The Application of Approximate Message Passing in Low Rank Matrix Estimation

Wanshan Li

1 Background

Low-rank matrix estimation is a very important problem in high-dimensional statistics and statistical theory for complex-structured data like random graph or network. In a very rough way, we might conclude that there are two main directions in the research of low-rank matrix estimation.

One direction, represented by recent works [4, 1], is to analyze the statistical properties of an M-estimator with a close-form definition under some sufficient conditions. For instance, [4] study spectral estimators and [1] focus on an M-estimator minimizing the penalized approximation error.

The other direction, represented by a series of works [2, 3, 6, 5, 7], is to seek for the optimal estimator and fundamental limits of the estimation of low-rank matrices. Rather than studying an clearly-defined M-estimator, this series of works either apply approximate message passing to design an estimator via iterations ([2, 3, 6, 5, 7]), or study the lower bound of the estimation problem, i.e., without discussing any specific estimator ([5]).

The goal of my project is to see how the AMP procedure is designed and used to study the challenging problems in low-rank matrix estimation by understanding the results in [7].

2 A review of Montanari & Venkataramanany 2019

2.1 Summary

The main idea of the application of AMP in low rank matrix estimation is to correct the solution path of the nonlinear power iteration $\boldsymbol{x}_{PI}^{t+1} = \boldsymbol{A} f_t(\boldsymbol{x}_{PI}^t)$ so that precise analysis of the distribution is possible, and consequently f_t can be picked in a smart way to improve the signal-to-noise ratio of the solution at each iterates. It initializes the procedure at the spectral estimate, so intuitively it must be able to get closer to the information boundary than simple spectral estimates.

The main results of [7] are Theorem 5 and its special case Theorem 1. Theorem 1 has two applications:

1 To derive sharp asymptotics in the moderate SNR regime (λ, ε of order one) of the sparse spike model, where $\boldsymbol{x}_0$ has at most $n\varepsilon$ nonzero entries for some $\varepsilon \in (0, 1)$.

2 To build a Bayes AMP algorithm which can give achieve the Bayes-optimal error under a certain condition.

Essentially the first application just exploits the specific structure and then applies Theorem 1. I think the most interesting part there is how to construct suitable ψ. I will focus on the second application as it gives the asymptotic behavior as $t \to \infty$, although the part of $\lim_{t\to\infty}$ is not very hard as the involving state evolution is relatively clear and simple (which is kind of surprising).

The proof of Theorem 1 is straightforward given Lemma A.1 and Lemma B.3. The idea behind the proof is similar to the proof in the linear model case we discussed in the class: introduce a copy of the key random variable to decouple it from the "solution path" in AMP.

Theorem 2 and following corollary says that AMP can achieve Bayes-optimal error under a certain condition. But they just give one fairly simple example where such condition can be met. This seems to make the conclusion of Bayes-optimal restrictive in some way.

2.2 Main Theorem: Theorem 1

Let $\boldsymbol{x}_0 = \boldsymbol{x}_0(n)\mathbb{R}^n$ be a sequence of signals indexed by the dimension n, satisfying the following conditions:

(i) Their rescaled ℓ_2-norms converge $\lim_{n\to\infty} \|\boldsymbol{x}_0(n)\|_2/\sqrt{n} = 1$;

(ii) The empirical distributions of the entries of $\boldsymbol{x}_0(n)$ converges weakly to a probability distribution ν_{X_0} on $\mathbb{R}$, with unit second moment.

We then consider the following spiked model, for $W \sim \mathrm{GOE}(n)$, i.e., $\boldsymbol{W} = \boldsymbol{W}^\top$, $(W_{ij})_{i\le j\le n}$ are independent with $(W_{ii})_{i\le n} \overset{iid}{\sim} N(0, 2/n)$ and $(W_{ij})_{i<j\le n} \overset{iid}{\sim} N(0, 1/n)$.

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{x}_0\boldsymbol{x}_0^\top + W. \tag{1}$$

In order to estimate $\boldsymbol{x}_0$, we compute the principal eigenvector of $\boldsymbol{A}$, to be denoted by φ_1, and apply the following iteration, with initialization $\boldsymbol{x}^0 = \sqrt{n}\,\varphi_1$:

$$\boldsymbol{x}^{t+1} = \boldsymbol{A} f_t\left(\boldsymbol{x}^t\right) - b_t f_{t-1}\left(\boldsymbol{x}^{t-1}\right), \quad b_t = \frac{1}{n}\sum_{i=1}^{n} f_t'\left(\boldsymbol{x}_i^t\right) \tag{2}$$

Here $f_t(\boldsymbol{x}) = \left(f_t\left(x_1\right), \ldots, f_t\left(x_n\right)\right)^\top$ is a separable function for each t.

Theorem 1. *Consider the $k = 1$ spiked matrix model of* (1)*, with $\boldsymbol{x}_0(n) \in \mathbb{R}^n$ a sequence of vectors satisfying assumptions (i), (ii) above, and $\lambda > 1$. Consider the AMP iteration in* (2) *with initialization $\boldsymbol{x}^0 = \sqrt{n}\,\varphi_1$ (where, without loss of generality $\langle \boldsymbol{x}_0, \varphi_1\rangle \ge 0$). Assume $f_t : \mathbb{R} \mapsto \mathbb{R}$ to be Lipschitz continuous for each $t \in \mathbb{N}$.*

Let $(\mu_t, \sigma_t)_{t\ge 0}$ be defined via the recursion

$$\mu_{t+1} = \lambda\mathbb{E}\left[X_0 f_t\left(\mu_t X_0 + \sigma_t G\right)\right], \sigma_{t+1}^2 = \mathbb{E}\left[f_t\left(\mu_t X_0 + \sigma_t G\right)^2\right], \tag{3}$$

where $X_0 \sim \nu_{X_0}$ and $G \sim N(0,1)$ are independent, and the initial condition is $\mu_0 = \sqrt{1-\lambda^{-2}}$, $\sigma_0 = 1/\lambda$.

Then, for any function $\psi : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}$ with $|\psi(\boldsymbol{x}) - \psi(\boldsymbol{y})| \le C\left(1 + \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2\right)\|\boldsymbol{x}-\boldsymbol{y}\|_2$ for a universal constant $C > 0$, the following holds almost surely for $t \ge 0$:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi\left(x_{0,i}, x_i^t\right) = \mathbb{E}\left\{\psi\left(X_0, \mu_t X_0 + \sigma_t G\right)\right\}. \tag{4}$$

The key point in the proof of Theorem 1 is to use an approximate representation for the conditional distribution of $\boldsymbol{A}$ given $(\boldsymbol{\varphi}_1, z_1)$, given by

$$\tilde{A} = z_1\boldsymbol{\varphi}_1\boldsymbol{\varphi}_1^\top + P^\perp\left(\lambda v v^\top + \tilde{W}\right)P^\perp, \tag{5}$$

where $(\boldsymbol{\varphi}_1, z_1)$ are the principal eigenvector and eigenvalue of $\boldsymbol{A}$ in (1), $\tilde{\boldsymbol{W}} \sim \mathrm{GOE}(n)$ is independent of $\boldsymbol{W}$, and the matrix $\boldsymbol{P}^\top = \boldsymbol{I} - \boldsymbol{\varphi}_1\boldsymbol{\varphi}_1^\top$ is the projector onto the space orthogonal to $\boldsymbol{\varphi}_1$.

Lemma A.1. *Consider the modified spiked model* (5) *and let $\tilde{\boldsymbol{x}}^t$ be the AMP sequence obtained by replacing $\boldsymbol{A}$ with $\tilde{\boldsymbol{A}}$, namely*

$$\tilde{\boldsymbol{x}}^0 = \sqrt{n}\,\mathrm{sign}\left(\langle \boldsymbol{v}, \boldsymbol{\varphi}_1\rangle\right)\boldsymbol{\varphi}_1, \quad \tilde{\boldsymbol{x}}^{t+1} = \tilde{A} f\left(\tilde{\boldsymbol{x}}^t; t\right) - b_t f\left(\tilde{\boldsymbol{x}}^{t-1}; t-1\right). \tag{6}$$

Then the state evolution statement (4) *holds with $\boldsymbol{x}^t$ replaced by $\tilde{\boldsymbol{x}}^t$.*

Let $\boldsymbol{A}$ be a general spiked matrix given as

$$A = \sum_{i=1}^{k}\lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^\top + W, \tag{7}$$

with non-random $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and $\boldsymbol{v}_i \in \mathbb{R}^n$, and $\boldsymbol{W} \sim \mathrm{GOE}(n)$.

Now we consider the case where $\lambda_1 \geq \ldots \lambda_{k_+} > 1 > \lambda_{k_++1}$ and $\lambda_{k-k_-} > -1 > \lambda_{k-k_-+1} \geq \cdots \geq \lambda_k$. Let $\boldsymbol{z} = (z_1, \ldots, z_n)$ be the ordered eigenvalues of $\boldsymbol{A}$ with corresponding eigenvectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$. Denote $\hat{S} = \{1, \ldots, k_+\} \cup \{n-k_-+1, \ldots, n\}$, $S = \{1, \ldots, k_+\} \cup \{k-k_-+1, \ldots, k\}$, and $k_* = k_+ + k_-$. Let $\boldsymbol{\lambda}_S = (\lambda_i)_{i \in S}$, $\boldsymbol{z}_{\hat{S}} = (z_i)_{i \in \hat{S}}$, and $\Phi_{\hat{S}} = (\boldsymbol{\varphi}_i)_{i \in \hat{S}}$.

Lemma B.3. *With the definitions above, let*

$$\tilde{\boldsymbol{A}} \equiv \sum_{i \in S} z_i \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^\top + \sum_{i=1}^{k} \lambda_i \boldsymbol{P}^\perp \boldsymbol{v}_i \boldsymbol{v}_i^\top \boldsymbol{P}^\perp + \boldsymbol{P}^\perp \tilde{\boldsymbol{W}} \boldsymbol{P}^\perp, \tag{8}$$

where $\boldsymbol{P}^\top$ is the projector onto the orthogonal complement of the space spanned by $(\boldsymbol{\varphi}_i)_{i \in \hat{S}}$, and $\tilde{\boldsymbol{W}} \sim \mathrm{GOE}(n)$ is independent of $\boldsymbol{W}$. Let $\rho(x) = x + x^{-1}$, $\eta(x) = 1 - x^{-2}$, and define the event

$$\mathcal{E}_\varepsilon \equiv \left\{ \max_{i \leq k_*} |(\boldsymbol{z}_S)_i - \rho(\lambda_{S,i})| \leq \varepsilon, \min_{i \in S} \left\| \Phi_S^\top \boldsymbol{v}_i \right\|_2^2 - \eta(\lambda_i) \geq -\varepsilon \right\}. \tag{9}$$

Then there exists a constant $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$ there is $c(\varepsilon) > 0$, such that

$$\mathbb{P}\{\mathcal{E}_\varepsilon\} \geq 1 - \frac{1}{c(\varepsilon)} e^{-nc(\varepsilon)}. \tag{10}$$

Further (for a suitable version of the conditional probabilities):

$$\sup_{(\boldsymbol{z}_{\hat{S}}, \boldsymbol{\Phi}_{\hat{S}}) \in \mathcal{E}_\varepsilon} \left\| \mathbb{P}\left(\boldsymbol{A} \in \cdot | \boldsymbol{z}_{\hat{S}}, \boldsymbol{\Phi}_{\hat{S}}\right) - \mathbb{P}\left(\tilde{\boldsymbol{A}} \in \cdot | \boldsymbol{z}_{\hat{S}}, \boldsymbol{\Phi}_{\hat{S}}\right) \right\|_{TV} \leq \frac{1}{c(\varepsilon)} e^{-nc(\varepsilon)}. \tag{11}$$

Proof of Theorem 1. For any $\varepsilon \in (0, \varepsilon_0)$, Lemma B.3 bounds the total variation distance between the conditional joint distributions of $(\tilde{\boldsymbol{A}}, \boldsymbol{\varphi}_1)$ and $(\boldsymbol{A}, \boldsymbol{\varphi}_1)$ given $(\boldsymbol{\varphi}_1, z_1) \in \boldsymbol{\mathcal{E}}_\varepsilon$, where

$$\mathcal{E}_\varepsilon = \left\{ \left| z_1 - \left(\lambda + \lambda^{-1}\right) \right| \leq \varepsilon, \quad \left(\boldsymbol{\varphi}_1^\top \boldsymbol{v}\right)^2 \geq 1 - \lambda^{-2} - \varepsilon \right\}. \tag{12}$$

Since $\tilde{\boldsymbol{x}}^t$ and $\boldsymbol{x}^t$ are obtained by applying the same deterministic algorithm to $(\tilde{\boldsymbol{A}}, \boldsymbol{\varphi}_1)$ and $(\boldsymbol{A}, \boldsymbol{\varphi}_1)$ it follows that there exists a coupling of the laws of $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$ such that, for $(\boldsymbol{\varphi}_1, z_1) \in \boldsymbol{\mathcal{E}}_\varepsilon$

$$\mathbb{P}\left\{ \sum_{i=1}^{n} \psi\left(x_i^t, \tilde{v}_i\right) \neq \sum_{i=1}^{n} \psi\left(\tilde{x}_i^t, \tilde{v}_i\right) | z_1, \boldsymbol{\varphi}_1 \right\} \leq \frac{1}{c(\varepsilon)} e^{-nc(\varepsilon)} \tag{13}$$

for some constant $c(\varepsilon) > 0$. With this coupling, we therefore have

$$\mathbb{P}\left\{ \sum_{i=1}^{n} \psi\left(x_i^t, \tilde{v}_i\right) \neq \sum_{i=1}^{n} \psi\left(\tilde{x}_i^t, \tilde{v}_i\right) \right\} \leq \mathbb{P}\left(\mathcal{E}_\varepsilon^c\right) + \frac{e^{-nc(\varepsilon)}}{c(\varepsilon)} \leq \frac{2e^{-nc(\varepsilon)}}{c(\varepsilon)}. \tag{14}$$

Therefore by Borel-Cantelli, $\sum_{i=1}^n \psi\left(x_i^t, \tilde{v}_i\right) = \sum_{i=1}^n \psi\left(\tilde{x}_i^t, \tilde{v}_i\right)$ eventually almost surely. Theorem 1 hence follows by applying Lemma A.1. □

2.3 Bayes-optimal Estimation

In Theorem 2 the initialization in AMP (2) is modified to be

$$x^0 = \sqrt{n\lambda^2\left(\lambda^2 - 1\right)} \boldsymbol{\varphi}_1. \tag{15}$$

To define the optimal nonlinearity, consider the scalar denoising problem of estimating X_0 from the noisy observation $Y = \sqrt{\gamma} X_0 + G$ ($X_0 \sim \nu_{X_0}, G \sim N(0,1) \in \mathbb{R}$ are independent scalar random variables). The minimum mean square error is

$$\mathrm{mmse}(\gamma) = \mathbb{E}\left\{ \left[X_0 - \mathbb{E}\left(X_0 | \sqrt{\gamma} X_0 + G\right) \right]^2 \right\} \tag{16}$$

The state evolution of the effective signal-to-noise ratio is

$$\begin{aligned} \gamma_0 &= \lambda^2 - 1, \\ \gamma_{t+1} &= \lambda^2 \left\{1 - \mathrm{mmse}\left(\gamma_t\right)\right\}. \end{aligned} \tag{17}$$

The optimal non-linearity $f_t(\cdot)$ after t iterations is the minimum mean square error denoiser for signal-to-noise ratio γ_t:

$$f_t(y) \equiv \lambda F\left(y; \gamma_t\right), \tag{18}$$

$$F(y;\gamma) \equiv \mathbb{E}\left\{X_0 | \gamma X_0 + \sqrt{\gamma} G = y\right\}. \tag{19}$$

After t iterations, we produce an estimate of $\boldsymbol{x}_0$ by computing $\hat{\boldsymbol{x}}^t(\boldsymbol{A}) \equiv f_t\left(\boldsymbol{x}^t\right)/\lambda = F\left(\boldsymbol{x}^t; \gamma_t\right)$. Such choice is called Bayes AMP.

Theorem 2. *Consider the above setting and assumption (i)(ii) with $\lambda > 1$ and $F(\cdot;\gamma) : \mathbb{R} \mapsto \mathbb{R}$ to be Lipschitz continuous for any $\gamma \in (0, \gamma]^2$, and the state evolution in* (17)*. Then, for any pseudo-C-Lipschitz function ψ with a universal $C > 0$, almost surely we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi\left(x_{0,i}, x_i^t\right) = \mathbb{E}_{X_0, Z}\left\{\psi\left(X_0, \gamma_t X_0 + \gamma_t^{1/2} Z\right)\right\} \tag{20}$$

where $X_0 \sim \nu_{X_0}$ and $Z \sim N(0,1)$ are mutually independent, and $\langle \varphi_1, \boldsymbol{x}_0 \rangle \geq 0$. In particular, let $\gamma_{ALG}(\lambda)$ denote the smallest strictly positive solution of the fixed point equation $\gamma = \lambda^2\left[1 - \mathrm{mmse}(\gamma)\right]$. Then the AMP estimate $\hat{\boldsymbol{x}}^t(\boldsymbol{A}) = f_t\left(\boldsymbol{x}^t\right)/\lambda$ achieves

$$\begin{aligned} \lim_{t\to\infty} \lim_{n\to\infty} \frac{\left|\langle \hat{\boldsymbol{x}}^t(\boldsymbol{A}), \boldsymbol{x}_0 \rangle\right|}{\left\|\hat{\boldsymbol{x}}^t(\boldsymbol{A})\right\|_2 \left\|\boldsymbol{x}_0\right\|_2} &= \frac{\sqrt{\gamma_{\mathrm{ALG}}(\lambda)}}{\lambda}, \\ \lim_{t\to\infty} \lim_{n\to\infty} \frac{1}{n} \min_{s\in\{+1,-1\}} \left\|s\hat{\boldsymbol{x}}^t(\boldsymbol{A}) - \boldsymbol{x}_0\right\|_2^2 &= 1 - \frac{\gamma_{\mathrm{ALG}}(\lambda)}{\lambda^2}. \end{aligned} \tag{21}$$

Proof of Theorem 2. Using Theorem 1, we know that (4) holds with μ_t, σ_t given via state evolution

$$\mu_{t+1} = \lambda \mathbb{E}\left[X_0 f_t\left(\mu_t X_0 + \sigma_t G; t\right)\right], \quad \sigma_{t+1}^2 = \mathbb{E}\left[f_t\left(\mu_t X_0 + \sigma_t G; t\right)^2\right] \tag{22}$$

and initial condition $\mu_0 = (\lambda^2 - 1)$, $\sigma_0^2 = (\lambda^2 - 1)$. The goal is to pick f_t such that (20) holds with γ_t defined via SE (17).

By Cauchy-Schwarz inequality, the signal-to-noise ration μ_{t+1}/σ_{t+1} is maximized by setting $f(y;t) = f_{\text{Bayes}}\left(y;t\right)$ where

$$f_{\text{Bayes}}\left(y;t\right) = \lambda \mathbb{E}\left\{X_0 | \mu_t X_0 + \sigma_t G = y\right\}, \tag{23}$$

and with this $f_{\text{Bayes}}\left(y;t\right)$ we have

$$\mu_{t+1} = \sigma_{t+1}^2 = \lambda^2 \mathbb{E}\left\{\mathbb{E}\left\{X_0 | \mu_t X_0 + \sigma_t G\right\}^2\right\} = \lambda^2\left\{1 - \mathrm{mmse}\left(\mu_t^2/\sigma_t^2\right)\right\}. \tag{24}$$

Thus $\mu_t = \sigma_t^2$ for all $t \geq 0$. Let $\gamma_t \equiv \mu_t^2/\sigma_t^2$, then it satisfies the SE (17), and by $\mu_t = \sigma_t^2$ and (4) we know that (20) holds.

To prove (21), we can choose suitable $\psi(x,y)$ for fixed t and use Theorem 1, and then let $t \to \infty$. For the first equation, the construction of ψ is described in the proof of Proposition 2.1; for the second one the choice for a fixed t is $\psi(x,y) = (x - \mathbb{E}[X_0 | \gamma_t X_0 + \sqrt{\gamma_t} G = y])$. Under such choice the right hand side of the two equations will be $\sqrt{\gamma_t}/\lambda$ and $1 - \gamma_t(\lambda)/\lambda^2$.

The last step is to show $\gamma_t \to \gamma_{ALG}$. Let $M_\lambda(\gamma) = \lambda^2\{1 - \mathrm{mmse}(\gamma)\}$. Since $\mathrm{mmse}(\gamma) \leq \mathbb{E}[X_0 - c(\gamma X_0 + \sqrt{\gamma_0} G)]^2$ for any c, by Cauchy-Schwartz and $\int x^2 \nu_{X_0}(\mathrm{d}x) = 1$ we have $\mathrm{mmse}(\gamma) \leq (1+\gamma)^{-1}$. Hence $M_\lambda(\gamma) \geq \frac{\lambda^2\gamma}{1+\gamma}$. Since $\gamma \mapsto \mathrm{mmse}(\gamma)$ is non-increasing, we know that $M_\lambda \gamma$ is non-decreasing and $M_\lambda \gamma > \gamma$ for $\gamma \in (0, \gamma_{ALG})$, $\gamma_0 \leq \gamma_{ALG}$. Thus by $\gamma_{t+1} = M_\lambda \gamma_t$ we know that $\gamma_t \to \gamma_{ALG}$. □

In [5] they give the Bayes-optimal accuracy in the rank-one estimation problem, and the result can be summarized as the following proposition.

Proposition 2.2. *Consider the spiked matrix model with $\boldsymbol{x}_0(n)$ a vector with $i.i.d.$ entries with distribution ν_{X_0} with bounded support and $\int x^2 \nu_{X_0}(\mathrm{d}x) = 1$. Then there exists a countable set $D \subset \mathbb{R}_{\geq 0}$ such that, for $\lambda \in \mathbb{R} \setminus D$, the Bayes-optimal accuracy in the rank-one estimation problem is given by*

$$\lim_{n\to\infty} \sup_{\hat{\boldsymbol{x}}(\cdot)} \mathbb{E}\left\{ \frac{\langle \hat{\boldsymbol{x}}(\boldsymbol{A}), \boldsymbol{x}_0\rangle^2}{\|\hat{\boldsymbol{x}}(\boldsymbol{A})\|_2^2 \, \|\boldsymbol{x}_0\|_2^2} \right\} = \frac{\gamma_{Bayses}\,(\lambda)}{\lambda^2}, \tag{25}$$

where the supremum is over (possibly randomized) estimators. Here $\gamma_{Bayes}(\lambda)$ is the fixed point of the recursion $\gamma_{t+1} = \lambda^2 \left\{1 - \mathrm{mmse}\,(\gamma_t)\right\}$ that maximizes the following free energy functional

$$\Psi(\gamma, \lambda) = \frac{\lambda^2}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + I(\lambda), \tag{26}$$

where $I(\gamma) = \mathbb{E} \log \frac{\mathrm{p}_{Y|X_0}}{\mathrm{d}P_Y}(Y, X_0)$ is the mutual information for the scalar channel $Y = \sqrt{\gamma} X_0 + G$, with $X_0 \sim \nu_{X_0}$ and $G \sim N(0,1)$ mutually independent.

Theorem 2 precisely characterizes the gap between the Bayes-optimal estimation and message passing algorithms for rank-one matrix estimation. Ultimately we can summarize the following result.

Corollary 2.3. *Consider the setting of Theorem 2 and function $\Psi(\gamma, \lambda)$. Then Bayes-AMP asymptotically achieves the Bayes-optimal error (and $\gamma_{ALG}(\lambda) = \gamma_{Bayes}(\lambda)$) if and only if the global maximum of $\gamma \mapsto \Psi(\gamma, \lambda)$ over $(0, \infty)$ is also the first stationary point of the same function (as γ grows).*

As is pointed out in [7], one case where Corollary 2.3 holds is the two-points mixture

$$\begin{aligned} \nu_{X_0} &= \varepsilon \delta_{a_+} + (1-\varepsilon)\delta_{-a_-}, \\ a_+ &= \sqrt{\frac{1-\varepsilon}{\varepsilon}}, \quad a_- = \sqrt{\frac{\varepsilon}{1-\varepsilon}}. \end{aligned} \tag{27}$$

Notice that $\int x \nu_{X_0} \mathrm{d}x = 0$ and $\int x^2 \nu_{X_0} \mathrm{d}x = 1$. For ε close enough to $1/2$, $\gamma_{t+1} = \lambda^2 \left\{1 - \mathrm{mmse}\,(\gamma_t)\right\}$ only has one stable fixed point that is also the minimizer of the free energy function $\Psi(\gamma, \lambda)$. Hence $\gamma_{ALG}(\gamma) = \gamma_{Bayes}(\lambda)$ for all values of γ and the Bayes AMP is Bayes optimal.

3 Conclusion and Future work

In a rough sense I got the point that makes AMP "sharper" than other directions. The word "sharper" has two-fold meaning:

1 In $M_{ij} = M_{ij}^* + W_{ij}$, AMP approach assumes $\sigma = 1/\sqrt{n}$ but other approaches assume weaker noise, for instance [1] assumes $\sigma \lesssim 1/\sqrt{n \log^2 n}$.

2 It can achieve Bayes-optimal error under a certain condition. And even without that further condition, no polynomial-time algorithm is known that achieves estimation accuracy superior to the one guaranteed by Theorem 2.

It is pretty clear that AMP is "sharper" because it starts at the spectral estimates and it is flexible enough to pick the optimal function f_t (maximizing the signal-to-noise ratio) at each step t, as is illustrated in section 2.1 and the proof of Theorem 2 in [7]. But this seems to be too rough a conclusion, because it is not always the case that step-wise optimality leads to ultimate optimality. Perhaps this is the reason why it seems hard to find the case where the condition for Bayes-optimality holds. This leaves some future reading for me.

References

[1] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *arXiv:1906.04159*, 2019.

[2] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201, June 2014.

[3] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 12 2016.

[4] Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Asymptotic theory of eigenvectors for large random matrices. *arXiv:1902.06846*, 2019.

[5] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929, Apr 2019.

[6] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403, jul 2017.

[7] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *arXiv:1711.01682*, 2019.