree and then show that such an approximation induces closeness on the state evolution parameters of the polynomial AMP algorithm and the original AMP algorithm.

In order to understand the AMP algorithm a first step is to establish some results on a slight relaxation. For notational simplicity we absorb the $\frac{1}{\sqrt{n}} W$ to be just W. We consider

a general message passing algorithm according to

$$\begin{aligned} z_{i\to j}^0 &= x_i^0 \\ z_{i\to j}^{t+1} &= \sum_{\ell\in[N]\setminus j} W_{\ell i} f_t(z_{\ell\to i}^t) \\ z_i^{t+1} &= \sum_{\ell\in[N]} W_{\ell i} f_t(z_{\ell\to i}^t) \end{aligned}$$

Where we essentially pass messages between nodes nodes i to j excluding a "reversing" or "backtracking" term corresponding to the message passed from j to i in the previous step.

Notice that the state evolution results would be immediate if we had a related message passing algorithm y with

$$\begin{aligned} y_{i\to j}^0 &= x_i^0 \\ y_{i\to j}^{t+1} &= \sum_{\ell\in[N]\setminus j} W_{\ell i}^t f_{t-1}(y_{\ell\to i}^t) \\ y_i^{t+1} &= \sum_{\ell\in[N]} W_{\ell i}^t f_t(y_{\ell\to i}^t) \end{aligned}$$

Where each W^t is a GOE matrix independent from other iterates. Immediately we would be able to make simple statements about the state evolution because we need not be concerned with strong correlations between iterates due to them being hit by the same matrix W.

We can then relate AMP algorithms to this message passing algorithm y in the following two results, for $m \in \mathbb{N}$, there exists constants C and K such that

$$|E[(z_i^t)^m] - E[(y_i^t)^m]| \le CN^{-1/2} \tag{1}$$

and

$$|E[(z_i^t)^m] - E[(x_i^t)^m]| \le KN^{-1/2} \tag{2}$$

High Level approach to (1):

To show this first result we have a couple of high-level steps we take, these methods are quite common in AMP analysis so they are worth understanding

(1) Unroll the message passing algorithm
   (a) First, as each f_t is a polynomial we can consider its decomposition into its power terms $f_t(x) = \alpha_t^0 + \alpha_t^1 x + \cdots + \alpha_t^D x^D$.

(b) Then starting at the final t iterate we want to analyze the average of some test function ψ. For simplicity we will consider $\psi(x) = x$, write out its sum as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} z_i^t &= \sum_{\ell\in[N]} W_{\ell i} f_{t-1}(z_{\ell\to i}^{t-1}) \\
&= \sum_{k\in[D]} \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in[N]} W_{\ell i}\alpha_{t-1}^k (z_{\ell\to i}^{t-1})^k \\
&= \sum_{k\in[D]} \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in[N]} W_{\ell i}\alpha_{t}^k (z_{\ell\to i}^{t-1})^k \\
&= \sum_{k\in[D]} \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in[N]} W_{\ell i}\alpha_{t-1}^k \left(\sum_{p\in[N]\setminus j} W_{\ell i} f_{t-2}(z_{p\to\ell}^{t-2})\right)^k \\
&= \sum_{k\in[D]} \alpha_{t-1}^k \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell\in[N]}\sum_{p_1,\cdots,p_k\in[N]\setminus j} W_{\ell i}\prod_{q=1}^{k} W_{p_q i} f_{t-2}(z_{p_q\to\ell}^{t-2})
\end{aligned}$$

(c) notice that this unrolling had two steps, first we picked a power term for f and this dictated the number of new matrices and then we converted the power to a sum over a labelled product of matrices. A canonical structure to represent this is that our choice of degree for f is constructing a node with a $d+1$ degree constraint and then attaching d new edges which were previously unconnected. Recursively doing this we can see that any such message passing algorithm is a linear combination of tree graphs, where for each tree we sum over a labellings on the nodes (constrained to be some non-reversing structure) that induces a label on the edges W. In notation we have with $\mathcal{A}$ representing a finite set of trees and ϕ as all possible labellings for a tree,

$$\frac{1}{n}\sum_{i=1}^{n}\psi(z_i^t) = \sum_{A} c_{\mathcal{A}}\left(\frac{1}{n}\sum_{\phi:[|V(T)|]\to[n]}\prod_{(u,v)\in E(T)} W_{\phi(u),\phi(v)}\right)$$

For example in the above unrolling we have a tree like

Thus we can now focus on each of these trees individually. For a specific tree T, we can pull out the $\frac{1}{\sqrt{n}}$ normalization from each W edge in the tree and

pull out a leading factor of $n^{-(|E(T)|/2+1)}$, seen as:

$$\frac{1}{n}\sum_{\phi:[|V(T)|]\to[n]}\prod_{(u,v)\in E(T)} W_{\phi(u),\phi(v)} = \frac{1}{n}\sum_{\phi:[|V(T)|]\to[n]}\prod_{(u,v)\in E(T)} \frac{1}{\sqrt{n}} W_{\phi(u),\phi(v)}$$
$$= \frac{1}{n^{|E(T)|/2+1}}\sum_{\phi:[|V(T)|]\to[n]}\prod_{(u,v)\in E(T)} W_{\phi(u),\phi(v)}$$

(2) We can consider a partition τ on the node labels which identify blocks of nodes give the same label. We then define a labelling restricted to τ as $\phi|\tau$, where each label u, v which share a block in τ have $\phi(u) = \phi(v)$. This reduces our summation to:

$$\frac{1}{n^{|E(T)|/2+1}}\sum_{\phi:[|V(T)|]\to[n]}\prod_{(u,v)\in E(T)} W_{\phi(u),\phi(v)} = \frac{1}{n^{|E(T)|/2+1}}\sum_{\text{all } \tau}\sum_{\phi|\tau}\prod_{(u,v)\in E(T)} W_{\phi(u),\phi(v)}$$

Our goal is to show for which partitions τ does the summation over $\phi|\tau$ cancel out leading order of $n^{-(|E(T)|/2+1)}$. These non-vanishing partitions correspond to labellings which induces a pairing of the edges W and when taking the quotient graph over nodes of the same label is a tree. For example here are two partitions τ, τ', and their corresponding quotient graphs for a tree. One is a pairing and the other is not,

(3) By the non-backtracking criterion we know that two edges which are paired cannot be adjacent. With all of the non-vanishing labellings having the above characterization, we have that any labelling which paired two W's from different generations (if we were using the y message passing then this means we paired W^s to $W^{s'}$ where $s \neq s'$) will have a cycle after taking the quotient graph and thus this is a vanishing labelling. Meaning that message passing y is the message passing algorithm z in the large n limit.

(4) As we have only pairings an non-vanishing then this algorithm essentially only relies on the second moment of each matrix in the limit and we can replace each generation of W with a independent copy and still retail the limiting behavior. Thus showing that in the limit the two algorithms are the same, teasing out the $N^{-1/2}$ order comes from any vanishing labelling shrinks at this rate.

High Level approach to (2):

This second step involves relating the message passing algorithm z to our AMP algorithm x. At a high level this simply comes from expanding out our AMP iteration to the following

form

$$\begin{aligned} x_i^{t+1} &= \sum_{j\in[N]} W_{ij} f_t(x_j^t) - \sum_{j\in[N]} (W_{ij})^2 f_t'(x_j^t) f_{t-1}(x_i^{t-1}) \\ &= \sum_{j\in[N]} W_{ij} f_t(x_j^t) - \sum_{j\in[N]} W_{ji} W_{ij} f_t'(x_j^t) f_{t-1}(x_i^{t-1}) \end{aligned}$$

Where the 2nd term in the right hand side is based off the Onsager correction. By unrolling both terms on the right hand side we see that the Onsager correction will cancel out non-vanishing labellings which do not respect the non-reversing structure of the message passing algorithm z. Thus in the high dimensional limit they are equivalent.

Proving the State Evolution

All that is left is to show the state evolution results for the y message passing algorithm. We will show the two statements inductively on t:

$$\begin{aligned} \lim_{n\to\infty} E[(y_{i\to j}^t)^m] &= E[(Z^t)^m] \\ \frac{1}{n}\sum_{i=1}^n (y_{i\to j}^t)^m &\xrightarrow{p} E[(Z^t)^m] \end{aligned}$$

Where $Z^t \sim N(0,\sigma_t^2)$. The base case for both statements is satisfied by assumption on the limiting properties of the initialization. For $t \geq 1$ we define $\mathfrak{F}_t$ to be the sigma algebra generated by the prior iterates. We have,

$$\begin{aligned} \lim_{n\to\infty} E[y_{i\to j}^{t+1}|\mathfrak{F}_t] &= \lim \sum_{\ell\in[N]\setminus j} E[W_{\ell i}^t] f_{t-1}(y_{\ell\to i}^t) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \lim_{n\to\infty} E[(y_{i\to j}^{t+1})^2|\mathfrak{F}_t] &= \lim E[(\sum_{\ell\in[N]\setminus j} W_{\ell i}^t f_{t-1}(y_{\ell\to i}^t))^2] \\ &= \lim \sum_{\ell\in[N]\setminus j} \sum_{k\in[N]\setminus j} E[W_{\ell i}^t W_{ki}^t] f_{t-1}(y_{\ell\to i}^t) f_{t-1}(y_{k\to i}^t) \\ &= \lim \sum_{\ell\in[N]\setminus j} E[(W_{\ell i}^t)^2] f_{t-1}(y_{\ell\to i}^t)^2 \\ &= \lim \frac{1}{n} \sum_{\ell\in[N]\setminus j} f_{t-1}(y_{\ell\to i}^t)^2 \\ &= E[f_{t-1}(Z^t)^2] \qquad \text{(By the induction hypothesis)} \\ &= \sigma_{t+1}^2 \end{aligned}$$

As each $y_{i\to j}^{t+1}$ when conditioned on $\mathfrak{F}_t$ is a sum of gaussian random variable, then it is guassian distributed and thus determined by its mean and variance, allowing us to conclude that

$$\lim_{n\to\infty} E[(y_{i\to j}^t)^m] = E[N(0,\sigma_t^2)^m]$$

As for the limiting statement we need to show that

$$|E[(y_{i\to j}^{t+1})^m (y_{k\to j}^{t+1})^m] - E[(y_{i\to j}^{t+1})^m]E[(y_{i\to j}^{t+1})^m]| \to 0$$

Once this is true then we have

$$\begin{aligned}
&\lim_{n\to\infty} V[\frac{1}{n}\sum_{i=1}^{n}(y_{i\to j}^{t+1})^m] \\
&\leq \lim \frac{1}{n^2}\sum_{i,k} |E[(y_{i\to j}^{t+1})^m (y_{k\to j}^{t+1})^m] - E[(y_{i\to j}^{t+1})^m]E[(y_{i\to j}^{t+1})^m]| \\
&\to 0
\end{aligned}$$

Using Chebychev's inequality and our expectation results we have

$$\frac{1}{n}\sum_{i=1}^{n}(y_{i\to j}^{t})^m \xrightarrow{p} E[(Z^t)^m]$$

Thus, all we are left to show is:

$$|E[(y_{i\to j}^{t+1})^m (y_{k\to j}^{t+1})^m] - E[(y_{i\to j}^{t+1})^m]E[(y_{i\to j}^{t+1})^m]| \to 0$$

By unrolling each of these algorithms we can represent these terms as summations over partitions on m trees rooted at i and k. We think of the quotient graph G of these collection of trees together.

We can then begin to consider the possibilities for G. We have the following:

(1) if G can be disconnected into sub-graphs containing k and i, G_k and G_i. This means that terms within the expectation are independent and cancel out.
(2) G is a vanishing labelling and then both terms have 0 impact.
(3) G is a connected tree, notice that we have fixed two indices so there are $N^{|V(G)|-2}$ choices (we need labels i and k). It can be shown that each of these graph have expected product of size $N^{-(|V(G)|-1)}$, and thus they are vanishing as well.

Simple Universality Statements

Since we had to show such a strong statement about which labellings are non-vanishing. We also get, for free, that this algorithm is universal so long as we replace W with any other matrix so long as it is symmetric with element-wise mean 0 and variance 1 (i.e. a Wigner Matrix). In fact this has a "free independence" interpretation but this is unneeded for our analysis, so we leave this remark here.

Extension to Matrix Iterates

For ease of understanding, we have been restricting to the simple case of a single AMP iterate. We can extend our results to a d simultaneous AMP iteration of the following form:

$$\mathbf{x}^{t+1} = \frac{1}{\sqrt{n}} Y F_t(\mathbf{x}^t) - F_{t-1}(\mathbf{x}^{t-1})\mathbf{B}_t^T$$

Where $\mathbf{x} \in \mathbb{R}^{n \times d}$ and $F_t : \mathbb{R}^d \to \mathbb{R}^d$ is applied row-wise to our iterates. We define $\mathbf{B}_t$ as,

$$\mathbf{B}_t = E[DF_t(\mu_t \Theta + G_t)]$$

Where $DF_t = \{ \partial_i F_{t,j} \}_{i,j \in [d]}$. Now instead of tracking two state evolution parameters we track a vector and matrix $\boldsymbol{\mu_t}$ and $\boldsymbol{\Sigma_t}$, defined iteratively as:

$$\begin{aligned}(\boldsymbol{\mu_t}, \boldsymbol{\Sigma_t}) &= (0,0)\\ (\Theta, G_t) &\sim \pi_\Theta \otimes N(0, \boldsymbol{\Sigma_t})\\ \boldsymbol{\mu_{t+1}} &= E[\Theta F_t(\boldsymbol{\mu_t} + G_t)]\\ \boldsymbol{\Sigma_{t+1}} &= E[F_t(\boldsymbol{\mu_t} + G_t)F_t(\boldsymbol{\mu_t} + G_t)^T]\end{aligned}$$

We can also have a post processing function $g_t(\boldsymbol{x}^t)$ (applied row-wise).

This AMP algorithm can be proved using the tools from above and will be important to prove our equivalence theorems.

2. The Bayes AMP algorithm can be arbitrarily approximated by any polynomial

We Will Learn: We will consider the Bayes AMP algorithm and justify a polynomial approximation of the non-linearity $f_t = E[\theta | \sqrt{q_t} G + q_t \theta]$ as this function has finite expectation and the space of polynomials is dense under Gaussian measure. We then show that a suitable polynomial approximation induces a suitable approximation of the state evolution parameters. Under this state evolution approximation we can get arbitrarily close to the error of the Bayes AMP after a fixed number of iterations, sat t. Such an approximation is a t way composition of polynomial which will have bounded degree if each approximation of f_t is of bounded degree.

Our first goal will to be to show the following upper bound on the error of the best D degree polynomial.

Theorem, Bayes AMP can be approximated by a polynomial estimator

Assume[3] that π_Θ is independent of n and $E[\Theta] \neq 0$. For any $\epsilon > 0$, there exists an degree $D(\epsilon)$ estimator $\hat{\theta}_D$ such that

$$\lim_{n \to \infty} \frac{1}{n} E[||\hat{\theta}_D - \theta||^2] \leq E[\theta^2] - q_{AMP} + \epsilon$$

Alternatively said, any Bayes AMP algorithm run for a bounded number of iterations can approximated arbitrarily closely by a polynomial of bounded degree.

Proof. First we establish that a sufficient condition for this statement is that there exists an AMP algorithm with bounded degree polynomial non-linearities f_t with state evolution

[3]This assumption on the expectation can be removed if we change the error metric to be invariant under sign change or if one can extend this proof to allow for spectral initialization (see future directions)

parameters

$$|\mu_t - q_t| \leq \varepsilon' \quad |\sigma_t^2 - q_t| \leq \varepsilon'$$

Consider t^* such that for the Bayes AMP iterations q_t it is that largest choice of t such that $q_{t^*} \geq q_{amp} - \frac{\varepsilon}{2}$. We are then provided the following statement by state evolution:

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} E\left[||\hat{\theta}^{t_*}(Y) - \theta||^2\right] &= E\left[\left(\theta - f_{t_*}\left(\mu_{t_*}\theta + \sigma_{t_*}G\right)\right)^2\right] \\
&= E\left[\theta^2\right] - 2\mu_{t_*+1} + \sigma_{t_*+1}^2 \\
&\leq E\left[\theta^2\right] - 2q_{t_*+1} + 2\epsilon_0 + q_{t_*+1} + 2\epsilon_0 \\
&\leq E\left[\theta^2\right] - q_{t_*+1} + 3\epsilon_0 \\
&\leq E\left[\theta^2\right] - q_{AMP} + \epsilon/2 + 3\epsilon_0
\end{aligned}$$

Which satisfies our claim when $\epsilon_0 < \epsilon/8$. Such an algorithm has degree $\prod_{i=1}^{t_*}(D_i + 1)$ which is bounded.

Now we are left to show that such a sufficient condition is possible. First we analyze $|\mu_t - q_t|$. We strive for a proof by induction on t, of which the base case is satisfied by definition.

To codify this induction argument into notation we define a sequence $\epsilon_{t,k} \downarrow 0$ indexed by k. For $s \leq t$ we define the function $f_s = f_{k,s}$ (a polynomial) such that we can satisfy the induction hypothesis at the t step with $\epsilon_0 = \epsilon_{t,k}$. Or, stated equivalently, we can arbitrarily approximate iterates up to t with a finite polynomial. Denoting the bayes non-linearity as f_t^B, we then have that

$$\begin{aligned}
|\mu_{k,t+1} - q_{t+1}| &= |E\left[\theta\left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(q_t\theta + \sqrt{q_t}G\right)\right)\right]| \\
&\leq |E\left[\theta\left(f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(q_t\theta + \sqrt{q_t}G\right)\right)\right]| \\
&\quad + |E\left[\theta\left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)\right]| \\
&= B_1(k) + B_2(k)
\end{aligned}$$

To analyze these terms further we need the following convenient facts[4].

(1) $f_t^B(x) = E[\theta | q_t\theta + \sqrt{q_t}G = x]$ is continuous
(2) $|f_t^B(x)| \leq C(1 + |x|)$
(3) there exists a polynomial f_t such that

$$\sup_u |f_t(u) - f_t^B(u)| e^{\frac{-u^2}{4\tau^2}} \leq \frac{\varepsilon_0}{4E[\theta^2]^{1/2}}$$

Utilizing these three facts we can continue our analysis.

[4]As these don't have to do with AMP or really any high-dimensional methods, I chose to omit the proof of these results and we just take then for granted here.

By the induction hypothesis $\mu_{k,t}\theta + \sigma_{k,t}G \overset{a.s.}{\to} q_t\theta + \sqrt{q_t}G$ as we let $k \to \infty$. Thus we have

$$\begin{aligned}
\lim_{k\to\infty} B_1(k) &= \lim_{k\to 0} |E\left[\theta\left(f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(q_t\theta + \sqrt{q_t}G\right)\right)\right]| \\
&= |\lim_{k\to\infty} E\left[\theta\left(f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(q_t\theta + \sqrt{q_t}G\right)\right)\right]| \quad \text{(Continuity of } |\cdot|\text{)} \\
&= |E\left[\lim_{k\to\infty} \theta\left(f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(q_t\theta + \sqrt{q_t}G\right)\right)\right]| \\
&\quad (f^B(x) \le C_t(1+|x|) \text{ which has finite expectation, allowing us to interchange limit and expectation}) \\
&= |0| \\
&\quad (\text{As } f^B \text{ is continuous, the inputs to } f^B \text{ converge almost surely implies the difference goes to 0})
\end{aligned}$$

Meaning we can choose a k' such that $\forall k \ge k'$, $B_1(k) \le \frac{\epsilon_0}{2}$. Now to analyze $B_2(k)$, without loss of generality assume that we have chosen k large enough such that $\max(\mu_{k,t}, \sigma_{k,t}^2) \le 2q_t$. As we assumed that $\pi(\theta)$ is sub-Gaussian with parameter τ. We first make the observation that $Z_{k,t} = \mu_{k,t}\theta + \sigma_{k,t}G$ is sub-Gaussian with parameter

$$\mu_{k,t}^2\tau^2 + \sigma_{k,t}^2 \le 4(q_t^2\tau^2 + q_t) \coloneqq \tau_{t,k}^2$$

Using Cauchy Schwartz we can bound $B_2(k)$ with

$$B_2(k) \le E[\theta^2]^{1/2} E\left[\left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right) - f_t^B\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2\right]^{1/2}$$

Now using out third convenient fact we can choose f_t such that

$$\sup_u |f_t(u) - f_t^B(u)| e^{\frac{-u^2}{4\tau^2}} \le \frac{\varepsilon_0}{4E[\theta^2]^{1/2}}$$

or equivalently, with $\tau = \tau_{t,k}$,

$$\sup_u |f_t(u) - f_t^B(u)| \le \frac{e^{\frac{u^2}{4\tau^2}}\varepsilon_0}{4E[\theta^2]^{1/2}}$$

by expanding the expectation this means

$$\left[\int \left(f_{k,t}\left(Z_{k,t}\right) - f_t^B\left(Z_{k,t}\right)\right)^2 \partial F_{Z_{k,t}}\right]^{1/2} \le \left[\int \left(\frac{e^{\frac{z_{k,t}^2}{4\tau_{t,k}^2}}\varepsilon_0}{4E[\theta^2]^{1/2}}\right)^2 \partial F_{Z_{k,t}}\right]^{1/2} = \frac{\epsilon_0}{4E[\theta]^{1/2}} E\left[e^{\frac{z_{k,t}^2}{2\tau_{t,k}^2}}\right]^{1/2}$$

Thus,

$$\begin{aligned}
B_2(k) &\le \frac{\epsilon_0 E[\theta^2]^{1/2}}{4E[\theta]^{1/2}} E\left[e^{\frac{z_{k,t}^2}{2\tau_{t,k}^2}}\right]^{1/2} \\
&\le \frac{\epsilon_0}{4} E\left[e^{\frac{z_{k,t}^2}{\tau_{t,k}^2}}\right]^{1/4} && \text{(By Jensen's Inequality)} \\
&\le \frac{\epsilon_0}{4} 2^{1/4} && \text{(By Sub-Guassianity)} \\
&\le \frac{\epsilon_0}{2}
\end{aligned}$$

We can continue the analysis on σ^2 with

$$\begin{aligned}|\sigma^2_{k,t+1} - q_{t+1}| &= |E\left[\left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2 - \left(f^B_t\left(q_t\theta + \sqrt{q_t}G\right)\right)^2\right]| \\ &\leq |E\left[\left(f^B_t\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2 - \left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2\right]| \\ &\quad + |E\left[\left(f^B_t\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2 - \left(f^B_t\left(q_t\theta + \sqrt{q_t}G\right)\right)^2\right]| \\ &= A_1(k) + A_2(k)\end{aligned}$$

Thus we are left to bound A_1 and A_2. The analysis of A_1 is equivalent to B_1. We can still exchange limit and expectation as we have $f^B(x)^2 \leq (1+|x|)^2$, giving us finite expectation. We also have $\mu_{k,t}\theta + \sigma_{k,t}G \overset{a.s.}{\to} q_t\theta + \sqrt{q_t}G$ implying that the difference (by the continuity of f^B and $\cdot^2$),

$$\left(f^B_t\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2 - \left(f_{k,t}\left(\mu_{k,t}\theta + \sigma_{k,t}G\right)\right)^2 \overset{a.s.}{\to} 0$$

Thus bounding A_1. For bounding A_2 we need a new version of (3) with the square $(f^B)^2$, it shouldn't be surprising such a method exists. Thus we can bound A_2 the same as B_2. Thus we conclude the proof. ☭

3. No polynomial attains a minimum MSE lower than the Bayes AMP algorithm

Now we are left to show the corresponding upper bound,

Theorem, No polynomial estimator outperforms AMP Under the same assumptions on Θ as our previous theorem, for any constant D we have

$$\lim_{n\to\infty} \inf_{\hat{\theta}\in L_{\leq D}} \frac{1}{n} E[||\hat{\theta} - \theta||^2] E[\theta^2] - q_{AMP}$$

The interpretation here is pretty straightforward, any low degree estimator for some large but bounded degree D will never outperform AMP. Even more so, we will show that any of these low degree estimators can be reduce to a message passing algorithm which is known sub-optimal to the Bayes AMP. Combining this result with the one in the previous section we have established that the threshold for success in the low degree regime is equivalent to the success of Bayes AMP.

Proof. See the Remainder of this project ☭

3.1. The Best polynomials in $(\star)$ are tree structured.

We Will Learn: The goal of this section is to show that there exists a polynomial that is "tree structured" that achieves the optimal MSE amongst all polynomial estimators. Clearly if such a statement were to be true than AMP seems like a very canonical method to mimic polynomial estimators. This result is proved by constructing the polynomial as a solution to a linear equation of orthonormal coefficients. Hopefully you will notice many similarities with the low-degree estimation lecture

Lets consider a basis for tree polynomials that we hope can be a good estimator. Let $\mathcal{T}_{\leq D}$ be the set of rooted trees with at most D edges (up to some root preserving isomorphism). When we refer to the root $\circ$, we mean a specially designated node that is in any tree $T \in \mathcal{T}_{\leq D}$. We also define a *labelling rooted at 1* on a $T \in \mathcal{T}_{\leq D}$ as a function $\phi : V(T) \to [n]$ as a labelling of each node in the tree T constrained to $\phi(\circ) = 1$.

We say a labelling ϕ is *non-reversing* (also notated as $y \in nr(T)$) if for each distinct vertices $u, v \in T$ with the same have either

(1) $d(u, v) \geq 2$
(2) $d(u, v) = 2$ and u, v are children of a common parent

For example we have:

We can then define for each $T \in \mathcal{T}_{\leq D}$ a corresponding polynomial which is the evaluation of the tree graph sum over the set of non-reversing labels, meaning

$$\mathcal{F}_T(Y) = \frac{1}{\sqrt{|nr(T)|}} \sum_{\phi \in nr(T)} \prod_{(u,v) \in E(T)} Y_{\phi(u),\phi(v)}$$

This allows us to construct the space $R[Y]^{\mathcal{T}}_{\leq D}$ as the set of all polynomials p such that

$$p(Y) = \sum_{T \in \mathcal{T}_{\leq D}} \hat{p}_T \mathcal{F}_T(Y)$$

for any $\hat{p}_T \in \mathbb{R}$. Hopefully it is clear why this basis of polynomials is favorable to an AMP analysis.

Of course there are some preliminary comments to make. First we can reduce our analysis to just finding a good estimator for, say θ_1. By symmetry of the MSE, any estimator that

achieves the minimum MSE for estimating θ_1 will be able to achieved the minimum MSE is estimating the vector θ. We can begin our proof with the goal of showing that the best estimator for θ_1, in terms of minimizing MSE, is tree structured.

If we could show the following result then we would be very close to showing our low degree - AMP equivalence. Stated formally, for any π_θ, ψ, D we want to find $p \in R[Y]^{\mathcal{T}}_{\leq D}$ such that

Theorem, The best low degree polynomial is tree structured

$$\lim_{n\to\infty} \inf_{g \in L_{\leq D}} E[(q(Y) - \psi(\theta_1))^2] = \lim_{n\to\infty} E[(p(Y) - \psi(\theta_1))^2]$$

The reason for rooted labels at 1 is that we want to estimate $\psi(\theta_1)$ and we evaluate tree polynomials in reference to the final output being generated by the root.

We can use a more simplified model than $(\star)$, this new model will replace Z with $\tilde{Z}$ where $\tilde{Z}_{i,j} \sim Z_{ij}$ and $\tilde{Z}_{i,i} \sim N(0,1)$. This is justified on the left side as any tree structured polynomial doesn't use the diagonal since we are reducing the amount of noise in the model we expect the best low degree polynomial to be better with noise $\tilde{Z}$ than the best low degree polynomial with noise Z.

Lemma, $\leq$ Direction

We can immediately conclude that

$$\lim_{n\to\infty} \inf_{g \in L_{\leq D}} E[(q(Y) - \psi(\theta_1))^2] \leq \lim_{n\to\infty} E[(p(Y) - \psi(\theta_1))^2]$$

Proof. As we have that any polynomials $p \in R[Y]^{\mathcal{T}}_{\leq D}$ is a degree D polynomial. Unfortunately the other direction is much harder to grasp, in fact the proof technique shown here is relatively unique in the sense that it constructs a much easier to use basis then $R[Y]^{\mathcal{T}}_{\leq D}$ and then shows a change of basis argument that allows a polynomial in this simpler basis to be written as a polynomial in $R[Y]^{\mathcal{T}}_{\leq D}$. ∎

Constructing this simpler Basis

We Will Learn: We will define a slightly more general setting for our basis, that of distinct labellings over multi-graphs (known as embeddings). This basis is conveniently[a] invariant under permutations of the rows and columns of Y (excluding 1). Utilizing an extension of the Hunt-Stein theorem we have that the minimizing MSE polynomial must be spanned by this basis. For simplicity I will assume some background on the Hermite polynomials

[a] Perhaps purposefully is a better word here

Unfortunately the class of non-reversing polynomials is somewhat awkward for this analysis so we will need to construct a surrogate basis and then have a clean-up step to convert functions in this surrogate basis to the original non-reversing basis. This new basis will be based on the class of multi-graphs with $\leq D$ edges, denoted $\mathcal{G}_{\leq D}$ (in order to maintain the $\leq D$ degree polynomial constraint) and is summed over the class of embeddings. We say that $\phi : V(T) \to \mathbb{N}$ is an *embedding* for T if each distinct node gets a distinct label and we have $\phi(\circ) = 1$.

It is immediate that this is a super-set of our non-reversing basis over tree graphs as every non-reversing labelling ϕ. We can see this as any labelling ϕ can induce a multi-graph G by taking the quotient graph over all nodes of the same label. In fact for any ϕ we have its image $\alpha = \mathbf{im}(\phi, G)$ which consist of a vector in $\mathbb{N}^{\binom{n}{2}}$ where the α_{ij} (with $i < j$ counts the multiplicity of edges in quotient graph of G induced by matching nodes based on ϕ between nodes i and j.

For a given α, we can consider α as an edge code-book where the i, j-th element refers to the number of edges from i to j. In this case we can consider the set of each connected component of α, denoted $C(\alpha)$. Notice that each sub-graph $\Gamma \in C(\alpha)$ can be viewed as a vector or graph, which ever is convenient. For each α we will define the centered graph Hermite polynomial

$$\mathcal{H}_\alpha = \prod_{\Gamma \in C(\alpha)} (h_\Gamma(Y) - E[h_\Gamma(Y)])$$

Where we have

$$h_\Gamma(Y) = \prod_{(\gamma_1,\gamma_2) \in E(\Gamma)} h_{\alpha_\gamma}(Y_{\gamma_1,\gamma_2})$$

where h_k is the k-th Hermite polynomial and the expectation is over Y in $(\star)$. We can now sum over all of our embeddings to have

$$\mathcal{H}_G(Y) = \frac{1}{\sqrt{|emb(G)|}} \sum_{\phi \in emb(G)} \mathcal{H}_{\mathbf{im}(\phi,G)}(Y)$$

This polynomial has the symmetrization property where it is the linear subspace that is invariant under swapping of non-one rows/columns of Y. This can be seen as for a given embedding ϕ the swapping of a rows or columns $i, j \neq 1$ in Y is equivalent an embedding where each coordinate in reference to the label i and j are swapped. As we sum over all embedding, we maintain this symmetry property.

The Hunt-Stein theorem tells us that the best estimator for this task of estimating a function of θ_1 is invariant under the permutation of the non-one rows and columns of Y. Of course this is in terms of all estimators, not just low degree but this paper provides the extension to low-degree estimators (a result which is of independent interest). Meaning that if we define the space:

$$\mathbb{R}[Y]^{sym}_{\leq D} = \{f(Y) = \sum_{G \in \mathcal{G}_{\leq D}} a_G \mathcal{H}_G(Y)\}$$

Then we have that

$$\inf_{q \in L_{\leq D}} E[(q(Y) - \phi(\theta_1))^2] = \inf_{q \in \mathbb{R}[Y]^{sym}_{\leq D}} E[(q(Y) - \phi(\theta_1))^2]$$

Showing that the optimal polynomial in this new basis is tree structure

We Will Learn: Under this new basis we can construct explicitly the minimizing MSE polynomial. We then have a magical property which drive the ability to prove this statement: Any coefficient for a basis element that is not a tree is vanishing and thus the best polynomial is tree structured in this new basis.

Now that we have this simplified space, we can define two vectors $c_n \in \mathbb{R}^{|\mathcal{G}_{\leq D}|}$ and $M_n \in \mathbb{R}^{|\mathcal{G}_{\leq D}| \times |\mathcal{G}_{\leq D}|}$ where

$$c_{n,A} = E[\mathcal{H}_A(Y)\psi(\theta_1)] \qquad M_{n,AB} = E[\mathcal{H}_A(Y)\mathcal{H}_B(Y)]$$

Note that these are essentially the projections of ψ onto the graph polynomial $\mathcal{H}_A$ and then a matrix tracking the dependencies between two elements of our basis.

We will state the following fact without proof[5], for some $C > 0$:

$$||M_n^{-1}||_{op} > C$$

Thus we are now able to solve the following equation[6]

$$\inf_{q \in \mathbb{R}[Y]^{sym}_{\leq D}} E[(q(Y) - \psi(\theta_1))^2] = \inf_{q \in \mathbb{R}[Y]^{sym}_{\leq D}} E[\psi(\theta_1)^2] + E[q(Y)^2] - 2E[q(Y)\psi(\theta_1)]$$

The solution of which is equivalent to

$$\begin{aligned}
&\arg\min_{\hat{q}} E[(\sum_A \hat{q}_A \mathcal{H}_A)(\sum_B \hat{q}_B \mathcal{H}_B)] - 2E[\sum_A \hat{q}_A \mathcal{H}_A(Y)\psi(\theta_1)] \\
&= \arg\min_{\hat{q}} \sum_{A,B} \hat{q}_A \hat{q}_B E[\mathcal{H}_A \mathcal{H}_B] - 2\sum_A \hat{q}_A E[\mathcal{H}_A(Y)\psi(\theta_1)] \\
&= \arg\min_{\hat{q}} \sum_{A,B} \hat{q}_A \hat{q}_B M_{AB} - 2\sum_A \hat{q}_A c_A \\
&= \arg\min_{\hat{q}} \hat{q}^T M \hat{q} - 2\hat{q}^T c
\end{aligned}$$

Taking the derivative wrt to $\hat{q}$ give us the minimzer as M is positive definite. With

$$\frac{\partial}{\partial \hat{q}} = 2M\hat{q} - 2c = 0 \implies \hat{q} = M^{-1}c$$

And thus our minimizing value is

$$\inf_{q \in \mathbb{R}[Y]^{sym}_{\leq D}} E[(q(Y) - \psi(\theta_1))^2] = E[\psi(\theta_1)^2] - c_n^T M_n^{-1} c_n$$

Even though we know the optimal solution, the actual polynomial that satisfies this is not too clear. Magically, the limiting forms of M and c have a block structure that lets us conclude that the minimizing q is always tree structured in this basis.

Lemma, Block structure of M and c

There exists limits M_∞ and c_∞ with

$$M_{n,AB} = M_{\infty,AB} + O(n^{-1/2}) \qquad c_{n,A} = c_{\infty,A} + O(n^{-1/2})$$

Where we have, with reorganizing the matrix M_∞ and c_∞ such that A, B that are trees are in the first row/columns and rows respectively, the following block structure

$$M_\infty = \begin{bmatrix} P_\infty & 0 \\ 0 & Q_\infty \end{bmatrix} \qquad c_\infty = \begin{bmatrix} d_\infty \\ 0 \end{bmatrix}$$

Interestingly, the choice to normalize each hermite polynomial by its expectation is chosen so this result holds.

[5]The proof for this is not to difficult but is rather tedious and not vital to the understanding of this result

[6]Specifically for this derivation we neglect the n subscript

High-Level Approach to c_∞ (and M_∞)

Proof. The analysis required to find the asymptotics for c_∞ is relatively straightforward. Our goal is to calculate

$$c_A = E[\mathcal{H}_A(Y)\psi(\theta_1)] = \frac{1}{\sqrt{|emb(A)|}} \sum_{\phi \in emb(A)} E[\mathcal{H}_{\mathbf{im}(\phi,A)}(Y)\psi(\theta_1)] = \sqrt{|emb(A)|}E[\mathcal{H}_\alpha(Y)\psi(\theta_1)]$$

The last step is justified as the expectation of $\mathcal{H}$ depends only on the shape of A, not the exact embedding ϕ (Recall that changing the embedding is only changing which variables are assigned to which edges, each of the variables are equally distributed so there expectation is determined only by the structure of A). Thus the last equality above is for some embedding α on A. We have the number of embeddings as

$$|emb(A)| = \binom{n-1}{|V(A)|-1}(|V(A)-1|)! = n^{|V(A)|-1}(1+o(1))$$

As we choose, for each non-rooted node, from $2:n$ as our pool of labels and then assign these labels to each node in the graph. As D is bounded (and there for $V(A)$) we have the asymptotic growth $\binom{n}{k} = \frac{n^k}{k!}(1+o(1))$.

Now we can analyze $E[\mathcal{H}_\alpha(Y)\psi(\theta_1)]$, clearly this will be 0 if A has a connected component, say C, not containing the root, as we would have

$$\begin{aligned}
E[\mathcal{H}_\alpha(Y)\psi(\theta_1)] &= E[\prod_{\gamma\in C}(h_\gamma - E[h_\gamma])\psi(\theta_1)]E[\prod_{\gamma\notin C}(h_\gamma - E[h_\gamma])\psi(\theta_1)] \\
&= E[\prod_{\gamma\in C}(h_\gamma - E[h_\gamma])]E[\psi(\theta_1)]E[\prod_{\gamma\notin C}(h_\gamma - E[h_\gamma])\psi(\theta_1)] \\
&= 0 \cdot E[\psi(\theta_1)]E[\prod_{\gamma\notin C}(h_\gamma - E[h_\gamma])\psi(\theta_1)] \\
&= 0
\end{aligned}$$

With some facts from the Hermite polynomials we can show that

$$E[\mathcal{H}_\alpha(Y)\psi(\theta_1)] = C_A n^{-\frac{1}{2}|E(A)|}$$

and thus

$$c_A = n^{\frac{1}{2}(|V(A)|-1-|E(A)|)}(1+o(1))$$

As only trees have $|V(A)| = 1 + |E(A)|$ then our result follows.

The analysis of M_∞ is more complicated as we now have to analyze

$$M_{AB} = \frac{1}{\sqrt{\phi_A \in emb(A), \phi_B \in emb(B)}} E[\mathcal{H}_{\mathbf{im}(\phi_A,A)}\mathcal{H}_{\mathbf{im}(\phi_B,B)}]$$

Here it is natural to analyze the intersection structure of the two embeddings (i.e. when the labels for nodes in embedding A and embedding B are equal), using some graph theoretic knowledge (alongside some facts about Hermite polynomials) we can eventually work down to

$$M_{AB} = n^{\varphi(A,B)}$$

With

$$\varphi(A,B) = \begin{cases} \leq -\frac{1}{2} & \text{if } A \in \mathcal{T}_{\leq D} \text{ and } B \in \mathcal{G}_{\leq D} \setminus \mathcal{T}_{\leq D} \text{ or vice versa} \\ \leq 0 & \text{Otherwise} \end{cases}$$

I leave the remainder of this analysis to the paper. ☙

Now we can defined $r(Y)$ as the restriction of the optimal $q(Y)$ found earlier to only tree graphs with

$$r(Y) = \sum_{T \in \mathcal{T}_{\leq D}} \hat{r}(Y)\mathcal{H}_T(Y) \qquad \hat{r} = P_\infty^{-1} d_\infty$$

As a direct result of all the ingredients we have here, we have, we can redo our analysis for the optimal q in the limiting case to get

$$d_\infty^T P_\infty^{-1} d_\infty = c_\infty^T M_\infty^{-1} c_\infty \qquad \text{(By Matrix Multiplication)}$$

and thus,

$$\lim_{n\to\infty} E[(r(y) - \psi(\theta_1))^2] = E[\psi(\theta_1)^2] - c_\infty^T M_\infty c_\infty$$

Which is knocking on the door of our original goal.

Change of basis and putting everything together

> **We Will Learn:** Now that we were able to show that the minimizing MSE polynomial is tree structured in our simpler basis we construct a change of basis up to vanishing error. We can then wrap everything up to show that the best polynomial is tree structured in a basis favorable to our AMP analysis

We have been fortunate enough to be able to show that there exists a tree structured polynomial in some basis of graph polynomials. Unfortunately, to be able to use message passing we have to have our polynomial be in terms of the $\mathcal{F}_T$ basis. Luckily such a result holds

Lemma, Change of Basis

For any fixed $A \in \mathcal{T}_{\leq D}$, there exists n-independent coefficients m_{AB} such that

$$\mathcal{H}_A(Y) = \sum_{B \in \mathcal{T}_{\leq D}} m_{AB}\mathcal{F}_B(Y) + \mathcal{E}_A(Y)$$

where $E[\mathcal{E}_A(Y)^2] = o(1)$.

For notaional convience we let $Y^\phi = \prod_{(i,j)\in E(B)} Y_{\phi(i)\phi(j)} = Y^{\mathbf{im}(\phi,B)}$. At a high-level this is done by writing

$$\mathcal{H}_A(Y) = \frac{1}{\sqrt{|emb(A)|}} \sum_{\phi \in emb(A)} Y^\phi - \sqrt{|emb(A)|}E[Y^{\phi^*}]$$

This simplification occurs as when A is a tree each edge occurs only once and thus we have only single degree herminite polynomials. We have ϕ^* is an arbitrary embedding, this is permissible as the expectation of Y is independent of the actual embedding as every upper

triangular Y is equally distributed. $\sqrt{|emb(A)|}E[Y^{\phi^*}]$ is just a constant so we can project it onto $F_\varnothing = 1$, we are left to show the leading term can be projected. We have

$$\begin{aligned}\frac{1}{\sqrt{|emb(A)|}}\sum_{\phi\in emb(A)} Y^\phi &= \frac{1}{\sqrt{|emb(A)|}}\sum_{\phi\in nr(A)} Y^\phi - \frac{1}{\sqrt{|emb(A)|}}\sum_{\phi\in nr(A)\setminus emb(A)} Y^\phi \\ &= \sqrt{\frac{|nr(A)|}{|emb(A)|}}\mathcal{F}_A(Y) - \frac{1}{\sqrt{|emb(A)|}}\sum_{\phi\in nr(A)\setminus emb(A)} Y^\phi \\ &= \mathcal{F}_A(Y) + \left(\sqrt{\frac{|nr(A)|}{|emb(A)|}} - 1\right)\mathcal{F}_A(Y) - \frac{1}{\sqrt{|emb(A)|}}\sum_{\phi\in nr(A)\setminus emb(A)} Y^\phi\end{aligned}$$

The middle term disappear as embeddings make up $1 - o(1)$ proportion of non-reversing labellings (there are n^D embeddings all other remaining labellings have a strictly smaller count in terms of the exponent of n). It is left to show that the last term can be suitably controlled, see the paper for how to do this.

We can now write

$$\begin{aligned}r(Y) &= \sum_{A\in\mathcal{T}_{\leq D}} \hat{r}_A\mathcal{H}_A(Y) \\ &= \sum_{A\in\mathcal{T}_{\leq D}} \hat{r}_A \sum_{B\in\mathcal{T}_{\leq D}} m_{AB}\mathcal{F}_B(Y) + o(1) \qquad (\text{As } |\mathcal{T}_{\leq}| = O(1) = \hat{r}_A) \\ &= \sum_{B\in\mathcal{T}_{\leq D}} (\sum_{B\in\mathcal{T}_{\leq D}} \hat{r}_A m_{AB})\mathcal{F}_B(Y) \\ &:= \sum_{B\in\mathcal{T}_{\leq D}} \hat{p}\mathcal{F}_B(Y) \\ &= p(Y)\end{aligned}$$

And with some careful book keeping we can extend all our our $r(Y)$ results to $p(Y)$. This isn't too difficult as everything we are dealing with is bounded independent of n.

3.2. **The evaluation of a tree structured polynomial can be related to a Message passing algorithm.**

We Will Learn: Using what we leaned from [Bayati et al., 2015], we will construct a message passing algorithm that can be well approximated by AMP in the limit. This message passing algorithm recursively evaluates any low degree polynomial in at most time $O(Dn^2)$ (already a cool result). With a slight modification from our results in section 2 we can show that this algorithm has an AMP state evolution and therefore is always beaten by the Bayes-AMP.

Similar to [Bayati et al., 2015] we define a message passing algorithm s by [7]

$$s^0_{i\to j} = 0$$
$$s^{t+1}_{i\to j} = \frac{1}{\sqrt{n}} \sum_{[n]\setminus\{i,j\}} Y_{ik} F_t(s^t_{k\to i}) \qquad \text{(As messages } s_{i\to i} = 0 \text{ by convention here)}$$
$$s^{t+1}_i = \frac{1}{\sqrt{n}} \sum_{k\in[n]\setminus i} Y_{ik} F_t(s^t_{k\to i})$$
$$\hat{s}^{t+1}_i = F_t(s^{t+1}_i)$$

Note that we are actually defining a family of recursions here as we can construct an enumeration of $\mathcal{T}_{\leq D}$, $(T_1, \cdots, T_{|\mathcal{T}_{\leq D}|})$ and define a matrix type message passing algorithm where each column is each element of this enumeration. For convenience we will also define a graph for each tree T known as T_+ by the following algorithm:

(1) connect an edge to $\circ$ in T to a new node v_+
(2) make the root of this new graph T_+ to be $\circ = v_+$

We also define the class of children of root for T, $\mathcal{D}(T)$, as the set of sub-graphs which are connected to the root by a single edge. For example:

We define a special operation f^* as

$$F^*(s)(T) = \prod_{T'\in\mathcal{D}(T)} s(T')$$

This operation defined the value of a given tree T as the product of its sub trees.

Notice that the normalization of this message passing algorithm does not perfectly match our non-reversing polynomial basis, for example when we unroll this algorithm we see that

$$\mathcal{F}_{T,i\to j} = \frac{1}{n^{|E(T)|/2}} \sum_{\phi\in nr(T,i\to j)} \prod_{(u,v)\in E(T)} Y_{\phi(u),\phi(v)}$$

where $nr(T, i \to j)$ is a non-reversing label rooted at i and all direct children of the root do not have label j. Luckily we get luck as these normalizations are equivalent in the limit.

[7] Now is a good time to compare this recursions to section 2

For $nr(T)$ we have $(n-1)$ choices for each direct child of the of $\circ$ and $(n-2)$ for every other node. For $nr(T, i \to j)$ we have $(n-2)$ choices for each node.

In a tree each node (except the root, which is fixed) corresponds to one edge, thus

$$\frac{|nr(T, i \to j)|}{n^{|E(T)|}} = 1 + o(1) = \frac{|nr(T)|}{n^{|E(T)|}}$$

Wrapping this all together we have the following result:

Proposition, This Message Passing Algorithm works Let us have the message passing algorithm s with $F_t = F^*$ defined above then we have for a specific tree T with $t = depth(T)$(the tree depth),

$$\begin{aligned} s^t_{i\to j}(T) &= \mathcal{F}_{T_+, i\to i}(Y) \\ s^t_1(T) &= (1+o(1))\mathcal{F}_{T_+}(Y) \\ \hat{s}^t_1(T) &= (1+o(1))\mathcal{F}_T(Y) \end{aligned}$$

This can be seen by unrolling this message passing algorithm at $s^t_{i\to j}(T)$ as we did previously in section 1 (albeit complicated). This unrolling will naturally sum over all the nodes in a graph with the non-reversing structure. A good intuition for what is going on is that, when we calculate the iterate t, we want to have the evaluation of all trees T with depth t. Starting from the root, we have the multiplication with Y representing each of the edges with a unified label. Then the non-linearity finds the value of each sub trees (removing the root and edges attached to the root as they have already been unrolled) according to F^*, we have these values in the $t-1$ iterate. Inductively repeating the unrolling builds any tree of depth t in t iterates. To solidify this we know that the first iteration make all the single depth trees, closing the induction loop.

Summing over all trees T with linear coefficients $\hat{p}_T$ gives us the evaluation of the polynomial. We have $O(n^2)$ operations by enumeration. The AMP algorithm runs in the tree depth of T, which is at most D, thus this algorithm runs in $O(Dn^2)$ time. Riding on our results from section 2 (extended to the matrix case), e can immediately see that this algorithm has a state evolution, specifically with $(\Theta, G_t) \sim \pi_\Theta \otimes N(0, \Sigma_t)$

$$\begin{aligned} \lim_{n\to\infty} E[\psi(s^t_{1\to 2}, \theta_1)] = \lim_{n\to\infty} E[\psi(s^t_1, \theta_1)] &= E[\psi(\mu_t\Theta + G_t, \Theta)] \\ \lim_{n\to\infty} E[\psi(\hat{s}^t_1, \theta)] &= E[\psi(F_t(\mu_t\Theta + G_t), \Theta)] \end{aligned}$$

Now that we are fueled with the power of AMP, the exact coefficients of our vector c_∞ and matrix M_∞ easily with[8]:

$$\begin{aligned} MSE_{\leq D} &= \lim E[(\theta_1 - \sum_{T\in\mathcal{T}_{\leq D}} \hat{p}_T\mathcal{F}_T(Y))^2] \\ \lim E[\theta_1\mathcal{F}_T(Y)] &= E[\Theta F^*(\mu_t\Theta + G_t)(T)] = \mu_{t+1}(T) = c_{\infty,T} \\ \lim E[\mathcal{F}_{T'}(Y)\mathcal{F}_T(Y)] &= E[F^*(\mu_t\Theta + G_t)(T')F^*(\mu_t\Theta + G_t)(T)] = \sigma_{t+1}(T, T') = M_{\infty,TT'} \end{aligned}$$

[8]All lim are with respect to $n \to \infty$

Thus the optimal MSE given by a low degree estimator is

$$MSE_{\leq D} = \lim E[(\theta_1 - \sum_{T \in \mathcal{T}_{\leq D}} \hat{p}_T \mathcal{F}_T(Y))^2] = E[\Theta^2] - 2\hat{p}^T \mu_{t+1} + \hat{p}^T \Sigma_{t+1} \hat{p}$$

Which is equivalent to an AMP algorithm (under the extension to matrices) with $d = |\mathcal{T}_{\leq D}|$, $F_t = F^*$ and $g_t = \hat{p}^T F^*(x^t)$. Of which the Bayes-AMP estimator is known to be better than, the conclusion follows.

4. Future Directions / Applications / Relevant Literature

The main three papers that developed the tools needed for this analysis are [Bayati et al., 2015, Montanari and Wein, 2022, Feng et al., 2021]. Here we mention some possible extension and good resources to learn more. I have assigned stars to the difficulty of these possible extensions.

Immediate Applications(*) Under some very simple priors $\pi_\Theta(\lambda)$ (λ being a hyper parameter of the prior) once can derive the exact value of $q_A MP$ and for which hyper parameters λ allow a non-trivial Bayes-AMP MSE. Under this new result this analysis automatically derives the threshold for low-degree polynomials. An example of such an analysis is given in [Feng et al., 2021]. For an example which is low degree see [Montanari and Wein, 2022].

Universality (**) In many cases AMP algorithms have been extended to be "universal" where we can replace the matrix Z in $(\star)$ by some other Wigner matrix W. A common choice is some type of normalized Radmacher matrix for random graph models. For example [Deshpande et al., 2015] provides analysis of the stochastic block model (SBM) where AMP is utilized to find the asymptotic per-vertex mutual information between a vertex label and the entire SBM graph. This information is related to a phase transition of the information theoretic impossibility of recovery in SBM. This analysis required a nasty interpolation to model $(\star)$. It was shown in [Wang et al., 2023] that such an interpolation is unnecessary and the SBM model has the same state evolution and the analogous spiked Wigner model. To my knowledge there is no exact result that optimal algorithm is still Bayes AMP with a generic Wigner matrix but certainly (excluding issues of taking limits) seems very intuitive. Once this is establish then the upper bound proof can be immediately established, the lower bound proof relies heavily on the properties of the hermite polynomials which are chosen due to the Guassian measure of Z. A next step would be to try to redo this analysis with a similar set of orthonormal polynomials with $Z_{ij} \sim Rad(1/2)$.

Extensions to $D = \log(n)$, (*) Montanari explicitly mentions this in the paper than once could remove the $E[\Theta] \neq 0$ requirement if you could show the equivalence bounds for slowly growing degree, say $D = \log(n)$ (Montanari conjectures that may be true for $D = n^{.99}$). This would also be convenient since many times AMP needs a spectral iteration to have nice convergence properties, which requires running a "burn-in" of $\log(n)$ number of iterates. Technically this is somewhat of a burden to show since one needs to confirm that all of the approximations made throughout the paper have $o(\log(n))$ (which shouldn't be too tricky) and some conceptual issues with objects in the proof slowly growing. Overall the stretch to $\log(n)$ shouldn't be too much of a burden, just tedious. The extension to $n^{.99}$ will not be so trivial as there are some error rates here only of $o(n^{-1/2})$.

Extension to rectangular algorithms (*)

I am surprised that [Montanari and Wein, 2022] didn't include this in the very end of the paper. There are usually very straightforward way to embed a rectangular AMP algorithm inside of standard AMP algorithm. The benefit of this approach is that we could generalize our model ($\star$) to

$$Y = \frac{1}{\sqrt{n}} uv^t + \breve{Z}$$

where $Y \in \mathbb{R}^{m \times n}$ with $\frac{m}{n} \to \delta$, $\breve{Z}_{ij} \overset{iid}{\sim} N(0,1)$, $u_i \overset{iid}{\sim} \pi_U$ and $v_i \overset{iid}{\sim} \pi_V$. Perhaps there is much more difficulty as you have to have some awkward Hermite basis that has normalizations based on m, n, δ which is not straightforward. An example of embedding a rectangular AMP algorithm into a symmetric AMP algorithm can be seen in [Berthier et al., 2017].

Extensions to regression models (**)

($\star$) is one of two "fundamental" models for which we have an AMP algorithm. The second of which are high-dimensional GLM models which take the form of

$$Y_i = q(X_i^T \beta, \omega_i)$$

With $i \in [m]$ and $X_i, \beta \in \mathbb{N}$ and $\omega_i \in \mathbb{R}$ is some noise. We assume that $X_i \sim N(0, 1/n)$. The hope would be that we could perform some similar analysis to show that the best polynomial estimator in this model is tree-structured. A good first step would be to consider the trivial $q = I$ function. See [Feng et al., 2021] for an introduction to these models and [Tan and Venkataramanan, 2023] for how complicated they can get.

Extension to non-separable AMP (***)

Many cases it is not reasonable to think that the prior π_Θ is naturally an *iid* distribution. There may be very important correlations between our prior that can be codified as some outside information. Working this information into the algorithm is not immediately straightforward as the Bayes AMP functions become non-separable over a growing number of parameters. In this project we have discussed AMP under the notion of non-separable non-linear function, [Berthier et al., 2017] gave an analysis of separable AMP algorithms which are not just suited to this cause but many other interesting methods for model ($\star$). Unfortunately there is a lot of uncertainty in this direction as separable functions are rather delicate in the AMP analysis, the benefit may be worth it as they can cover a much larger class of algorithms. Personally I believe that these algorithms may be useful one day for analyzing local search algorithms (see point below)

Equivalence of other "hardness" classes (***)

It has been shown previously in [Celentano et al., 2020] that AMP is also known to characterize the ability for a class of algorithms with access to gradient information. It would be interesting to see how other definitions of hardness could be reduced to an AMP algorithm. Many algorithms can be reduce to message passing. The fact that message passing can be so versatile and that many message passing algorithms can be modified to an AMP algorithm is a very fruitful technique to keep in mind. I think it would be interesting if there were some way to convert local search algorithms, say MCMC, into a message passing algorithm. Although I am not sure to what extent other tools mentioned above need to catch up to make this possible.

References

[Bayati et al., 2015] Bayati, M., Lelarge, M., and Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2).

[Berthier et al., 2017] Berthier, R., Montanari, A., and Nguyen, P.-M. (2017). State evolution for approximate message passing with non-separable functions.

[Bolthausen, 2012] Bolthausen, E. (2012). An iterative construction of solutions of the tap equations for the sherrington-kirkpatrick model.

[Celentano et al., 2020] Celentano, M., Montanari, A., and Wu, Y. (2020). The estimation error of general first order methods.

[Deshpande et al., 2015] Deshpande, Y., Abbe, E., and Montanari, A. (2015). Asymptotic mutual information for the two-groups stochastic block model. *CoRR*, abs/1507.08685.

[Feng et al., 2021] Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. (2021). A unifying tutorial on approximate message passing.

[Montanari and Wein, 2022] Montanari, A. and Wein, A. S. (2022). Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation.

[Tan and Venkataramanan, 2023] Tan, N. and Venkataramanan, R. (2023). Mixed regression via approximate message passing.

[Wang et al., 2023] Wang, T., Zhong, X., and Fan, Z. (2023). Universality of approximate message passing algorithms and tensor networks.